Bayesian IV

Peter Rossi

Anderson School | UCLA

joint with Rob McCulloch, Tim Conley, and Chris Hansen

Motivation

IV problems are often done with a "small" amount of sample information (weak/many instruments).

It would seem natural to apply a small amount of prior information, e.g. price elasticities are unlikely to be outside (-1,-5).

Another nice example– instruments are not exactly valid. They have some small direct correlation with the outcome/unobservables.

BUT, Bayesian methods (until now) are tightly parametric. Do I always have to make the efficiency/consistency tradeoff as in std IV?

Overview

Consider parametric (normal) model first

Consider finite mixture of normals for error dist

Make the number of mixture components random and possibly “large”

Conduct sampling experiments and compare to state of the art classical methods of inference

Consider some empirical examples where being a non-parametric Bayesian helps!

Show how a Bayesian would deal with instruments that are not strictly valid.

The Linear Case

Linear Structural equations (perhaps in latent vars) are central in applied work. Many examples in both marketing and economics literatures. Derived Demand from referees!

This is a *relevant* and simple ex:

$$(1)\ x = \delta z + \varepsilon_1$$
$$(2)\ y = \beta x + \varepsilon_2$$
$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \sim N(0, \Sigma)$$

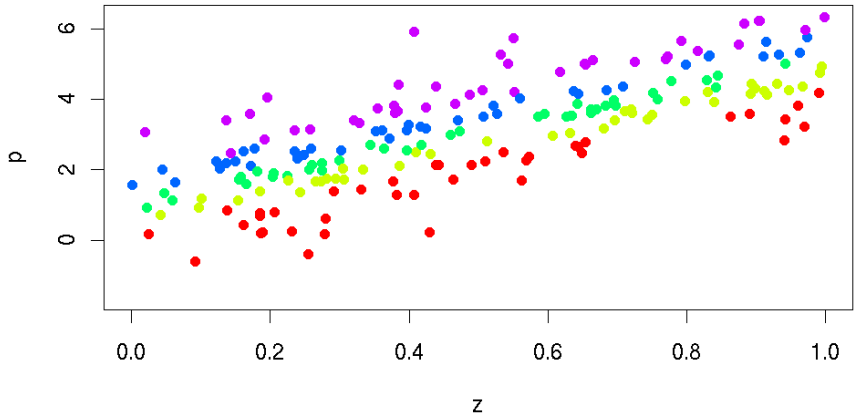
The Linear Case – IV Illustrated

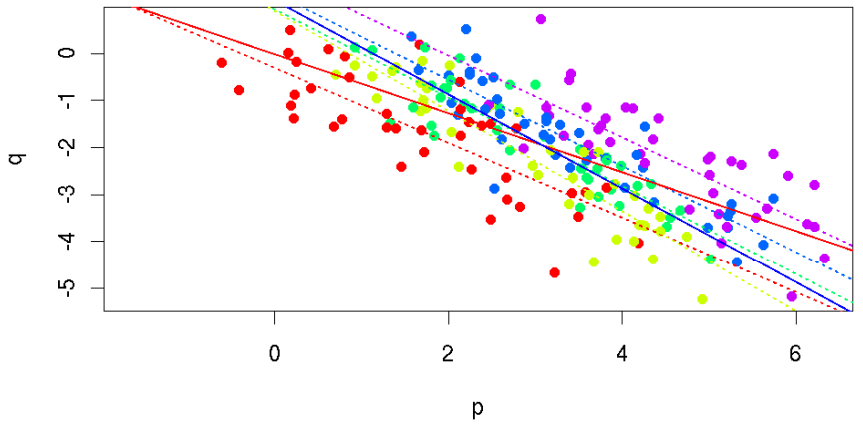
A simple example

$$p = 4z + \varepsilon_1$$

$$q = -1p + \varepsilon_2$$

$$corr(\varepsilon_1, \varepsilon_2) = .8$$





The Likelihood $\ell(\beta,\delta,\Sigma) = p(x,y|\beta,\delta,\Sigma)$

Derive the joint distribution of y, x | z.

$$(1)\ \ x = \delta z + \varepsilon_1$$

$$(2')\ y = \beta\delta z + (\beta\varepsilon_1 + \varepsilon_2)$$

or

$$(1)\ \ x = \pi_x z + v_1$$

$$(2')\ y = \pi_y z + v_2$$

$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \sim N(0,\Omega)$$

$$\beta = \frac{\pi_y}{\pi_x}$$

$$\Omega = A\Sigma A';\ A = \begin{bmatrix} 1 & 0 \\ \beta & 1 \end{bmatrix}$$

Priors

Which parameterization should you use?

Are independent priors acceptable?

$$p(\delta, \beta, \Sigma) = p(\delta)p(\beta)p(\Sigma)$$

$$p(\pi_x, \pi_y, \Omega) = p(\pi_x)p(\pi_y)p(\Omega)$$

reference prior situation

A Gibbs Sampler

(1) $\beta \mid \delta, \Sigma, x, y, z$

(2) $\delta \mid \beta, \Sigma, x, y, z$

(3) $\Sigma \mid \delta, \beta, x, y, z$

Tricks (`rivGibbs` in *bayesm*):

(1) given $\mathcal{G}$, convert structural equation into standard Bayes regression. We "observe" ε_1 Compute $\varepsilon_2 \mid \varepsilon_1$.

(2) given β, we have a two regressions with same coefficients or a restricted MRM.

Gibbs Sampler: beta draw $\beta \mid \delta, \Sigma, x, y, z$

Given δ, we observe ε_1. We rewrite the structural equation as

$$y = \beta x + \varepsilon_2 \mid \varepsilon_1$$

where $\varepsilon_2 \mid \varepsilon_1$ refers to the conditional distribution of ε_2 given ε_1.

$$\varepsilon_2 \mid \varepsilon_1 = \frac{\sigma_{12}}{\sigma_{11}} \varepsilon_1 + v_2; \quad \sigma_{v_2}^2 = \sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}}$$

$$\left(y - \frac{\sigma_{12}}{\sigma_{11}} \varepsilon_1 \right) \Big/ \sigma_{v_2} = \beta x + v_2 / \sigma_{v_2}$$

Gibbs Sampler: delta draw $\delta \mid \beta, \Sigma, x, y, z$

$$x = \delta z + \varepsilon_1$$

$$y = \beta(\delta z + \varepsilon_1) + \varepsilon_2 \text{ or } y = \delta(\beta z) + (\beta\varepsilon_1 + \varepsilon_2)$$

$$v = \begin{pmatrix} \varepsilon_1 \\ \beta\varepsilon_1 + \varepsilon_2 \end{pmatrix}; \; Var(v) = A\Sigma A' = \Omega = LL'$$

$$v = Lu; \quad Var(u) = I_2$$

Standardize the two equations and we have a restricted MRM (estimate by "doubling" the rows):

$$L^{-1}\begin{pmatrix} x \\ y \end{pmatrix} = L^{-1}\begin{bmatrix} z \\ \beta z \end{bmatrix}\delta + u$$

Identification Problems – "Weak" instruments

Suppose $\delta = 0$.

$$x = \varepsilon_1$$

$$y = \beta\varepsilon_1 + \varepsilon_2$$

$$\operatorname{cov}(x, y) = \beta\sigma_{11} + \sigma_{12}$$

or

$$\frac{\operatorname{cov}(x, y)}{\sigma_{11}} = \beta + \frac{\sigma_{12}}{\sigma_{11}}$$

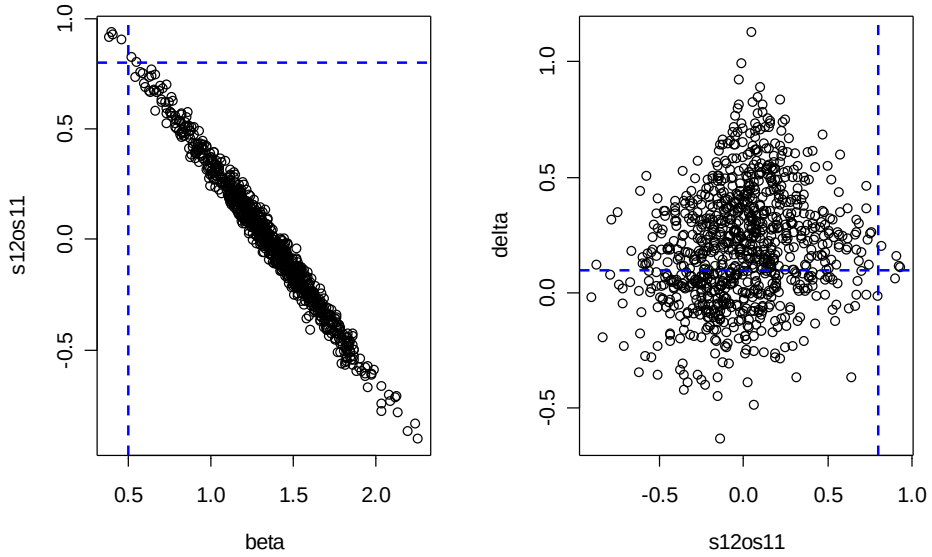
$\Rightarrow \delta$ small, trouble!

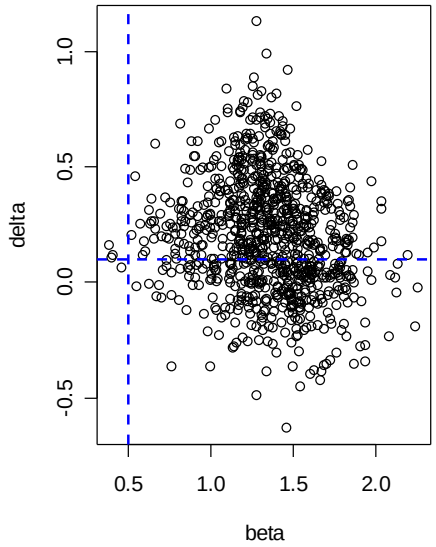
Weak Ins Ex

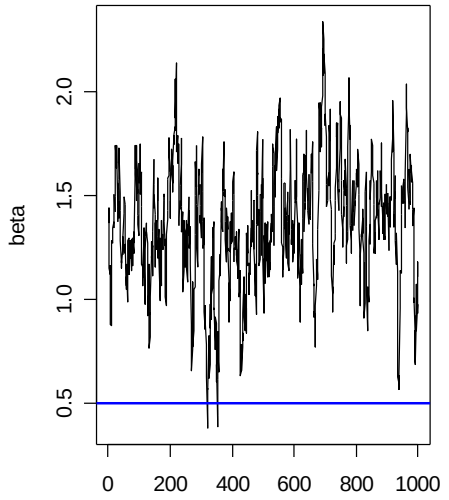
VERY weak (Rsq=0.01)

(rel num eff = 10)

Influence of very diffuse but proper prior on Σ -- shrinks corrs to 0.



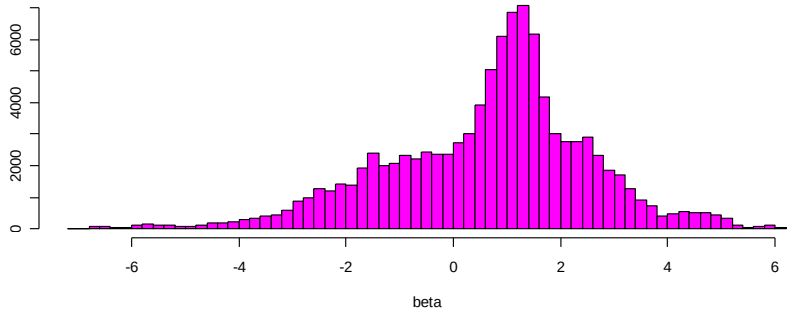


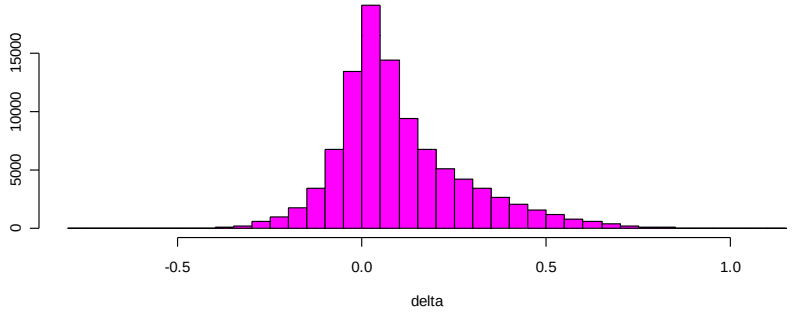


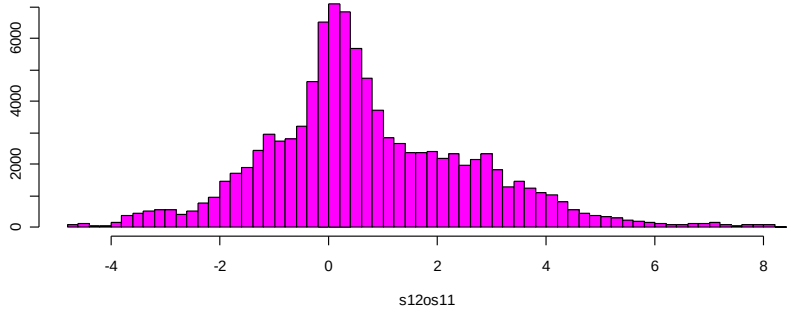
Weak Ins Ex

Posteriors based on
100,000 draws

Inadequacy of
Standard Normal
Asymptotic
Approximations!







Mixtures of Normal for Errors

Consider the instrumental variables model with mixture of normal errors with K components:

$$x = \delta z + \varepsilon_1'$$

$$y = \beta x + \varepsilon_2'$$

$$\begin{pmatrix} \varepsilon_1' \\ \varepsilon_2' \end{pmatrix} \sim N(\mu_{ind}, \Sigma_{ind})$$

$$ind \sim multinomial(p)$$

note: $\mathrm{E}\left[\varepsilon' \middle| ind\right] = \mu_{ind}$ and $E\left[\mathrm{E}\left[\varepsilon' \middle| ind\right]\right] = \sum_{k=1}^{K} p_k \mu_k$

A Gibbs Sampler

(1) $\beta \left| \delta, ind, \left\{ \mu_{k,} \Sigma_k \right\}, x, y, z \right.$

(2) $\delta \left| \beta, ind, \left\{ \mu_{k,} \Sigma_k \right\}, x, y, z \right.$

(3) $ind, \left\{ \mu_{k,} \Sigma_k \right\} \left| \delta, \beta, x, y, z \right.$

Tricks:

Need to deal with fact that errors have non-zero mean

Cluster observations according to ind draw and standardize using appropriate comp parameters.

Fat-tailed Example

Standard outlier model:

$$p' = (.95, .05)$$

$$\text{comp } 1: \ \varepsilon' \sim N(0, \Sigma_1) \quad \Sigma_1 = \begin{bmatrix} 1 & .8 \\ .8 & 1 \end{bmatrix}$$

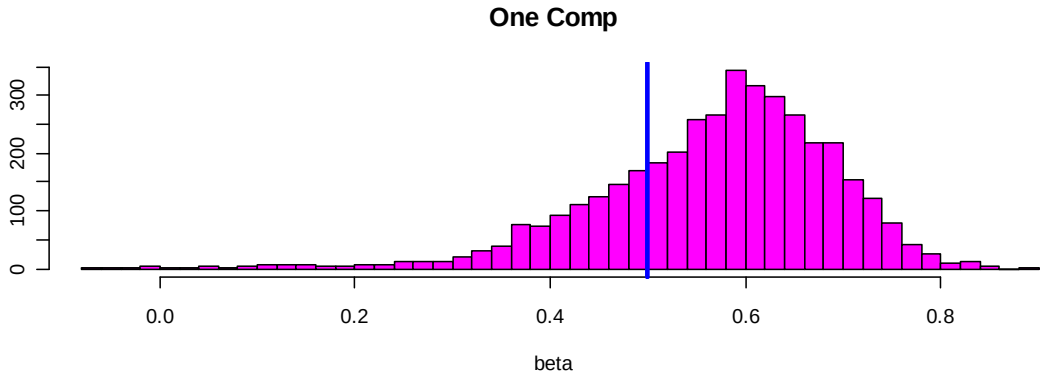
$$\text{comp } 2: \ \varepsilon' \sim N(0, \Sigma_2) \quad \Sigma_2 = M\Sigma_1$$

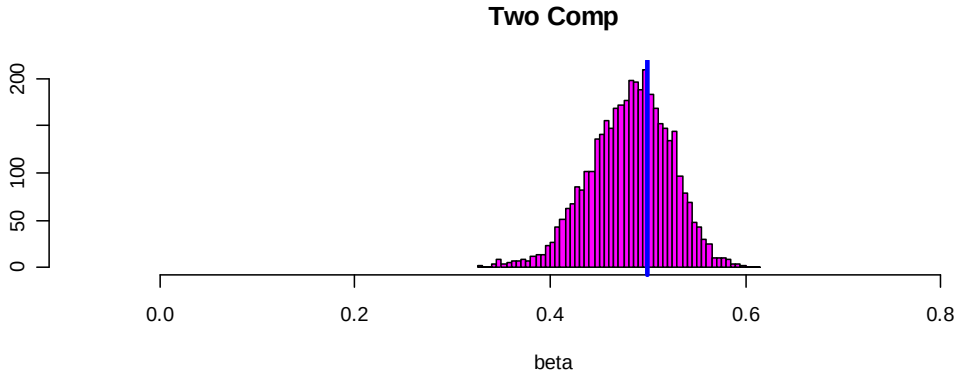
$$M >>> Var(\delta z)$$

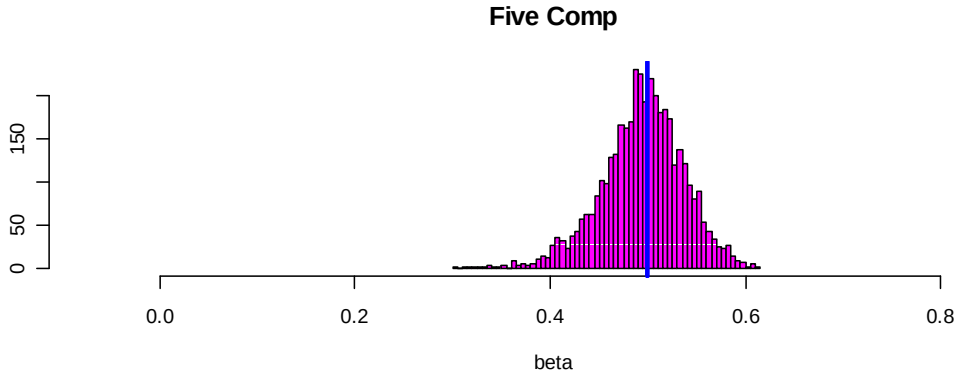
What if you specify thin tails (one comp)?

Fat Tails

$$\Sigma_2 = 200\Sigma_1$$







Number of Components

If I only use 2 components, I am cheating!

One practical approach, specify a relative large number of components, use proper priors.

What happens in these examples?

Can we make number of components dependent on data?

Dirichlet Process Model: Two Interpretations

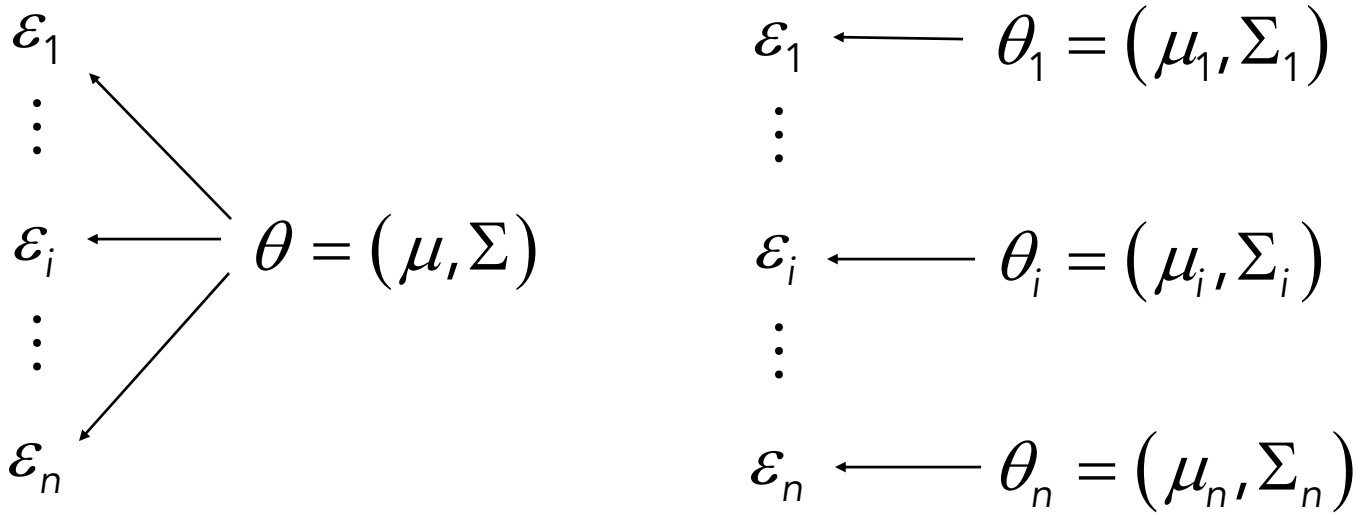
1). DP model is very much the same as a mixture of normals except we allow new components to be "born" and old components to "die" in our exploration of the posterior.

2). DP model is a generalization of a hierarchical model with a shrinkage prior that creates dependence or "clumping" of observations into groups, each with their own base distribution.

Outline of DP Approach

How can we make the error distribution flexible?

Start from the normal base, but allow each error to have it's own set of parms:



Outline of DP Approach

This is a very flexible model that accomodates: non-normality via mixing and a general form of heteroskedasticity.

However, it is not practical without a prior specification that ties the $\{\theta_i\}$ together.

We need shrinkage or some sort of dependent prior to deal with proliferation of parameters (we can't literally have n independent sets of parameters).

Two ways: 1. make them correlated 2. "clump" them together by restricting to I* unique values.

Outline of DP Approach

Consider generic hierarchical situation:

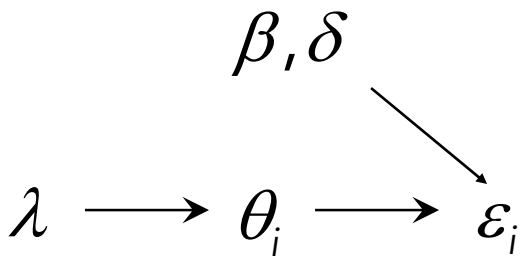
$\varepsilon_i \,|\, \theta_i, \beta, \delta$

$\theta_i \,|\, \lambda \sim G_0$

ε (errors) are conditionally independent, e.g. normal with $\theta_i = (\mu_i, \Sigma_i)$

One component normal model: $\theta_i = (\mu, \Sigma)$

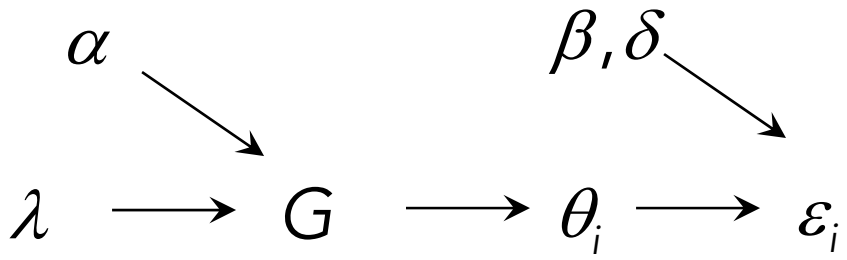
DAG:



DP prior

Add another layer to hierarchy – DP prior for theta

DAG:

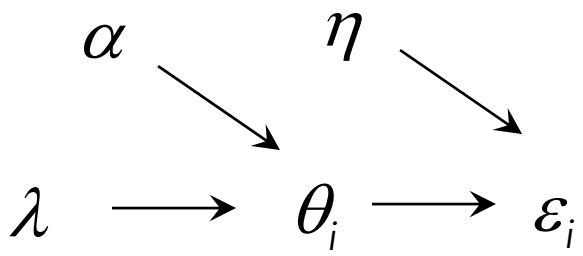


G is a Dirichlet Process – a distribution over other distributions. Each draw of G is a Dirichlet Distribution. G is centered on G_0 with tightness parameter α

DPM

Collapse the DAG by integrating out G

DAG:



$\{\theta_1, \ldots, \theta_n\}$ are now dependent with a mixture of DP distribution. Note: this distribution is not discrete unlike the DP. Puts positive probability on continuous distributions.

DPM: Sticking Breaking Representation

Each draw from G is a discrete distribution on a countable number of values of θ_j. The support points are iid draws from G_0.

The probability masses for this distribution are given by:

$$\pi_k = \omega_k \prod_{j=1}^{k-1} \left(1 - \omega_j\right)$$

$$\omega_0 = 0; \quad \omega_k \sim Beta\left(1, \alpha\right)$$

Declining pattern of weights. Rate of decline depends on α.

DPM: A Priori Distribution of Cluster Sizes

The "stick-breaking" representation reveals a possible limitation of the DPM prior. The weights decline rapidly.

Implies a prior view of a few large clusters.

$$\Pr\left[\theta_j = \theta_i^* \,\middle|\, \theta_1, \ldots, \theta_{j-1}, \alpha\right] = \frac{n_{i*}}{\alpha + j - 1}$$

n_{i*} is the size of the cluster with unique value θ_i^* among the collection, θ_1 , …, θ_{j-1}. The set of unique parameter values is $\theta_1^*, \ldots, \theta_{l*}^*$

DPM: Drawing from Posterior

Basis for a Gibbs Sampler:

$$\theta_j \left| \varepsilon, \theta_{-j} = \theta_j \right| \varepsilon_j, \theta_{-j}$$

Why? Conditional Independence!

This is a simple update:

There are "n" models for θ_j each of the other values of theta and the base prior. This is very much like mixture of normals draw of indicators.

DPM: Drawing from Posterior

n models and prior probs:

δ_i *with prior prob* $\dfrac{1}{\alpha + (n-1)}$ one of others

$G_0(\lambda)$ *with prior prob* $\dfrac{\alpha}{\alpha + (n-1)}$ "birth"

$$\theta_j \left| \theta_{-j}, \varepsilon_j, \lambda, \alpha \sim \begin{cases} q_0 & \theta_j \left| \varepsilon_j, G_0(\lambda) \right. \\ q_i & \delta_i \quad i \neq j \end{cases} \right.$$

DPM: Drawing from Posterior

$$q_0 = p\left(M_0 \middle| \varepsilon_j\right) = \int p\left(\varepsilon_j \middle| \theta_j\right) p\left(\theta_j | \lambda\right) d\theta_j \times p\left(M_0\right)$$

$$= \int p\left(\varepsilon_j \middle| \theta_j\right) G_0\left(\theta_j | \lambda\right) d\theta_j \times \frac{\alpha}{\alpha + (n-1)}$$

$$q_i = p\left(M_i \middle| \varepsilon_j\right) = p\left(\varepsilon_j \middle| \theta_i\right) \times \frac{1}{\alpha + (n-1)}$$

Note: q need to be normalized! Conjugate priors can help to compute q_0.

Assessing the DP prior

Two Aspects of Prior:

α-- influences the number of unique values of θ

G_0, λ -- govern distribution of proposed values of θ

e.g.

I can approximate a distribution with a large number of "small" normal components or a smaller number of "big" components.

Assessing the DP prior: choice of α

There is a relationship between α and the number of distinct theta values (viz number of normal components). Antoniak (74) gives this from MDP.

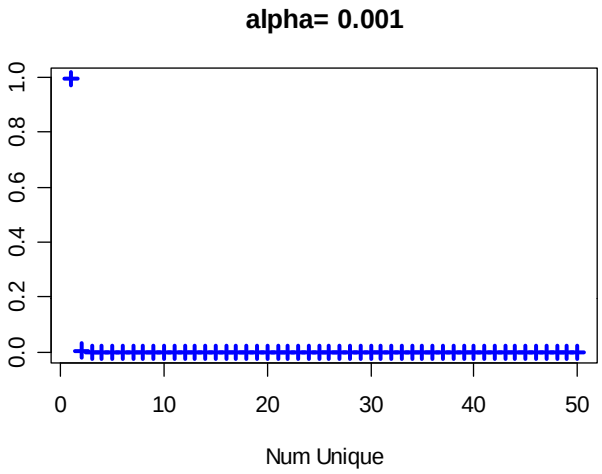
$$\Pr\left(l^* = k\right) = \left\|S_n^{(k)}\right\| \alpha^k \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)}$$

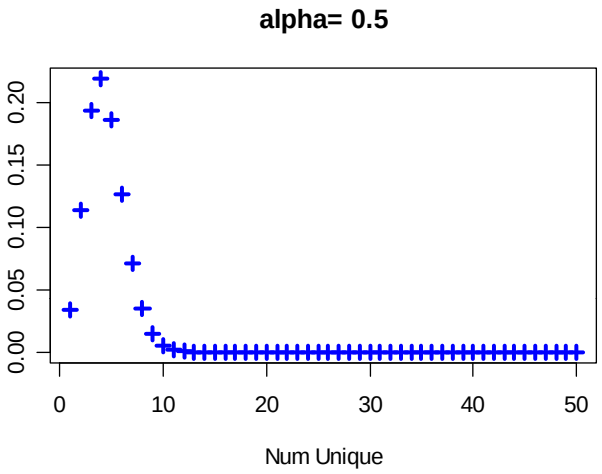
S are "Stirling numbers of First Kind." Note: S cannot be computed using standard recurrence relationship for n > 150 without overflow!

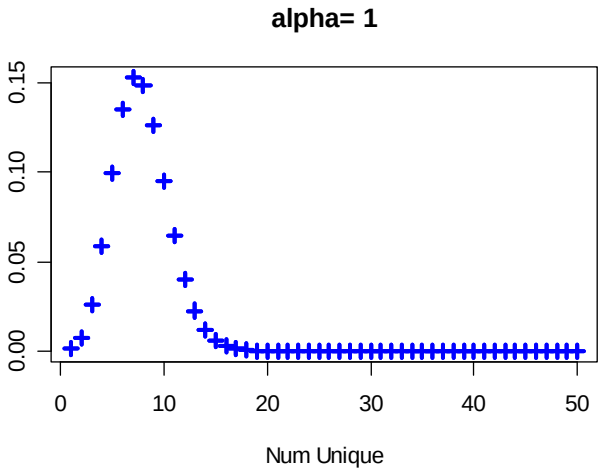
$$S_n^{(k)} \doteq \frac{\Gamma(n)}{\Gamma(k)} \left(\gamma + \ln(n)\right)^{k-1}$$

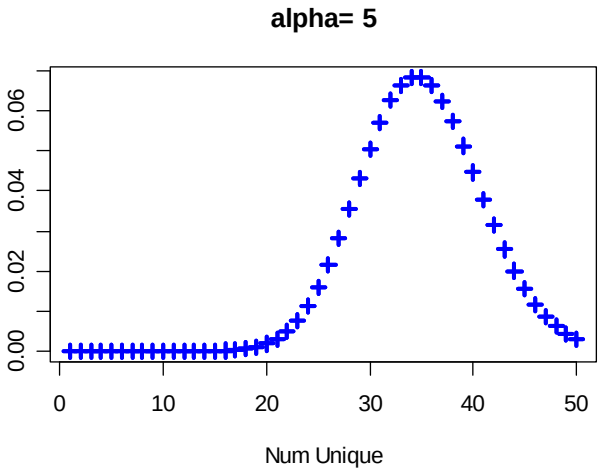
Assessing the DP prior: choice of α

For N=500









Assessing the DP prior: Priors on α

Fixing may not be reasonable. Prior on number of unique theta may be too tight.

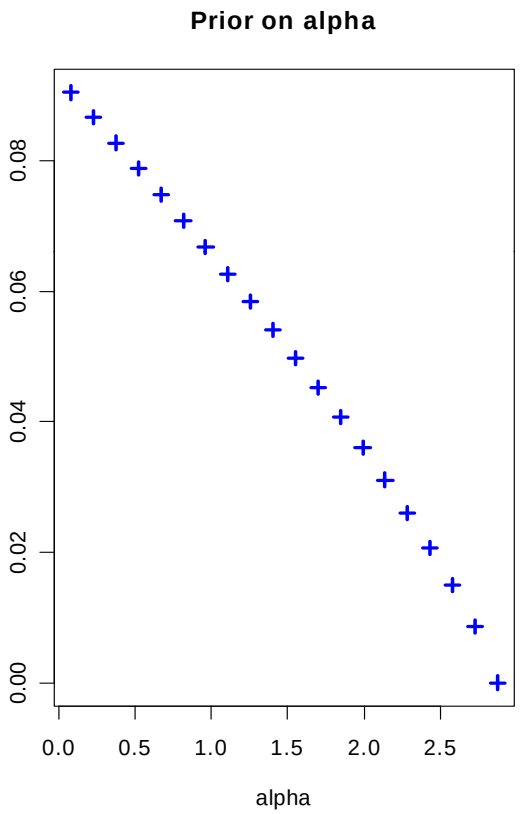
"Solution:" put a prior on alpha.

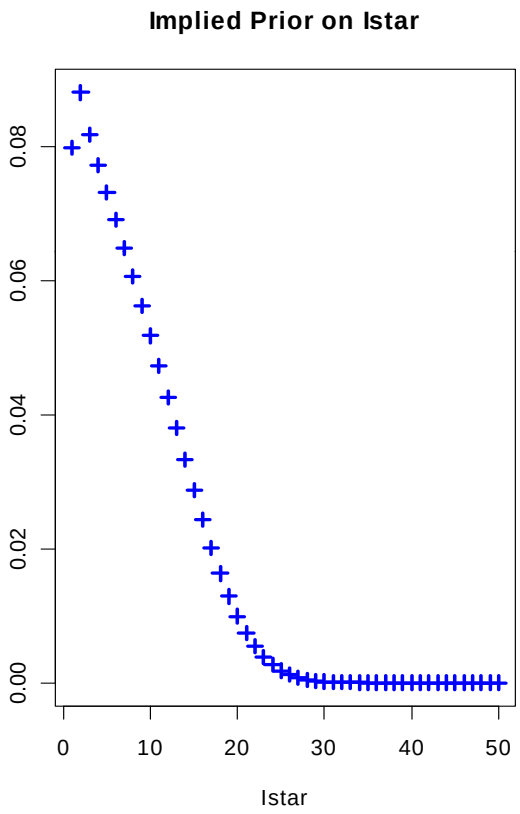
Assess prior by examining the priori distribution of number of unique theta.

$$p(I^*) = \int p(I^* | \alpha) p(\alpha) d\alpha$$

$$p(\alpha) \propto \left(1 - \frac{(\alpha - \underline{\alpha})}{(\bar{\alpha} - \underline{\alpha})} \right)^{\phi}$$

Assessing the DP prior: Priors on α





Assessing the DP prior: Choice of c

$$q_0 = p\left(M_0 \middle| \varepsilon_j\right) = \int p\left(\varepsilon_j \middle| \theta_j\right) G_0\left(\theta_j | \lambda\right) d\theta_j \times \frac{\alpha}{\alpha + (n-1)}$$

Both α and c determine the probability of a "birth."

Intuition:

1. Very diffuse settings of c reduce model prob.
2. Tight priors centered away from y will also reduce model prob.

Must choose reasonable values. Shouldn't be very sensitive to this choice.

Assessing the DP prior: Choice of α

$$G_0 : \mu \sim N\left(\bar{\mu}, a^{-1}\Sigma\right); \Sigma \sim IW\left(\upsilon, V\right)$$

Choice of λ made easier if we center and scale both y and x by the std deviation. Then we know much of mass ε distribution should lie in [-2,2] x [-2,2].

Set $V = vI_2$ and $\bar{\mu} = 0$

We need assess υ, v, a with the goal of spreading components across the support of the errors.

Assessing the DP prior: Choice of α

Look at marginals of μ and σ_1

Choose (υ, v, a)

э

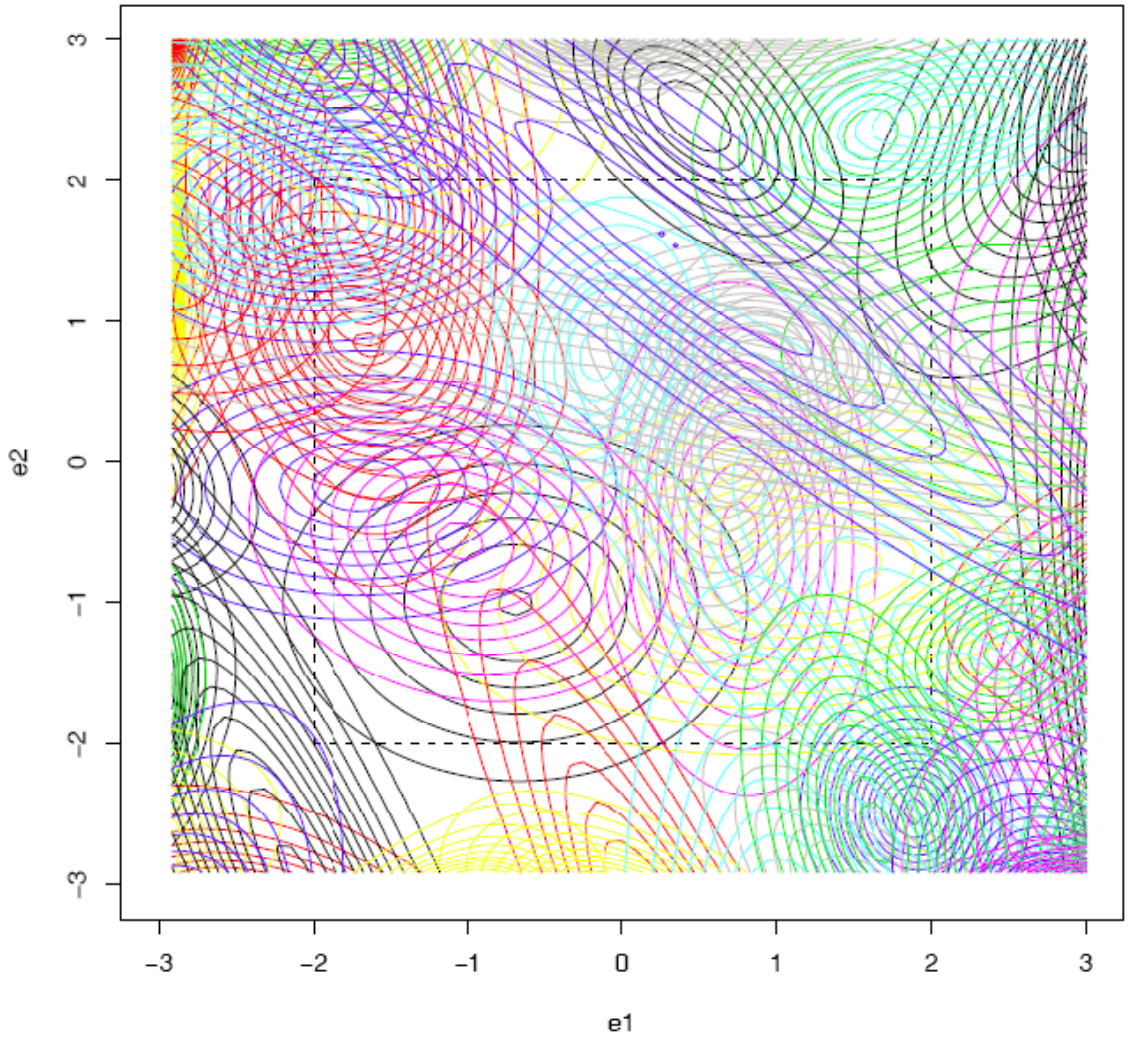
$\Pr[.25 < \sigma_1 < 3.25] = .8$

$\Pr[-10 < \mu < 10] = .8$

$\Rightarrow \upsilon = 2.004,\ v = .17,\ a = .016$

Very Diffuse!

Draws from G_0



Gibbs Sampler for DP in the IV Model

$$\beta \mid \delta, \{\theta_i\}, x, y, z$$

$$\delta \mid \beta, \{\theta_i\}, x, y, z$$

Same as for Normal Mixture Model

$$\{\theta_i\} \mid \delta, \beta, x, y, z$$

Doesn't Vectorize

$$\{\theta_i^*\} \mid ind, \delta, \beta, x, y, z$$

"Remix" Step

$$\alpha \mid I^*$$

Trivial (discrete)

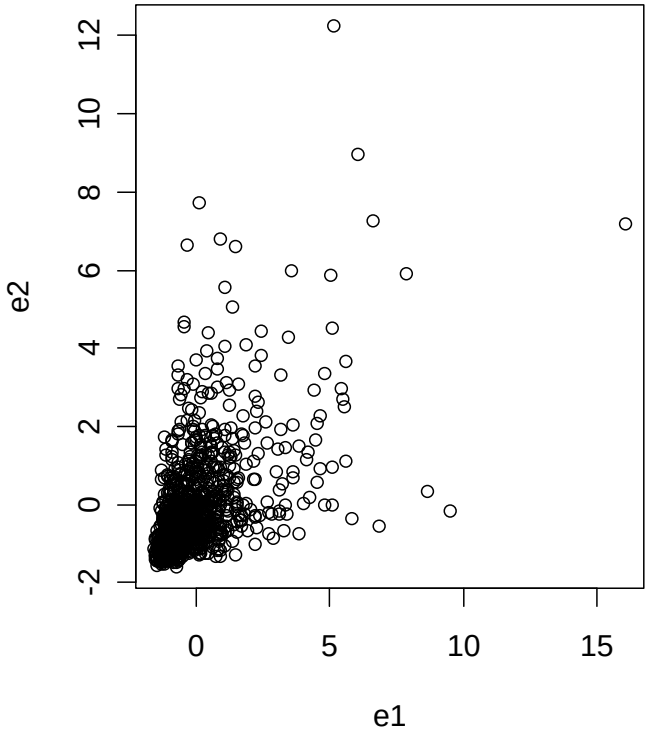
q computations and conjugate draws are can be vectorized (if computed in advance for unique set of thetas).

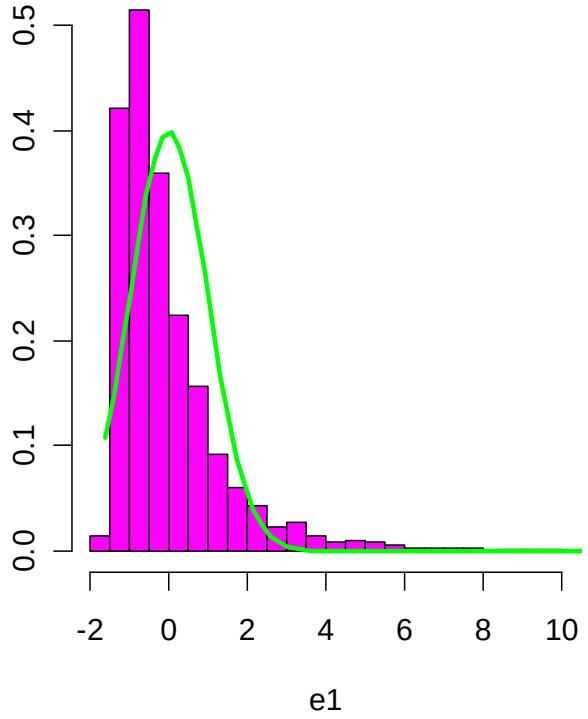
Sampling Experiments

1. How well do DP models accommodate departures to normality?
2. How useful are the DP Bayes results for those interested in “standard” inferences such as confidence intervals?
3. How do conditions of many instruments or weak instruments affect performance?

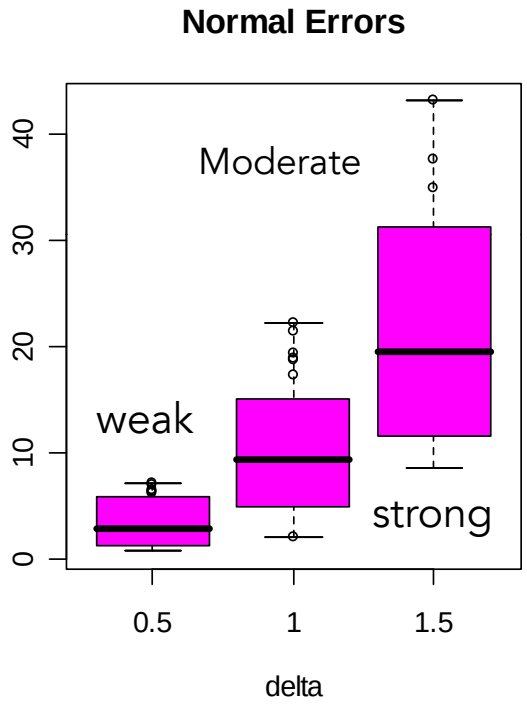
Sampling Experiments – Choice of Non-normal Alternatives

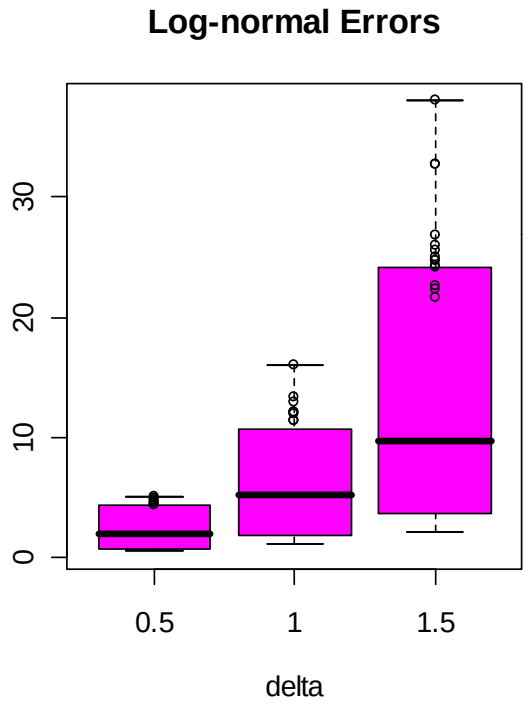
Let's start with skewed distributions. Use a translated log-normal. Scale by inter-quartile range.





Sampling Experiments – Strength of Instruments- F stats





k=10

Weak case is bounded away from zero. Our simulated datasets have information!

Sampling Experiments- Alternative Procedures

Classical Econometrician: "We are interested in inference. We are not interested in a better point estimator."

Standard asymptotics for various K-class estimators

"Many" instruments asymptotics (bound F as k, N increase)

"Weak" instrument asymptotics (bound F and fix k as N increases) Kleibergen (K), Modified Kleibergen (J), and Conditional Likelihood Ratio(CLR) (Andrews et al 06).

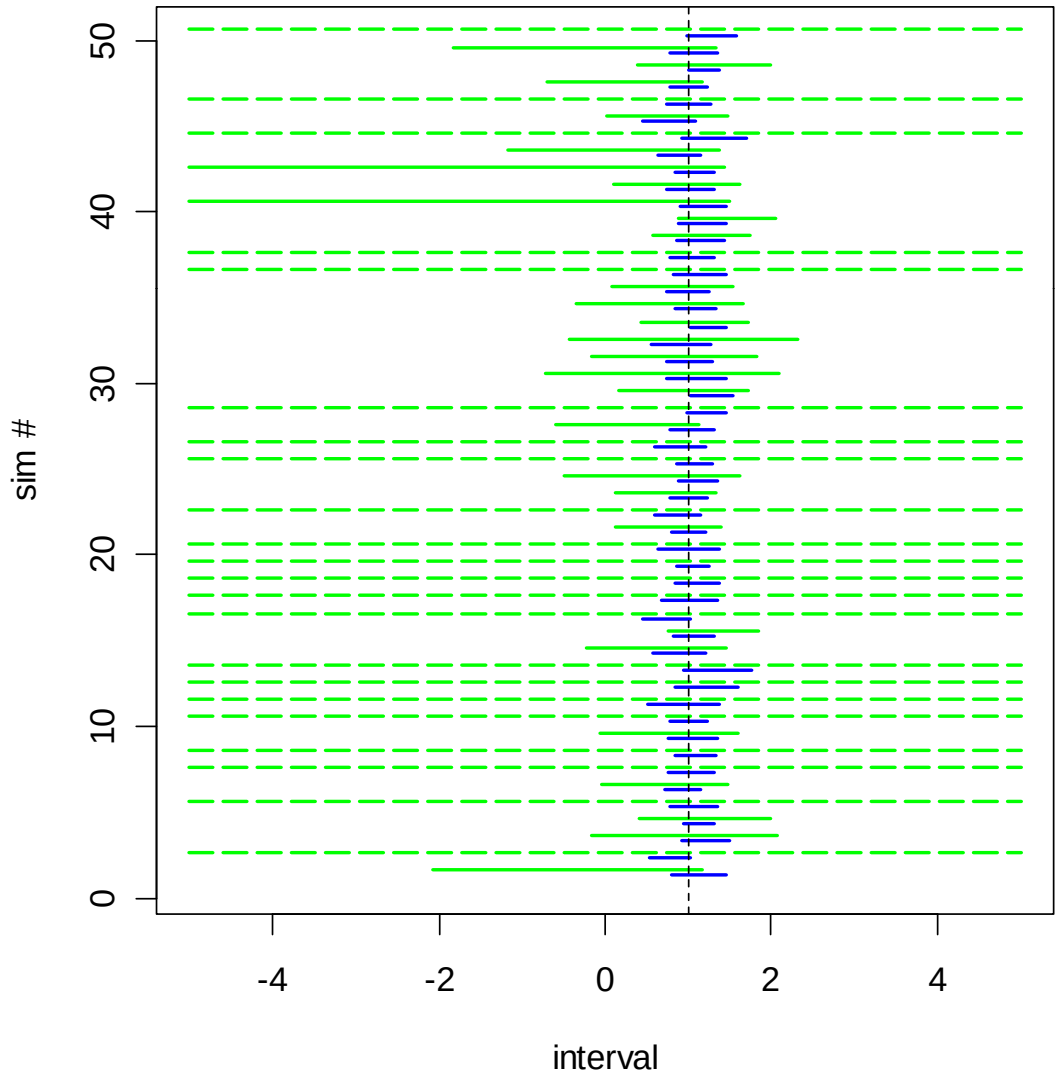
Sampling Experiments- Coverage of "95%" Intervals

N=100; based on 400 reps

	error dist	BayesDP	TSLS-STD	Fuller-Many	CLR
weak					
	Normal	0.83	0.75	0.93	0.92
	LogNormal	0.91	0.69	0.92	0.96
strong					
	Normal	0.92	0.92	0.95	0.94
	LogNormal	0.96	0.90	0.96	0.95

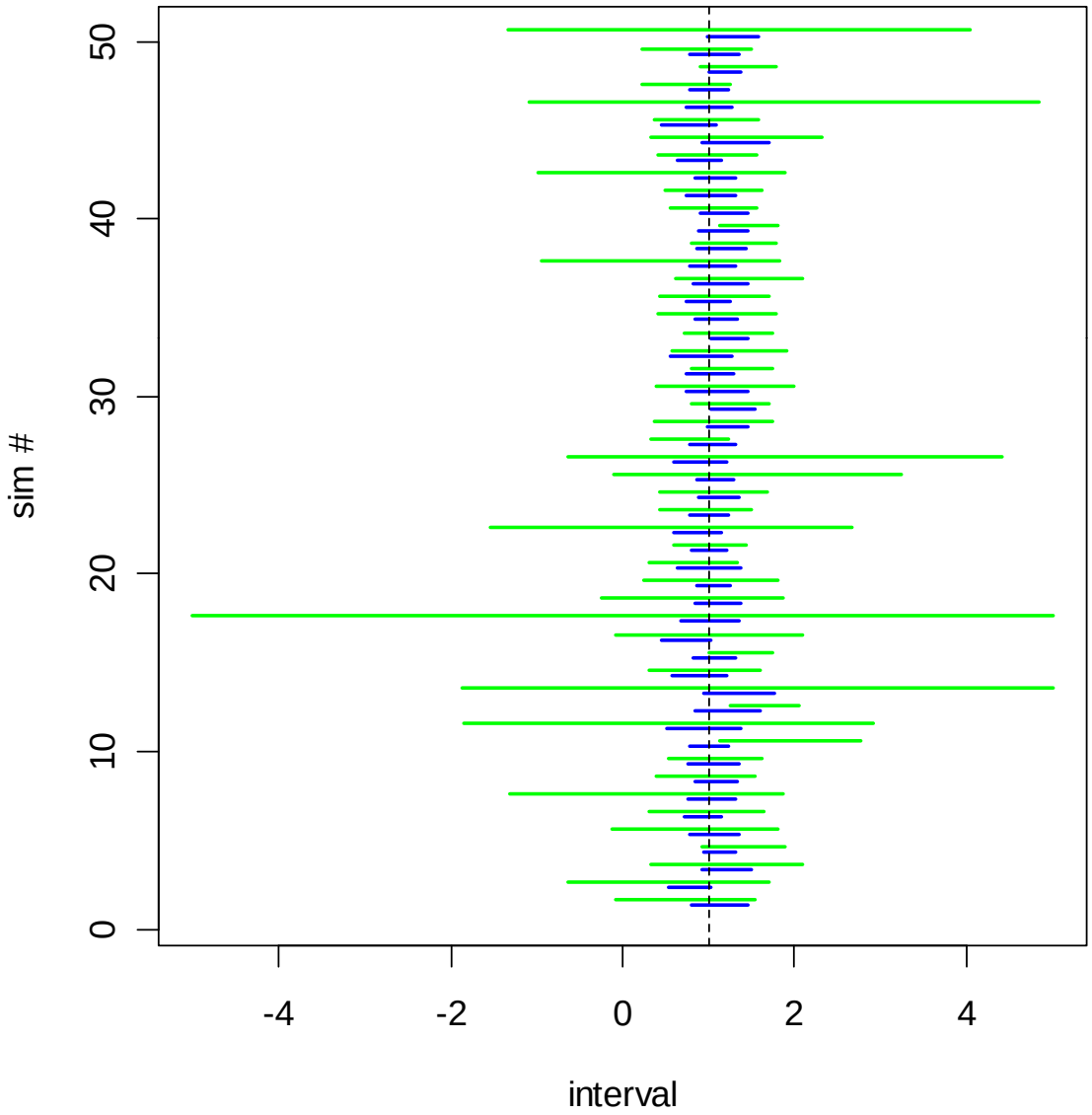
7% (normal) | 42 % (log-normal) are infinite length

Bayes Vs. CLR (Andrews 06)



Weak Instruments Log-Normal Errors

Bayes Vs. Fuller-Many



Weak Instruments Log-Normal Errors

A Metric for Interval Performance

Bayes Intervals don't "blow-up" – theoretically some should. However, it is not the case that > 30 percent of reps have no information!

Smaller and located closer to the true beta.

Scalar measure:

$$X \sim Unif(L, U)$$

$$E\left[|X - \beta|\right] = \int_L^U |x - \beta| \frac{1}{U - L} dx$$

Interval Performance

	error dist	BayesDP	TSLS-STD	Fuller-Many	CLR-Weak
weak					
	Normal	0.26	0.27	0.35	0.75
	LogNormal	0.18	0.37	0.61	1.58
strong					
	Normal	0.09	0.09	0.09	0.10
	LogNormal	0.07	0.14	0.14	0.16

Estimation Performance - RMSE

	Error Dist	Estimator			
		BayesNP	BayesDP	TSLS	F1
weak					
	Normal	0.24	0.24	0.26	0.29
	LogNormal	0.35	0.16	0.37	0.43
strong					
	Normal	0.07	0.07	0.08	0.08
	LogNormal	0.11	0.05	0.12	0.12

Estimation Performance - Bias

		BayesNP	BayesDP	TSLS	F1
weak					
	Normal	0.17	0.18	0.20	0.05
	LogNormal	0.26	0.09	0.28	0.09
strong					
	Normal	0.02	0.02	0.02	0.00
	LogNormal	0.04	0.01	0.06	0.01

An Example: Card Data

y is log wage.

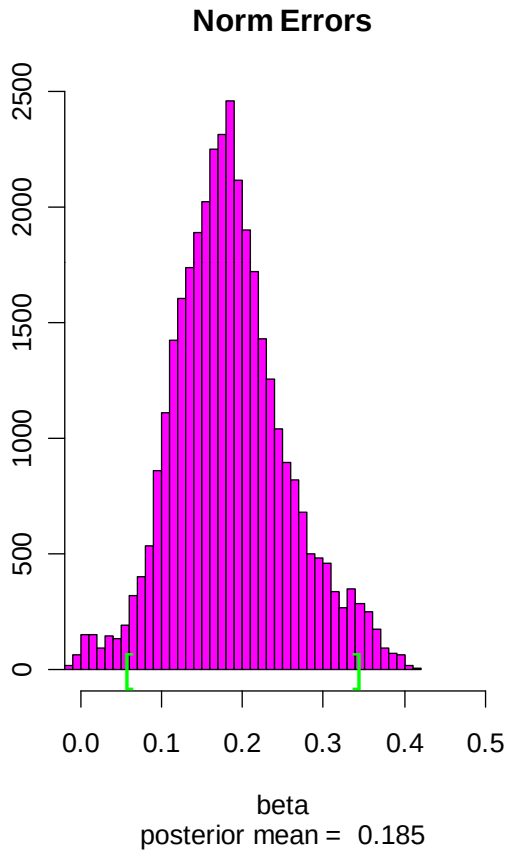
x education (yrs)

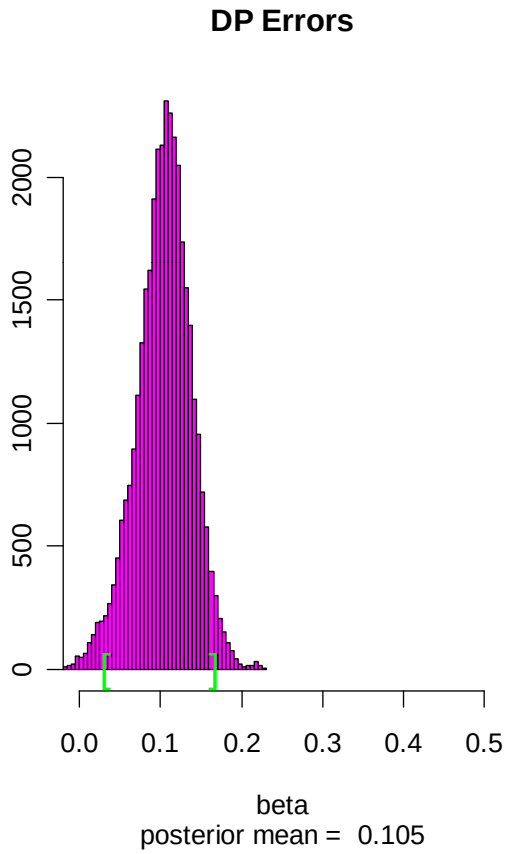
z is proximity to 2 and 4 year colleges

N=3010.

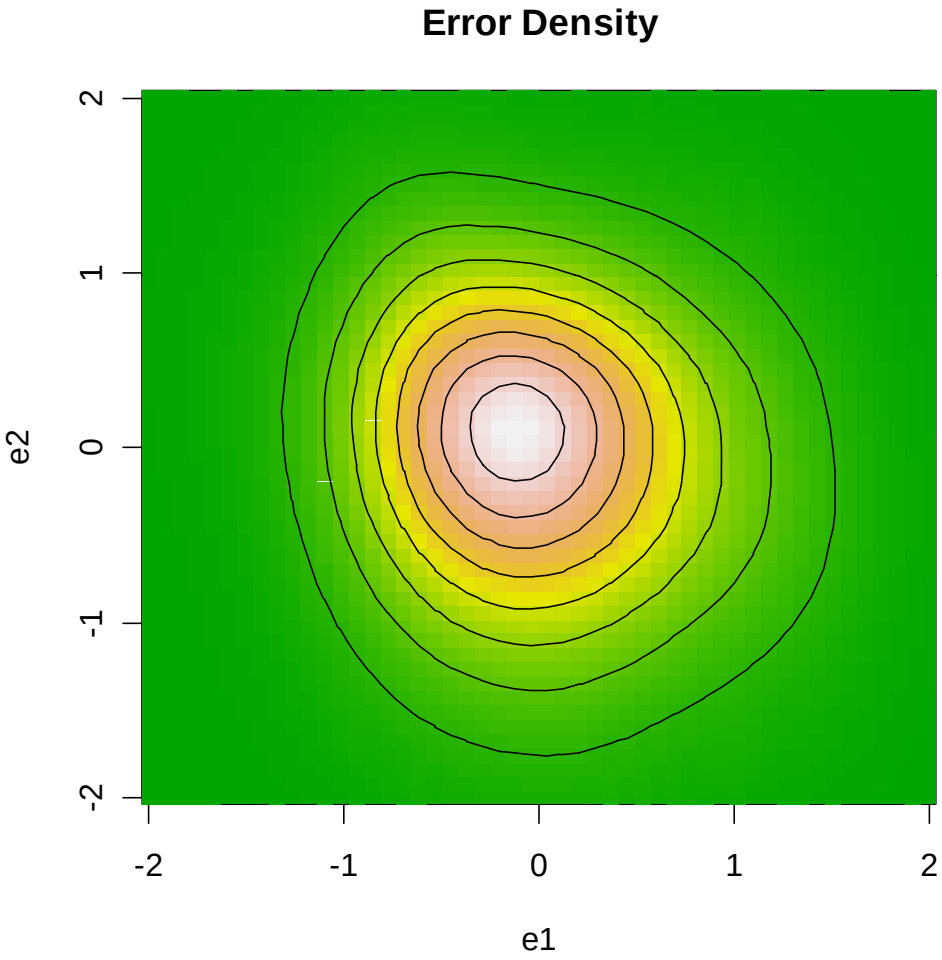
Evidence from standard models is a negative correlation between errors (contrary to the old ability omitted variable interpretation).

An Example: Card Data





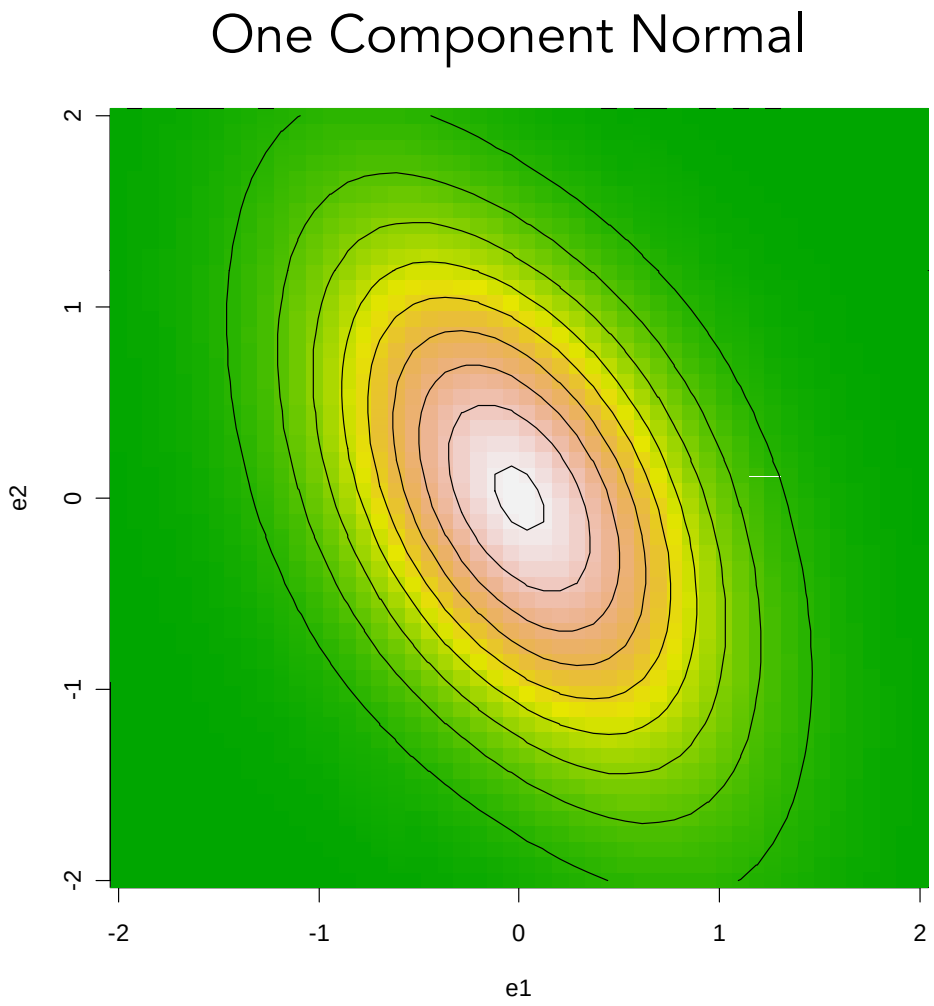
An Example: Card Data



Non-normal and low dependence.

Implies "normal" error model results may be driven by small fraction of data.

An Example: Card Data



One-component model is "fooled" into believing there is a lot "endogeneity"

Summary Bayes DP IV

BayesDP IV works well under the rules of the classical instruments literature game.

BayesDP strictly dominates BayesNP

Do you want much shorter intervals (more efficient use of sample information) at the expense of somewhat lower coverage in very weak instrument case?

"Plausibly Exogenous" Instruments

Many IV analyses use instruments that are can be viewed as "approximately" exogenous but that we might argue are not "strictly" exogeneous – e.g. wholesale prices, prices in other markets, other characteristics …

Yet these analyses impose *strict* exogeneity in estimation. Careful workers use other informal methods for assessing exogeneity such as regressing instruments on observables …

Can we help?

"Plausibly Exogenous" Instruments

Our goal: provide operational definition of "plausible" or approximate exogeneity

$$Y = X\beta + Z\gamma + \varepsilon$$
$$X = Z\Pi + V$$

γ is an unidentified parameter – models the relationship between instruments Z and structural error.

γ is a measure of the "direct" effect of instruments

"Plausibly Exogenous" Instruments

Standard approach (dogmatic prior): $\gamma = 0$

Our approach:

Put a prior on $\mathtt{j}$.

Sources of Prior information:

"direct" effect of instruments (e.g. A&K direct effect of qtr of birth)

prior beliefs that j is small relative to e

Answers the question: "How bad do the instruments need to be before key results change?"

"Plausibly Exogenous" Instruments

Given some prior information on γ, how should we conduct inference?

Full Bayes (using straightforward extensions of what we have done for strict exogeneity case.

Approximate Bayes:

prior-weighted frequentist intervals

intervals constructed via a "local" form of asymptotic experiment

Details: Conley, Hansen, Rossi, "Plausibly Exogenous" SSRN

"Plausibly Exogenous" Examples

Aggregate Share Model for Margarine Demand
(inspired by Chintagunta, Dube, Goh, 2003)

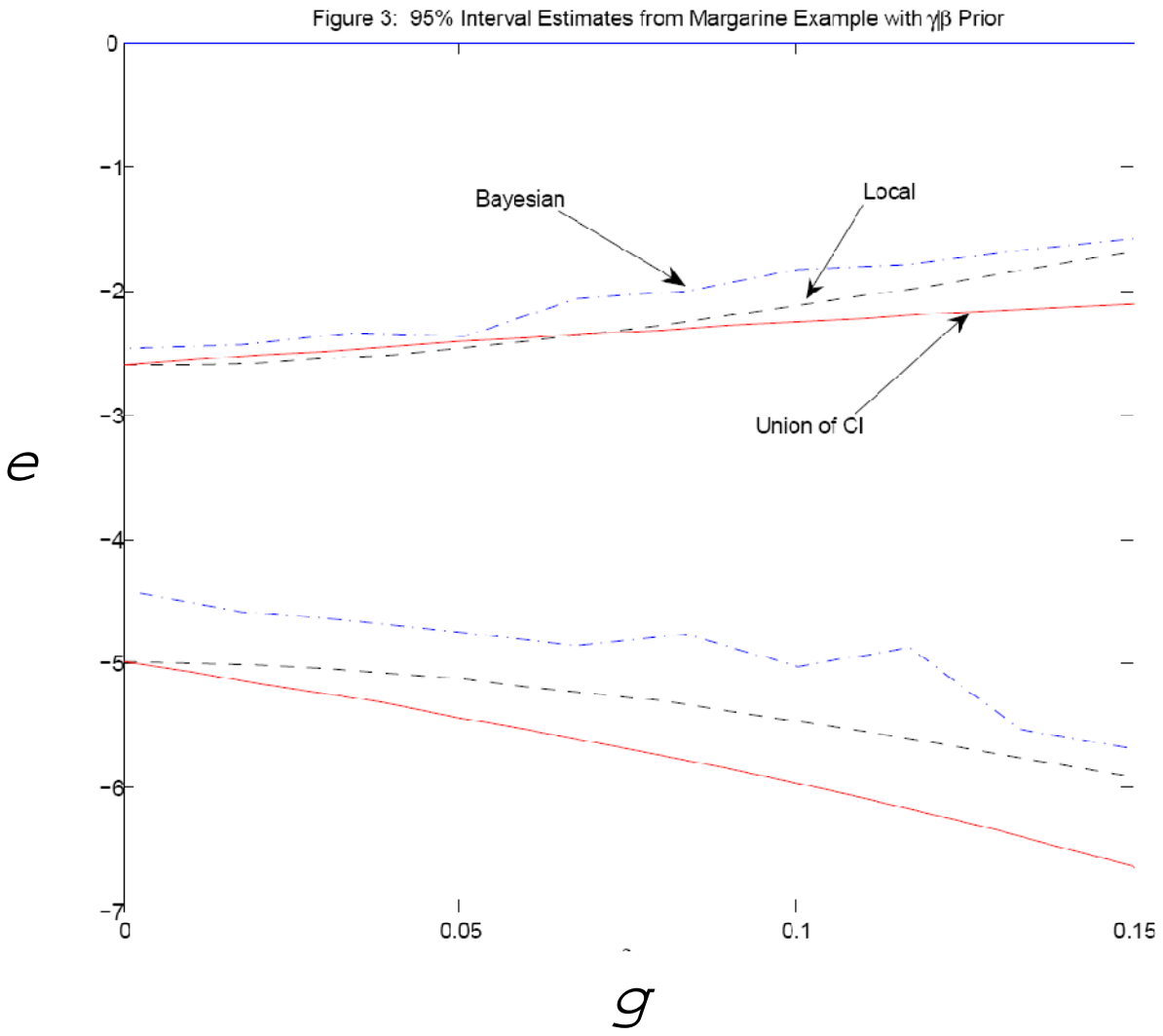
$$\log(share) = \beta \log(retail\,price) + X\lambda + Z\gamma + u$$

Wholesale price is "plausibly exogeneous" driven more by cost shocks then manufacturer-sponsored "demand" shocks.

Prior: direct effects of wholesale price less than price elasticity,

$$\gamma \,|\, \beta \sim N(0, \delta^2 \beta^2)$$

"Plausibly Exogenous" Examples



Small Sample: 117

intervals are insensitive to a large range of g values

"Plausibly Exogenous" Examples

Angrist and Krueger (1991)

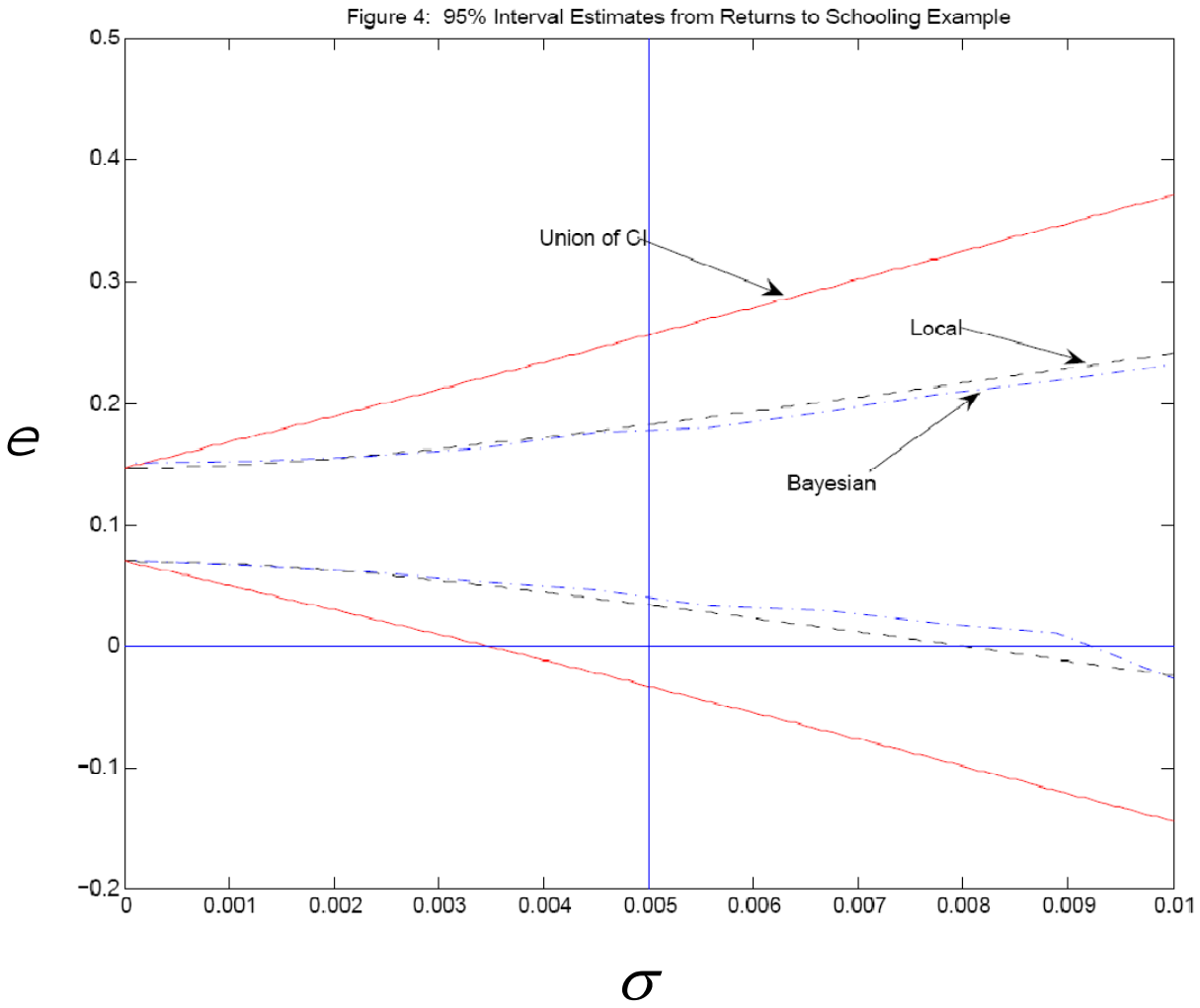
$$\log(wage) = \beta \log(school) + X\lambda + Z\gamma + u$$

Instruments: quarter of birth

Controls: year of birth/state dummies

Bound, Jaeger and Baker argue Q of B is not exogeneous and that direct effect could be of the order of 1 per cent. This motivates our choice of prior: $\gamma \sim N\left(0, \sigma^2 = (.005)^2\right)$

"Plausibly Exogenous" Examples



Large Sample: 329,509

Intervals are sensitive to the assumption of exogeneity

Conclusions

A "true" Bayesian IV approach is possible. Works well relative to "state of the art" frequentist methods

Prior information is important and prior sensitivity analysis is an excellent way to measure sample information