Bayesian Instruments Via Dirichlet Process Priors

Peter Rossi

GSB/U of Chicago

joint with Rob McCulloch, Tim Conley, and Chris Hansen

Motivation

IV problems are often done with a “small” amount of sample information (weak ins).

It would seem natural to apply a small amount of prior information, e.g. the returns to education are unlikely to be outside of (.01, .2).

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?

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!

The Linear Case

Linear Structural equations are central in applied work.

99-04, QJE/AER/JPE had 129 articles with Linear model, 89 with only one endog RHS var! This is a *relevant* and simple ex:

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

$$(2)\ \ y = \beta x + \varepsilon_2$$

(1) is a regression equation. $\varepsilon_1 | z \sim \varepsilon_1$

(2) is not !! $\text{cov}(x, \varepsilon_2) \neq 0 \Rightarrow \varepsilon_2 | x \nsim \varepsilon_2$

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

Derive the joint distribution of y, x | z.

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

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

or

$$(1)\ \ x=\pi_x z+v_1 \qquad \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \sim N\left(0,\Omega\right)$$

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

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

Identification Problems – “Weak” instruments

Suppose $\delta = .0$

$$x = \varepsilon_1$$

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

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

or

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

$\Rightarrow \delta$ small, trouble!

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

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

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

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

Tricks (`rivGibbs` in *bayesm*):

(1) given δ, convert structural equation into standard
Bayes regression. We “observe” Compute ε_1
. $\varepsilon_2 | \varepsilon_1$

(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 $\quad \delta \middle| \beta, \Sigma, x, y, z$

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

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

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

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

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

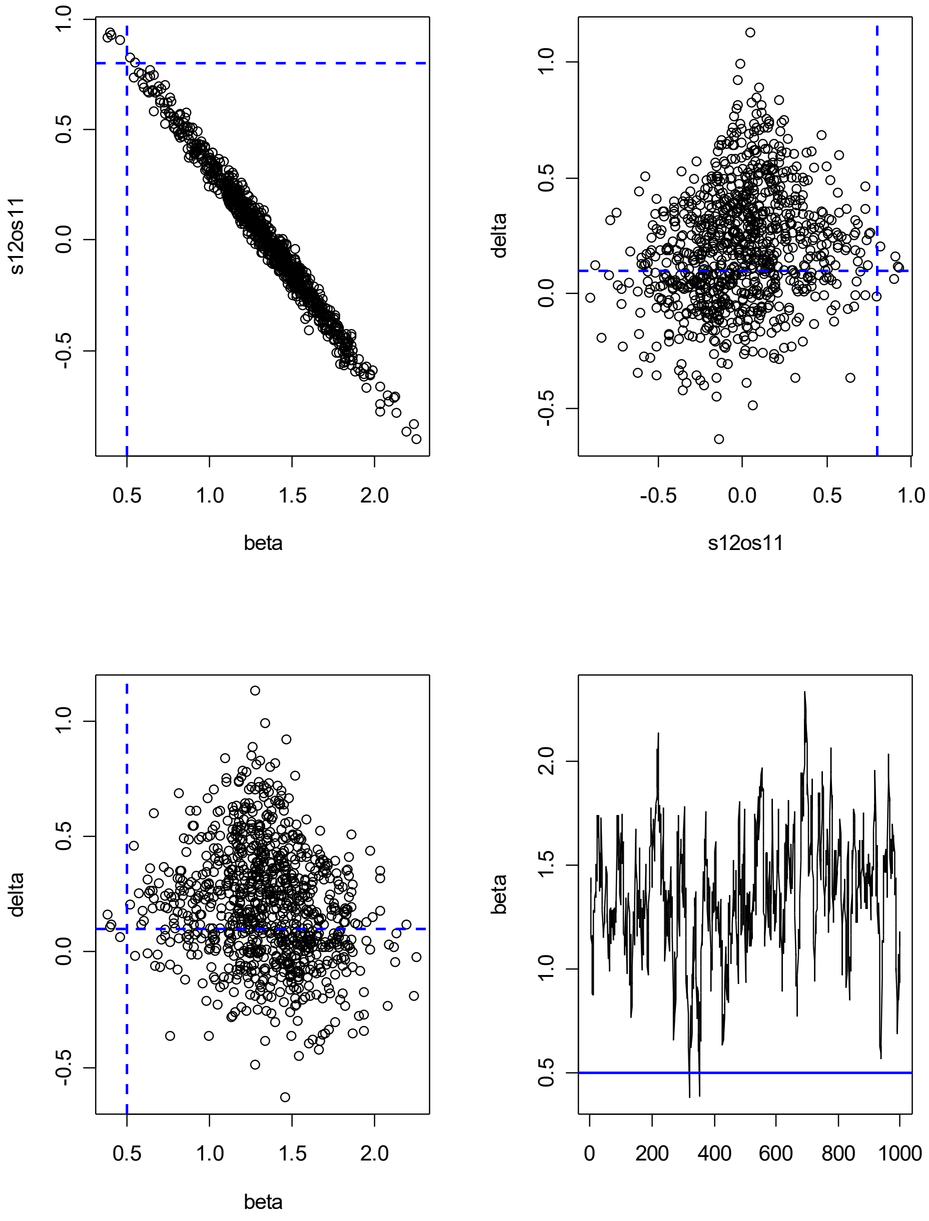
$$L^{-1}\begin{pmatrix} x \\ y \end{pmatrix} = L^{-1}\begin{bmatrix} z \\ \beta z \end{bmatrix}\delta + u$$

Weak Ins Ex

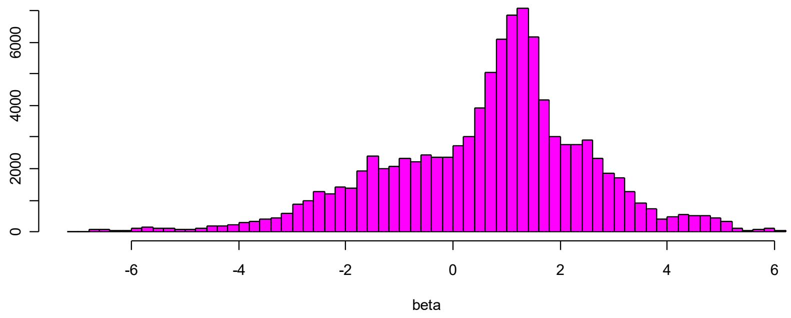
VERY weak (Rsq=0.01)

(rel num eff = 10)

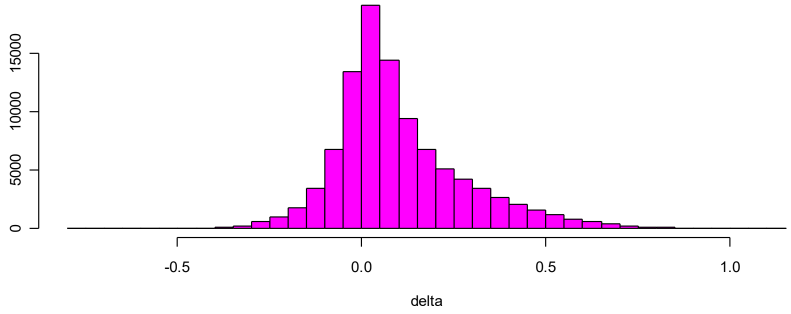
Influence of very diffuse but proper prior on Σ -- shrinks corrs to 0.



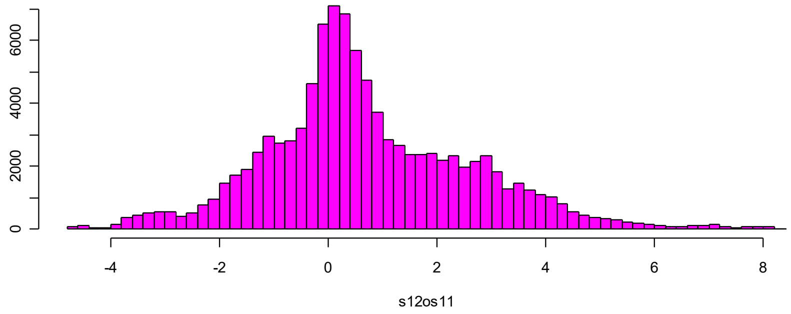
Weak Ins Ex



Posteriors based on 100,000 draws



Inadequacy of Standard Normal Asymptotic Approximations!



Using Mixtures of Normals

We can implement exact sample Bayesian inference with normal errors

However, our friends in econometrics tell us –

don't like making distribution assumptions.

willing to accept loss of efficiency.

willing to ignore adequacy of asymptotics issue or search for different asymptotic experiments (e.g. weak ins asymptotics).

Can we be "non-parametric" without loss of efficiency?

Mixtures of Normal for Errors

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

$$x = \delta z + \varepsilon_1^{'}$$

$$y = \beta x + \varepsilon_2^{'}$$

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

$$ind \sim multinomial(p)$$

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

Identification with Normal Mixtures

The normal mixture model for the errors is not identified. A standard unconstrained Gibbs Sampler will exhibit “label-switching.”

One View: Not an issue, error density is a nuisance parameter. Coefs of structural and instrument equations are identified.

Another View: Any function of the error density is identified. GS will provide the posterior distribution of the density ordinates.

Constrained samplers will often exhibit inferior mixing properties and are unnecessary here.

A Gibbs Sampler

$$(1)\quad \beta \middle| \delta, ind, \{\mu_{k,}\Sigma_k\}, x, y, z$$

$$(2)\quad \delta \middle| \beta, ind, \{\mu_{k,}\Sigma_k\}, x, y, z$$

$$(3)\quad ind, \{\mu_{k,}\Sigma_k\} \middle| \delta, \beta, x, y, z$$

Tricks:

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

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

Gibbs Sampler: beta draw $\beta\left|\delta, ind, \{\mu_k, \Sigma_k\}\right. x, y, z$

Very similar to one comp case, except the error terms have non-zero mean and keep track of which comp each obs comes from!

$$\varepsilon'_{2,i}\left|\varepsilon'_{1,i}\right. = E\left[\varepsilon'_{2,i}\left|\varepsilon'_{1,i}\right.\right] + v_{2,i}; \quad \sigma^2_{v_2,ind_i} = \sigma_{22,ind_i} - \frac{\sigma^2_{12,ind_i}}{\sigma_{11,ind_i}}$$

$$E\left[\varepsilon'_{2,i}\left|\varepsilon'_{1,i}\right.\right] = \mu_{2,ind_i} + \frac{\sigma_{12,ind_i}}{\sigma_{11,ind_i}}\left(\varepsilon'_{1,i} - \mu_{1,ind_i}\right)$$

$$\text{estimate: } \left(y_i - E\left[\varepsilon'_{2,i}\left|\varepsilon'_{1,i}\right.\right]\right)\Big/\sigma_{v_2,ind_i} = \beta x_i \big/ \sigma_{v_2,ind_i} + u_i$$

Gibbs Sampler: delta draw $\delta \Big| \beta, ind, \{\mu_k, \Sigma_k\} x, y, z$

Only trick now is to subtract means of "errors" and keep track of indicator.

As before, we move to "reduced" form with errors, v.

$$Var(v_i) = A\Sigma_{ind_i} A' = \Omega_{ind_i} = L_{ind_i} L_{ind_i}^{'}$$

$$L_{ind_i}^{-1} \begin{pmatrix} x_i - \mu_{1,ind_i} \\ y_i - \mu_{2,ind_i} - \beta\mu_{1,ind_i} \end{pmatrix} = L_{ind_i}^{-1} \begin{bmatrix} z_i \\ \beta z_i \end{bmatrix} \delta + u$$

Fat-tailed Example

Standard outlier model:

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

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

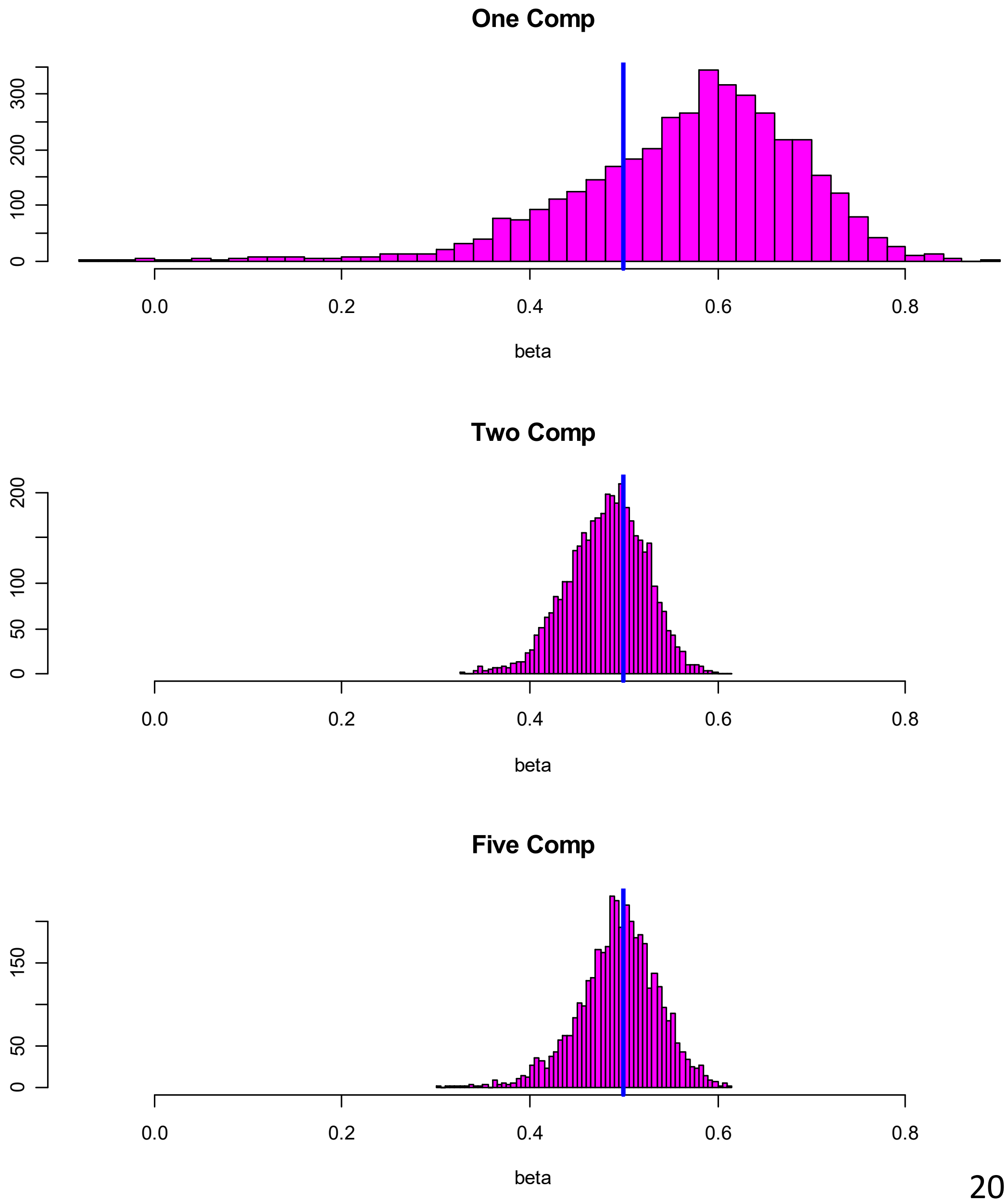
$$\text{comp 2: } \varepsilon' \sim N(0,\Sigma_2) \quad \Sigma_2 = M\Sigma_1$$

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

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! In the plots shown earlier I used 5 components.

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.

Ref: *Practical Nonparametric and Semiparametric Bayesian Statistics* (articles by West and Escobar/MacEachern)

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:

$$\begin{tikzcd}
\varepsilon_1 & \\
\vdots & \\
\varepsilon_i & \theta=(\mu,\Sigma) \arrow[l] \arrow[uul] \arrow[ddl] \\
\vdots & \\
\varepsilon_n &
\end{tikzcd}
\qquad
\begin{tikzcd}
\varepsilon_1 & \theta_1=(\mu_1,\Sigma_1) \arrow[l] \\
\vdots & \\
\varepsilon_i & \theta_i=(\mu_i,\Sigma_i) \arrow[l] \\
\vdots & \\
\varepsilon_n & \theta_n=(\mu_n,\Sigma_n) \arrow[l]
\end{tikzcd}$$

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 $\theta_i = \left(\mu_i, \Sigma_i\right)$

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

DAG:

$$\begin{tikzcd} & \beta,\delta \arrow[dr] & \\ \lambda \arrow[r] & \theta_i \arrow[r] & \varepsilon_i \end{tikzcd}$$

Note: thetas are indep (conditional on lambda)

DP prior

Add another layer to hierarchy – DP prior for theta

DAG:

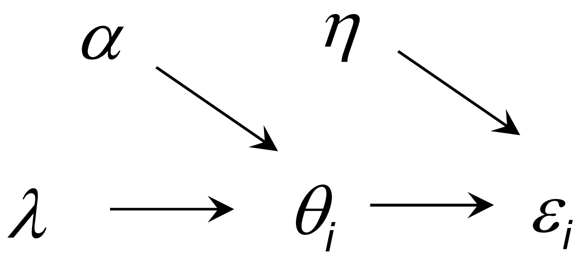
$$\begin{tikzcd}
\alpha \arrow[dr] & & \beta,\delta \arrow[dr] & \\
\lambda \arrow[r] & G \arrow[r] & \theta_i \arrow[r] & \varepsilon_i
\end{tikzcd}$$

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:



$\left\{\theta_1,\ldots,\theta_n\right\}$ 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: Drawing from Posterior

Basis for a Gibbs Sampler:

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

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”

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

DPM: Drawing from Posterior

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

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

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

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

DPM: Predictive Distributions or “Density Est”

$$p\left(\varepsilon_{n+1}\middle|\varepsilon_1,\ldots,\varepsilon_n\right)=\int p\left(\varepsilon_{n+1}\middle|\theta_{n+1}\right)p\left(\theta_{n+1}\middle|\varepsilon\right)d\theta_{n+1}$$

Note: this is like drawing from the first stage prior in hierarchical applications. We integrate out using the posterior distribution of the hyper-parameters.

$$p\left(\theta_{n+1}\middle|\varepsilon\right)=\int p\left(\theta_{n+1}\middle|\theta_1,\ldots,\theta_n\right)p\left(\theta_1,\ldots,\theta_n\middle|y\right)d\theta_1\cdots d\theta_n$$

Both equations are derived by using conditional independence.

DPM: Predictive Distributions or "Density Est"

$$\theta_{n+1} \middle| \theta \sim \begin{cases} \textit{with prob } \dfrac{\alpha}{\alpha+n}, \textit{ draw from } G_0(\lambda) \\ \textit{with prob } \dfrac{1}{\alpha+n}, \textit{ draw from } \delta_i \; i = 1,\ldots,n \end{cases}$$

Algorithm to construct predictive density:

1. draw $\theta_{n+1} \middle| \theta^r, \lambda$
2. construct $p\left(\varepsilon_{n+1} \middle| \theta_{n+1}^r\right)$
3. average to obtain predictive density

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.

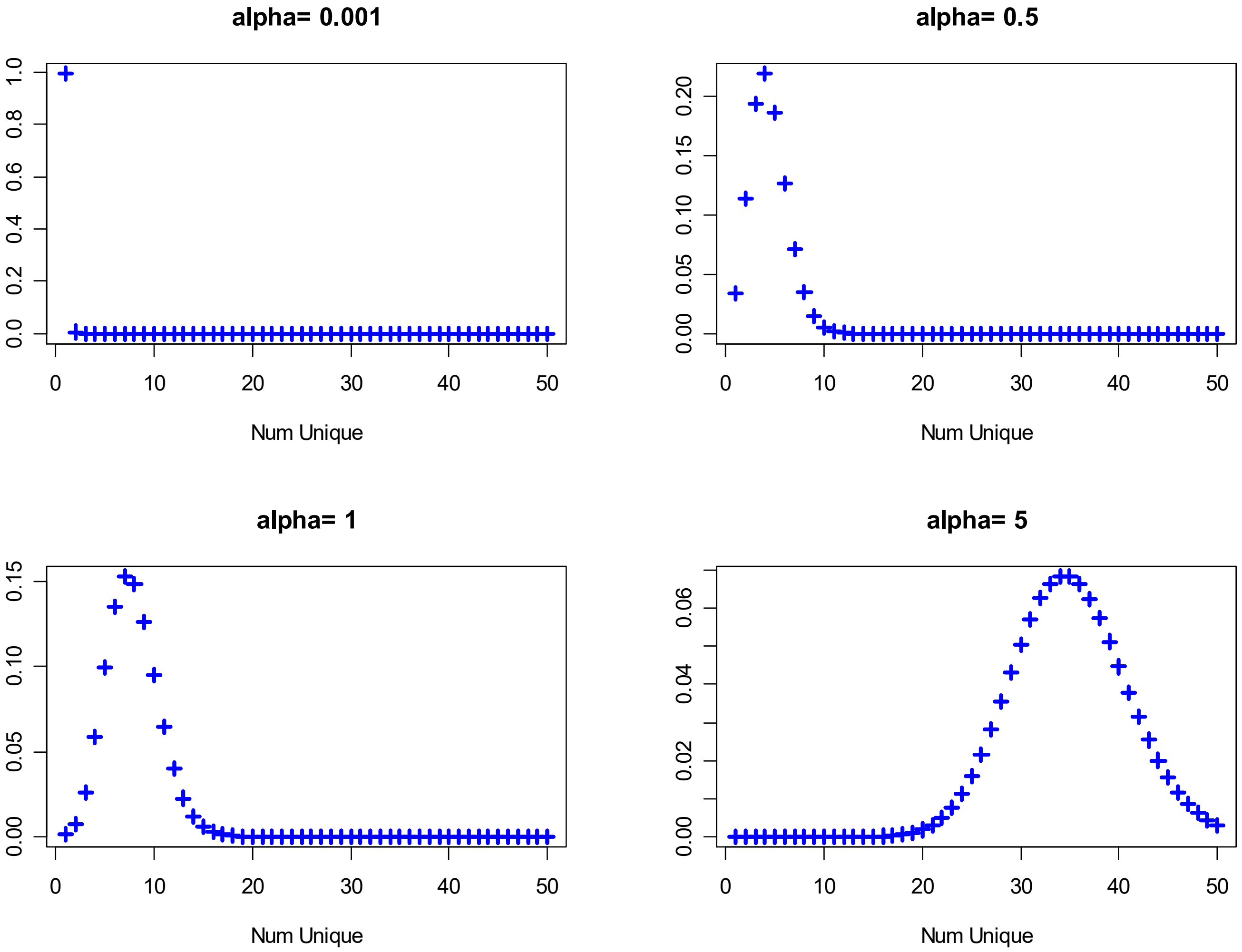
$$\Pr\left(I^{*} = k\right) = \left\| S_n^{(k)} \right\| \alpha^k \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)}$$

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

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

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.

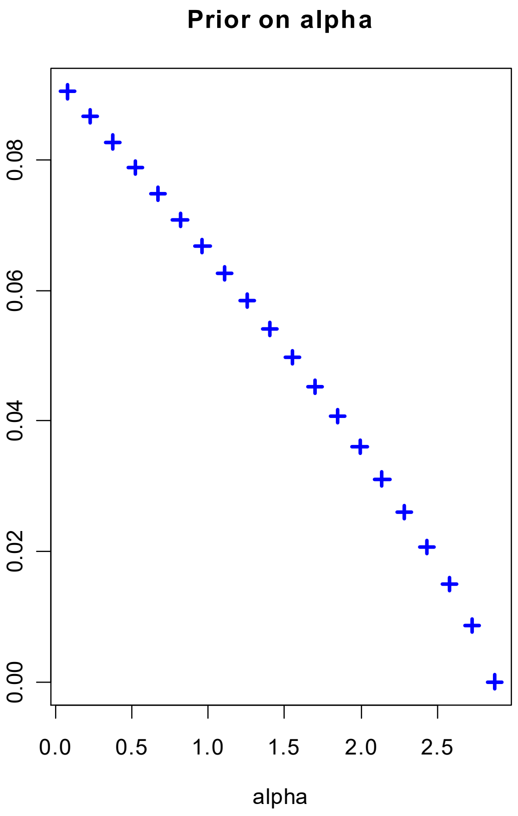
"Solution:" put a prior on alpha.

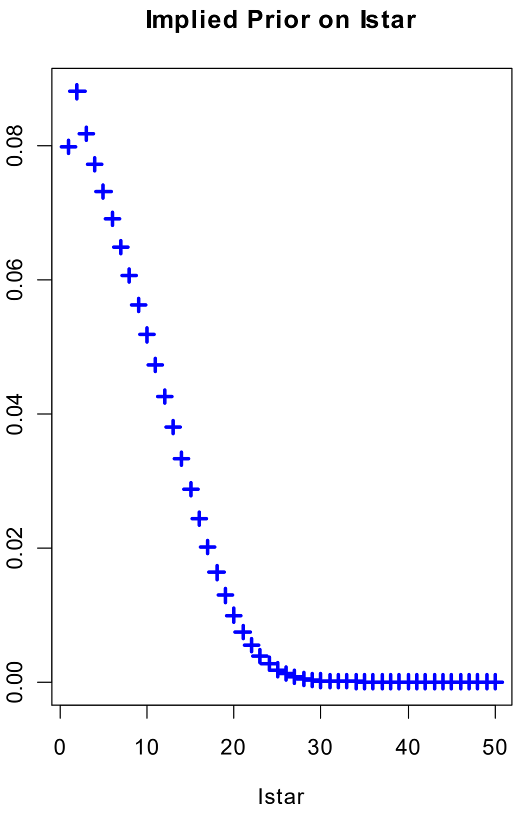
Assess prior by examining the priori distribution of number of unique theta.

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

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

Assessing the DP prior: Priors on α





Assessing the DP prior: Choice of λ

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

Both α and λ determine the probability of a “birth.”

Intuition:

1. Very diffuse settings of λ 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 λ

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

Choice of λ made easier if we center and sacle 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

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

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

Assessing the DP prior: Choice of λ

Look at marginals of μ and σ_1

Choose (υ, v, a)

∋

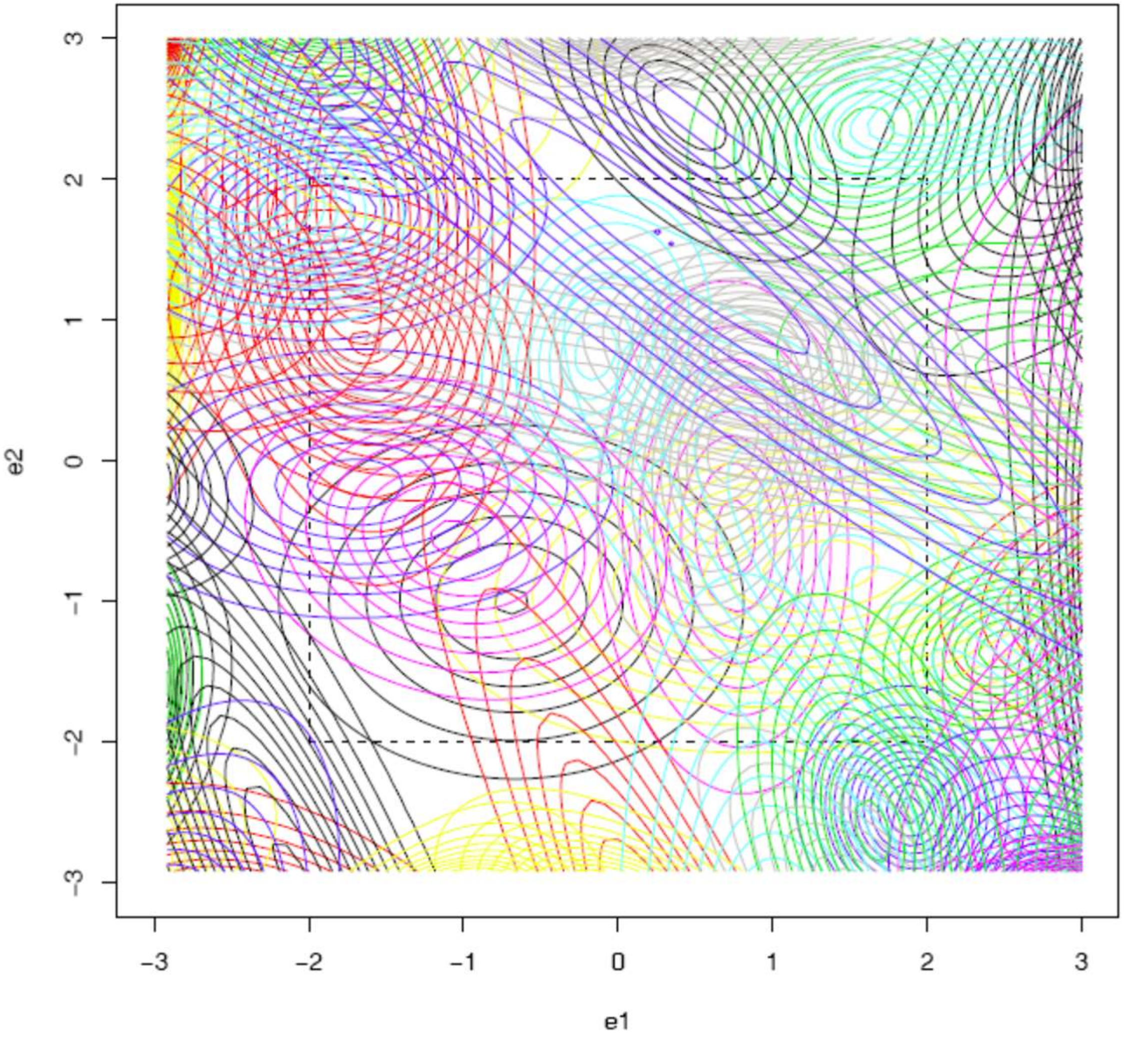
$$\Pr\left[.25 < \sigma_1 < 3.25\right] = .8$$

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

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

Very Diffuse!

Draws from G_0



Gibbs Sampler for DP in the IV Model

$\beta\left|\delta,\{\theta_i\},x,y,z\right.$	Same as for Normal Mixture Model
$\delta\left|\beta,\{\theta_i\},x,y,z\right.$	
$\{\theta_i\}\left|\delta,\beta,x,y,z\right.$	Doesn’t Vectorize
$\{\theta_i^*\}\left|ind,\delta,\beta,x,y,z\right.$	“Remix” Step
$\alpha\left|I^*\right.$	Trivial (discrete)

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

Coding DP and IV in R

$$\beta \,\middle|\, \delta, \{\theta_i\}, x, y, z$$

$$\delta \,\middle|\, \beta, \{\theta_i\}, x, y, z$$

Same as for Normal Mixture Model

$$\boxed{\{\theta_i\} \,\middle|\, \delta, \beta, x, y, z}$$

Doesn’t Vectorize

$$\left\{\theta_i^*\right\} \middle| ind, \delta, \beta, x, y, z$$

“Remix” Step

$$\alpha \,\middle|\, I^*$$

Trivial (discrete)

Coding DP and IV in R $\{\theta_i\} | \delta, \beta, x, y, z$

To draw indicators and new set of theta, we have to “Gibbs thru” each observation. We need routines to draw from the Base Prior and Posterior from “one obs” and base Prior (birth step).

We summarize each draw of using a list structure for the set of unique thetas and an indicator vector (length n).

We code the thetadraw in C but use R functions to draw from Base Posterior and evaluate densities at new theta value.

Coding DP and IV in R: inside rivDP

```
for(rep in 1:R) {
 # draw beta and gamma
       out=get_ytxt(y=y,z=z,delta=delta,x=x,
          ncomp=ncomp,indic=indic,comps=comps)
       beta = breg(out$yt,out$xt,mbg,Abg)

 # draw delta
      out=get_ytxtd(y=y,z=z,beta=beta,
             x=x,ncomp=ncomp,indic=indic,comps=comps,dimd=dimd)
      delta = breg(out$yt,out$xtd,md,Ad)

 # DP process stuff- theta | lambda
      Err = cbind(x-z%*%delta,y-beta*x)}
      DPout=rthetaDP(maxuniq=maxuniq,alpha=alpha,lambda=lambda,
              Prioralpha=Prior$Prioralpha,theta=theta,y=Err,

yden=reqfun$yden,q0=reqfun$q0,thetaD=reqfun$thetaD,GD=reqfun$GD)
      indic=DPout$indic
      theta=DPout$theta
      comps=DPout$thetaStar
      alpha=DPout$alpha
      Istar=DPout$Istar
      ncomp=length(comps)

}
```

Coding DP and IV in R: Inside rthetaDP

```
   # initialize indicators and list of unique thetas
   thetaStar=unique(theta);nunique=length(thetaStar)
   q0v = q0(y,lambda,eta)

   ydenmat[1:nunique,]=yden(thetaStar,y,eta)
   #  ydenmat is a length(thetaStar) x n array f(y[j,] | thetaStar[[i]]

   # use .Call to draw theta list
   theta=
.Call("thetadraw",y,ydenmat,indic,q0v,p,theta,thetaStar,lambda,eta,
                  thetaD=thetaD,yden=yden,maxuniq,new.env())
   thetaStar=unique(theta)
   nunique=length(thetaStar)
   newthetaStar=vector("list",nunique)

   #thetaNp1 and remix
   probs=double(nunique+1)
   for(j in 1:nunique) {
       ind = which(sapply(theta,identical,thetaStar[[j]]))
       probs[j]=length(ind)/(alpha+n)
       new_utheta=thetaD(y[ind,,drop=FALSE],lambda,eta)
       for(i in seq(along=ind)) {theta[[ind[i]]]=new_utheta}
       newthetaStar[[j]]=new_utheta
       indic[ind]=j
   }
   # draw alpha
```

Coding DP and IV in R: Inside thetadraw.C

```
/* start loop over observations */
   for(i=0;i < n; i++){
       probs[n-1]=NUMERIC_POINTER(q0v)[i]*NUMERIC_POINTER(p)[n-1];
       for(j=0;j < (n-1); j++){
             ii=indic[indmi[j]]; jj=i;      /* find element ydenmat(ii,jj+1) */
             index=jj*maxuniq+(ii-1);
             probs[j]=NUMERIC_POINTER(p)[j]*NUMERIC_POINTER(ydenmat)[index];
       }
       ind=rmultin(probs,n);

       if(ind == n){
          yrow=getrow(y,i,n,ncol);
          SETCADR(R_fc_thetaD,yrow);
          onetheta=eval(R_fc_thetaD,rho);
          SET_ELEMENT(theta,i,onetheta);
          SET_ELEMENT(thetaStar,nunique,onetheta);
          SET_ELEMENT(lofone,0,onetheta);
          SETCADR(R_fc_yden,lofone);
          newrow=eval(R_fc_yden,rho);
          for(j=0;j<n; j++)
          {NUMERIC_POINTER(ydenmat)[j*
               maxuniq+nunique]=NUMERIC_POINTER(newrow)[j];}
                     indic[i]=nunique+1;
                     nunique=nunique+1;}
       else {
          onetheta=VECTOR_ELT(theta,indmi[ind-1]);
          SET_ELEMENT(theta,i,onetheta);
          indic[i]=indic[indmi[ind-1]];
       }
    }
```

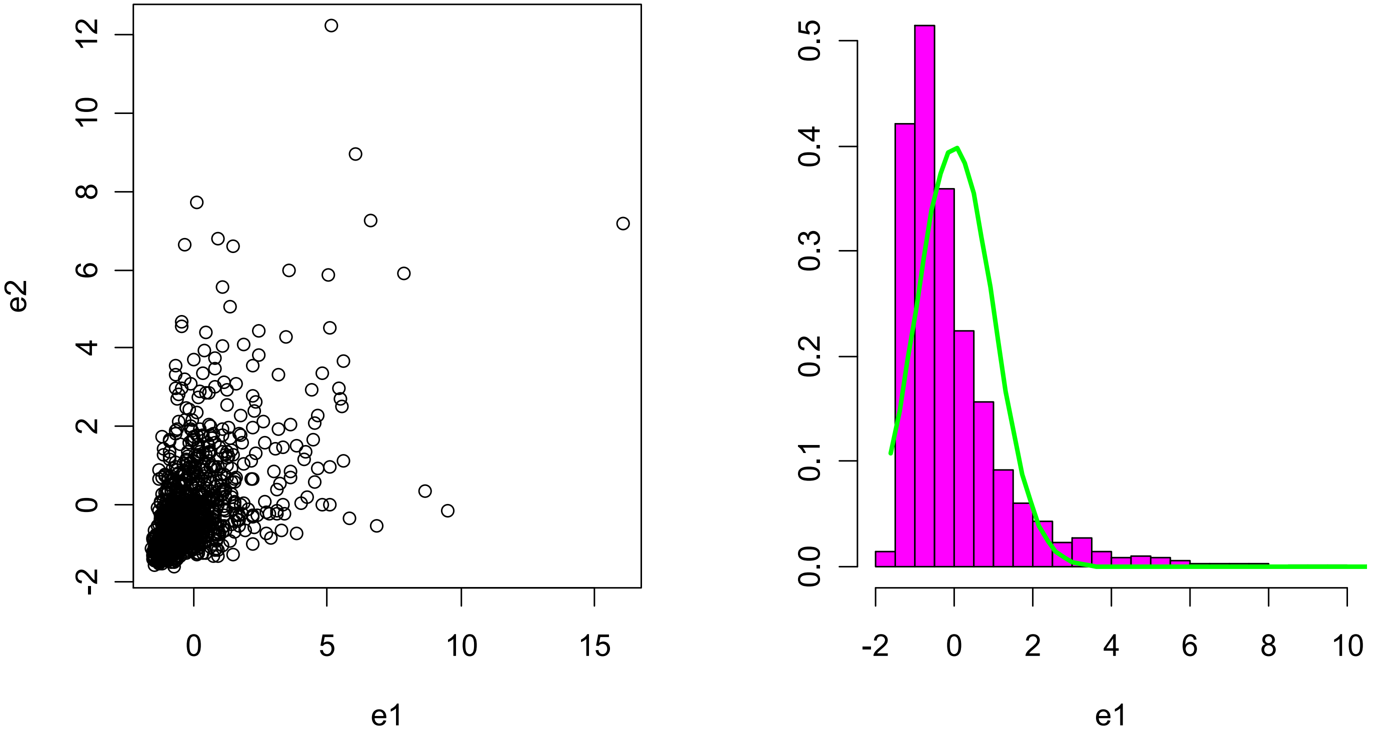
Sampling Experiments

Examples are suggestive, but many questions remain:

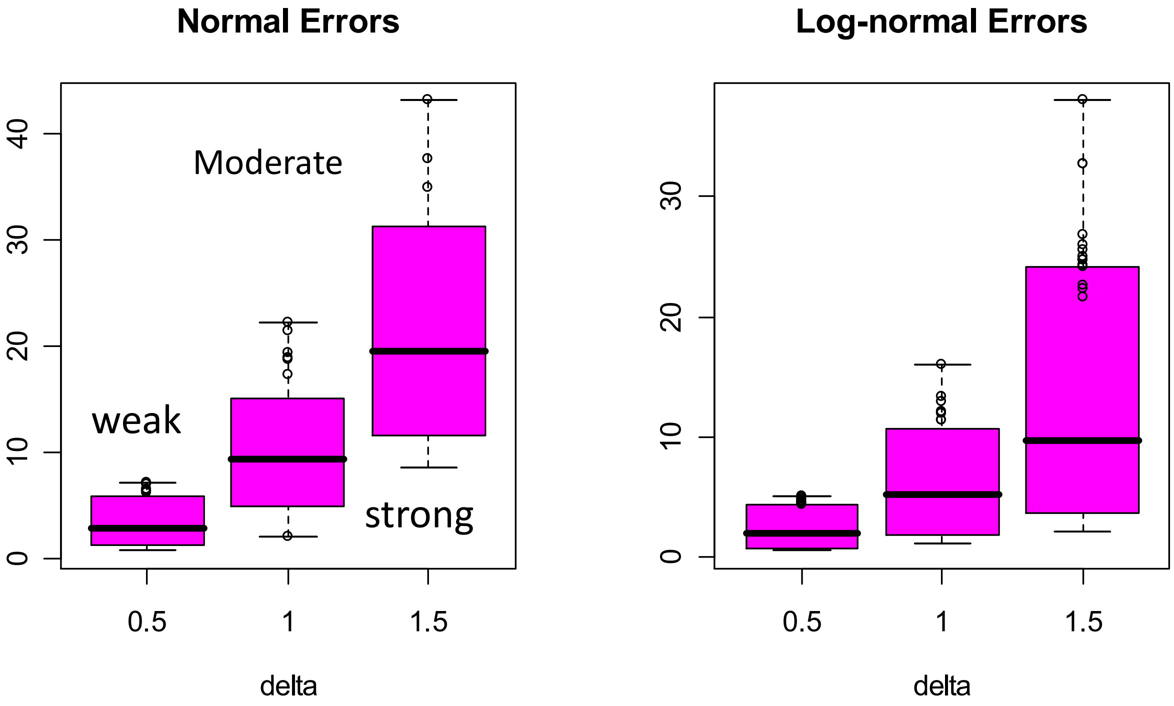
1. How well do DP models accommodate departures to normality?
2. How useful are the DP Bayes results for those interested in “standard” inferences such as confidence intervals?
3. How do conditions of many instruments or weak instruments affect performance?

Sampling Experiments – Choice of Non-normal Alternatives

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



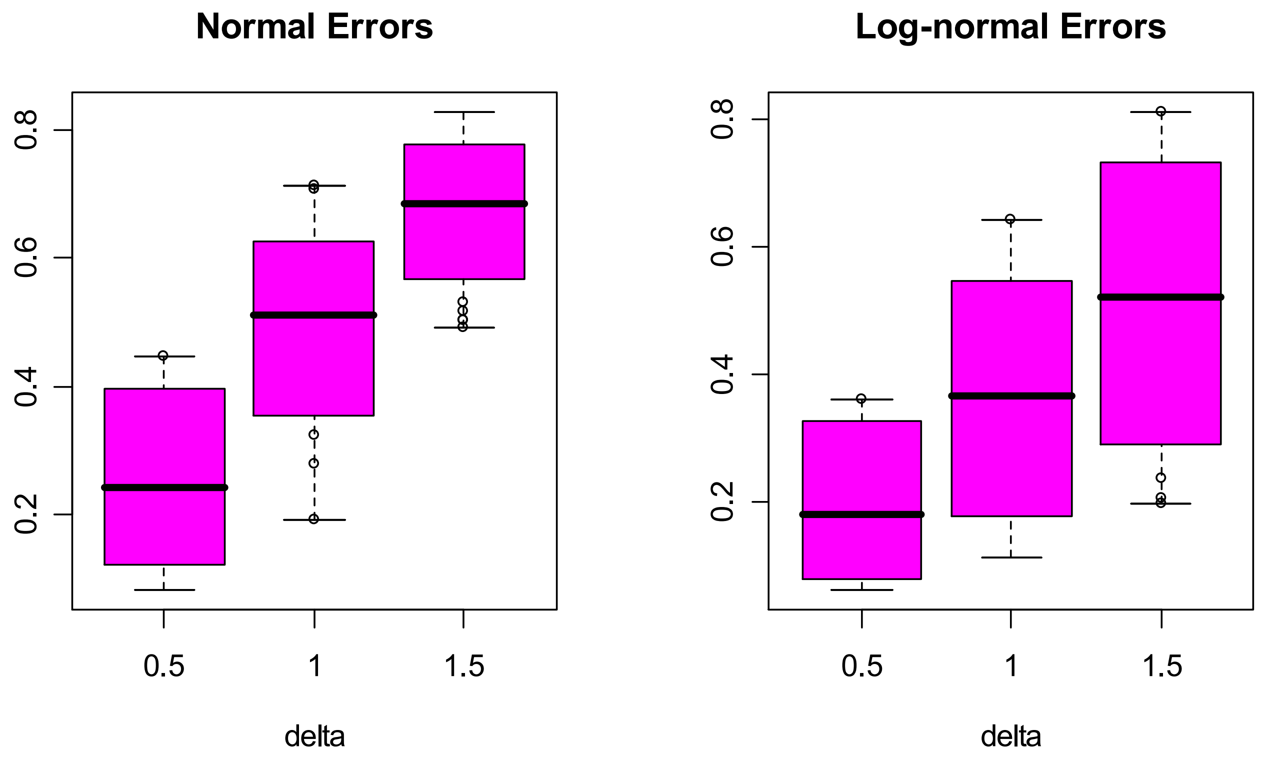
Sampling Experiments – Strength of Instruments- F stats



k=10

Weak case is bounded away from zero. We don’t have huge numbers of data sets with no information!

Sampling Experiments – Strength of Instruments- 1st Stage R-squared



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).

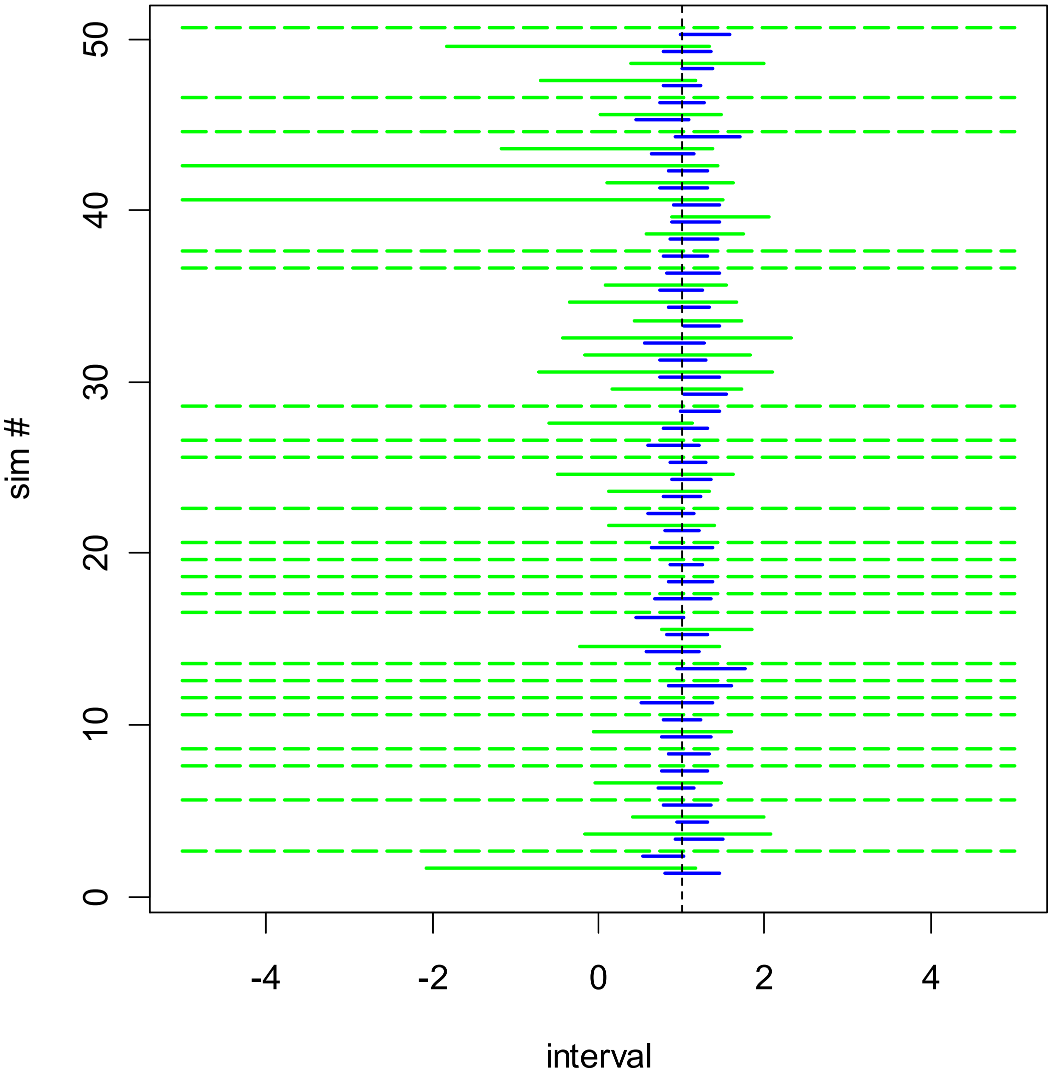
Sampling Experiments- Coverage of "95%" Intervals

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

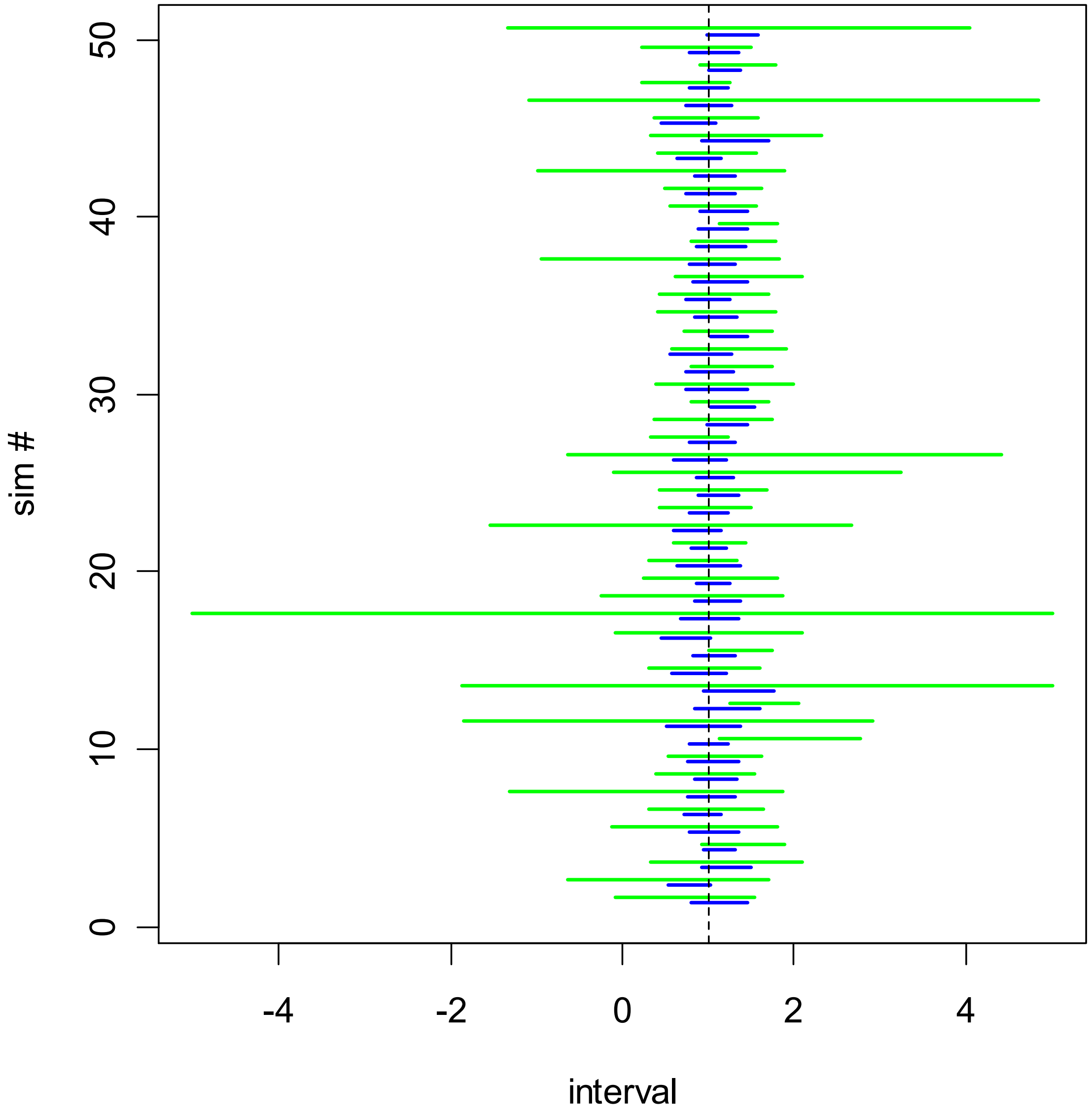
7% (normal) | 42 % (log-normal) are infinite length

Bayes Vs. CLR (Andrews 06)



Weak
Instruments
Log-Normal
Errors

Bayes Vs. Fuller-Many



Weak
Instruments
Log-Normal
Errors

Infinite Intervals vs. First Stage F-test

Weak Instruments | Log-Normal Errors Case

	Significant	Insignificant
Finite	190	42
Infinite	10	158

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!

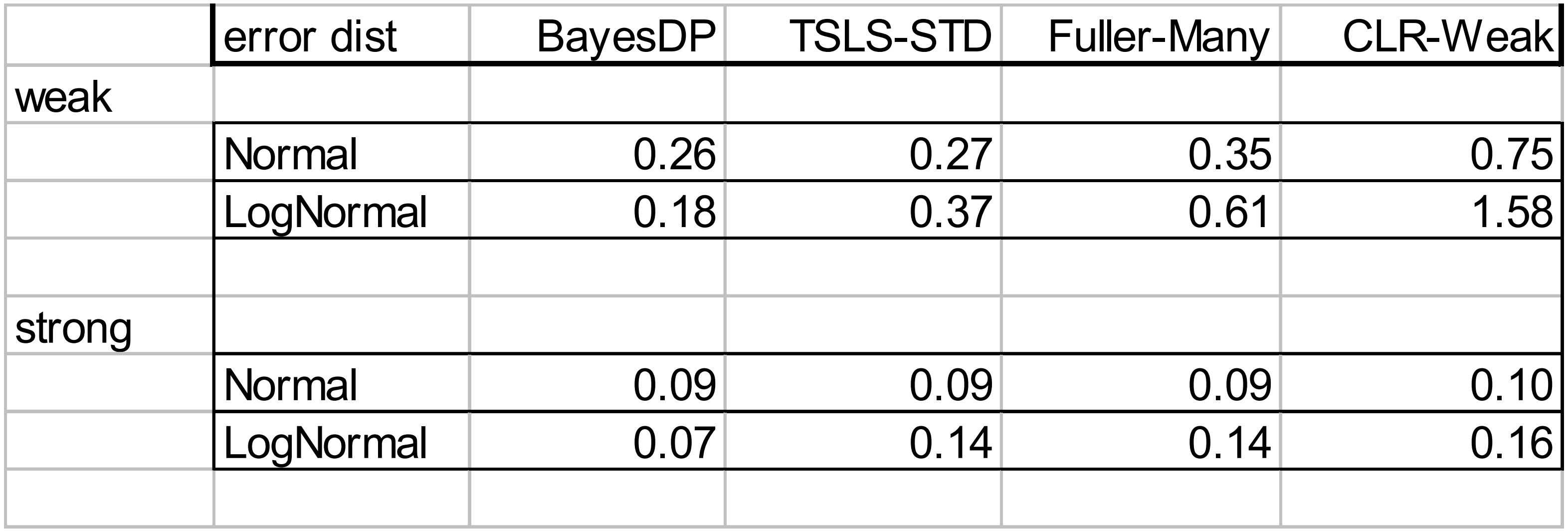
Smaller and located closer to the true beta.

Scalar measure:

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

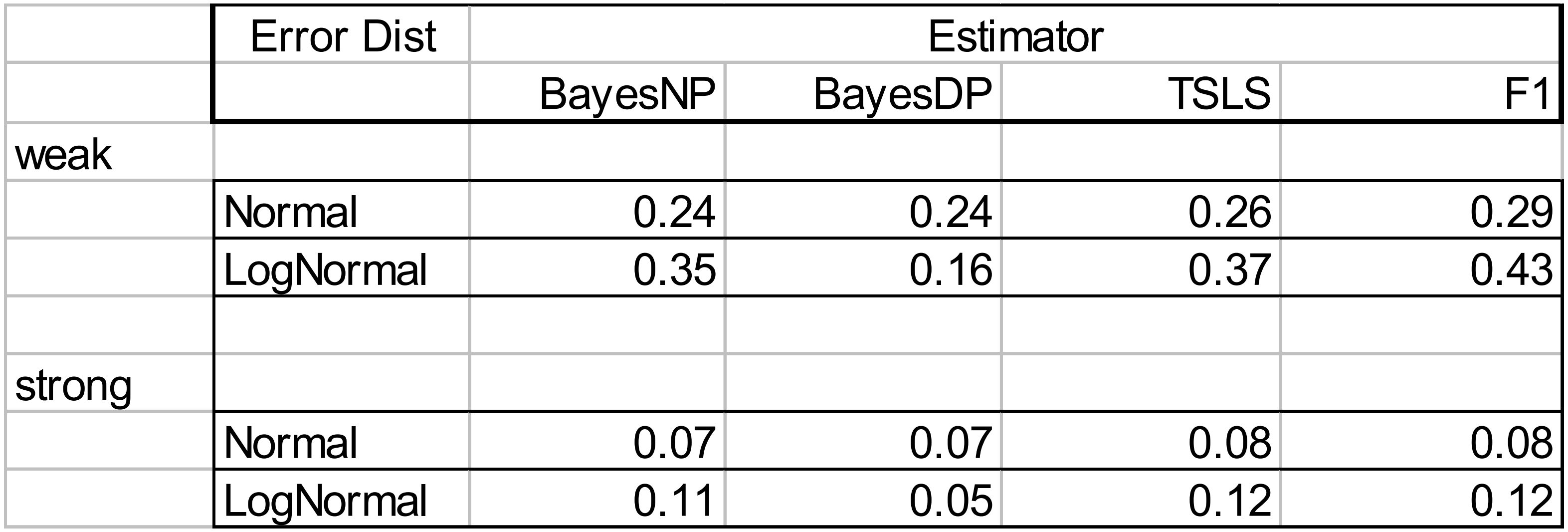
$$E[|X-\beta|] = \int_L^U |x-\beta| \frac{1}{U-L} dx$$

Interval Performance



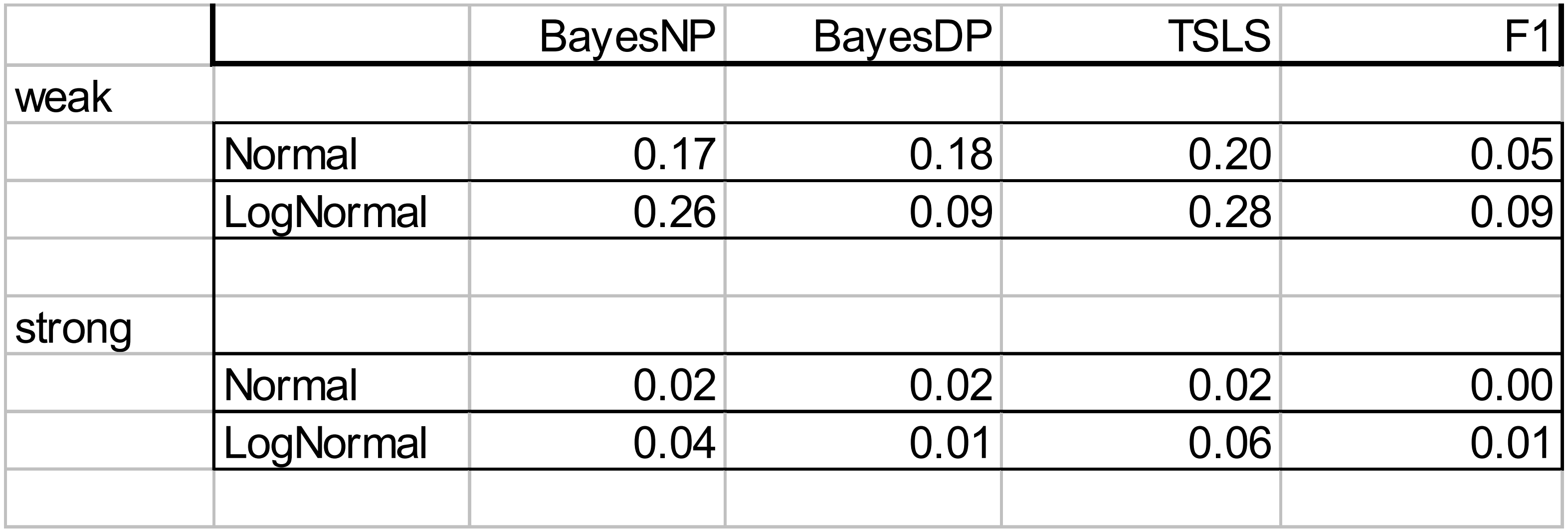
	error dist	BayesDP	TSLS-STD	Fuller-Many	CLR-Weak
weak					
	Normal	0.26	0.27	0.35	0.75
	LogNormal	0.18	0.37	0.61	1.58
strong					
	Normal	0.09	0.09	0.09	0.10
	LogNormal	0.07	0.14	0.14	0.16

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.

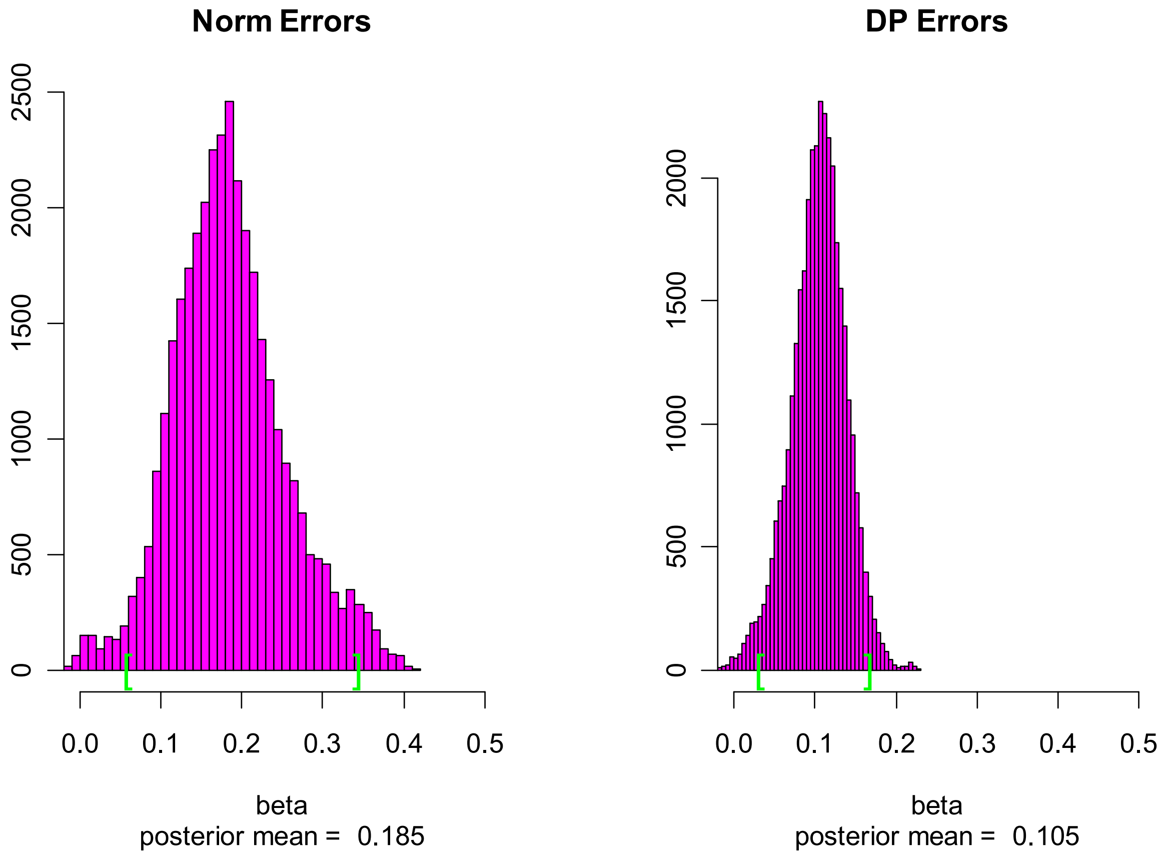
x education (yrs)

z is proximity to 2 and 4 year colleges

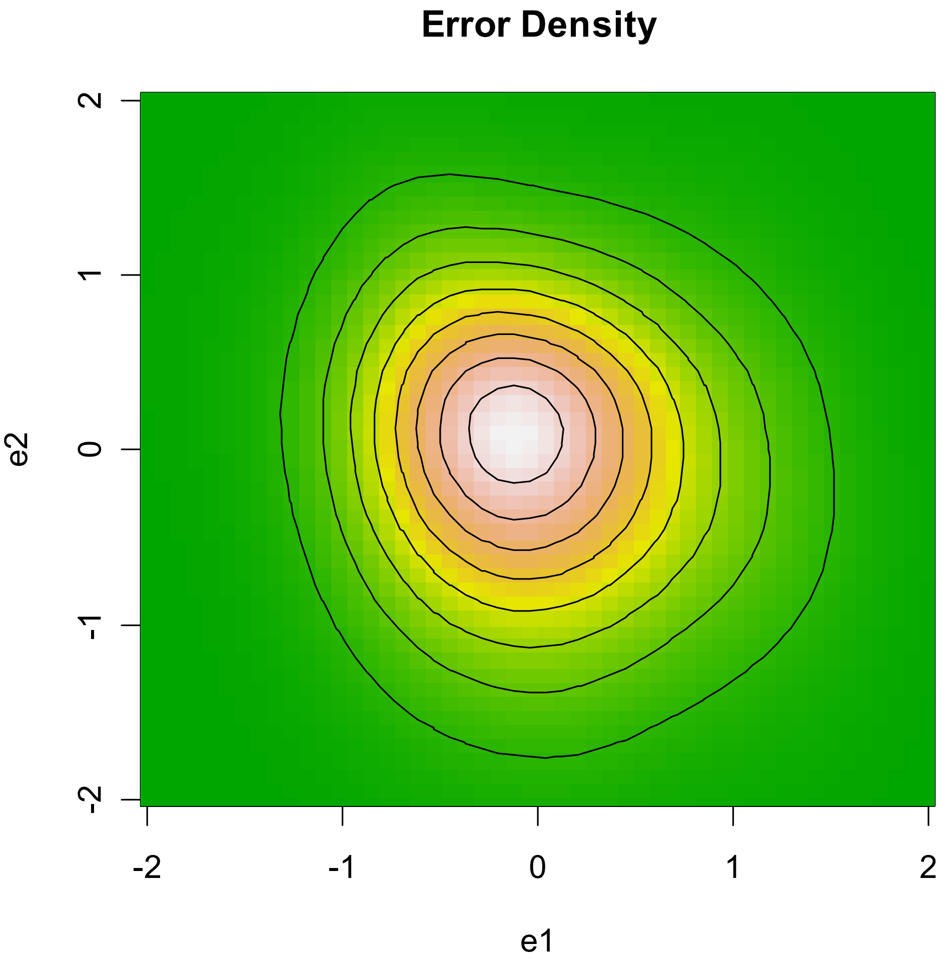
N=3010.

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

An Example: Card Data



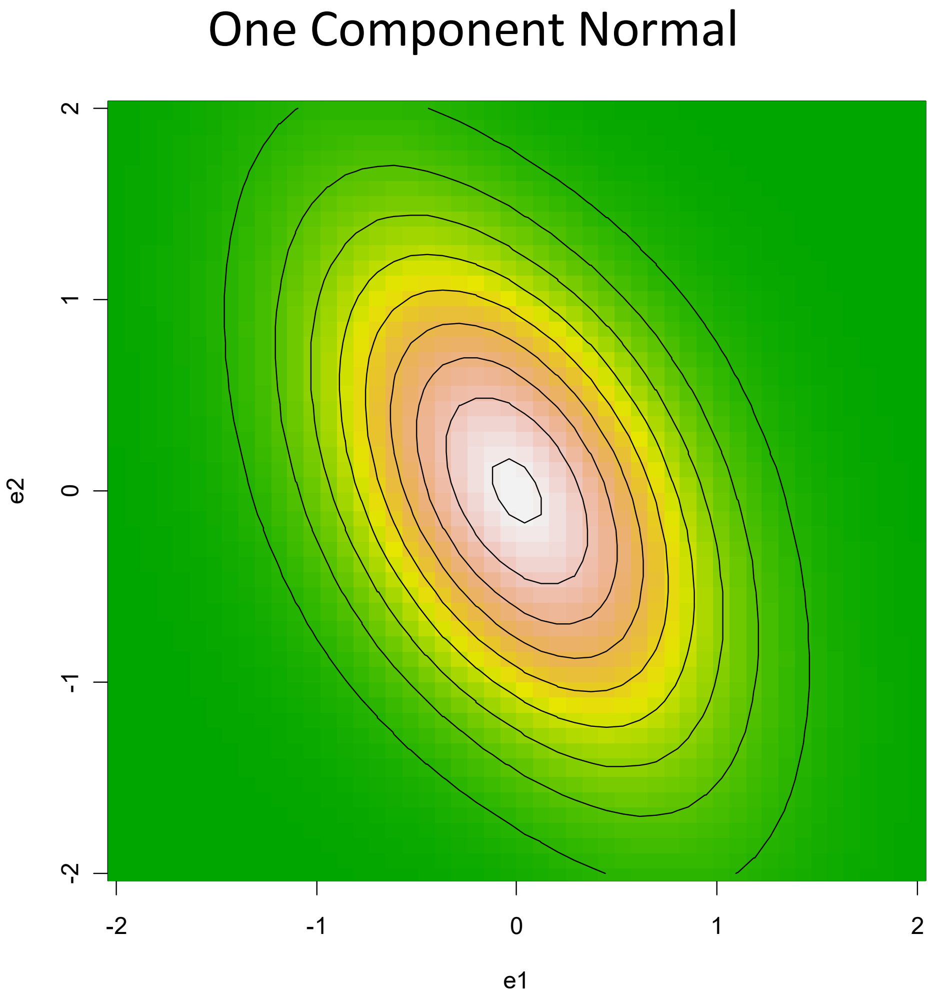
An Example: Card Data



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”

Conclusions

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?

General approach extends trivially to allow for nonlinear structural and reduced form equations via the same device of allowing clustering of parameter values.