Tech Report:
Setting Priors and Starting Values

Xun Cao, Michael D. Ward, and Randolph M. Siverson*

February 20, 2008

Introduction

This document discusses detailed MCMC estimation strategies for the estimation of a Bayesian generalized-bilinear-mixed-effects (`gbme`) model to study militarized inter-state conflict (MID) (Ward, Siverson and Cao 2007).

Programs are available on the replication Web site (http://dvn.iq.harvard.edu/dvn/dv/mward) to implement these procedures, but do not detail our strategy in a sufficiently transparent way. As a result, we have put together this short description that details our procedures.[1]

1 Step 1

We started by estimating the model for the year 2000 (Y2K). This year is the most recent in our database, has the largest number of countries, and is also the year with the largest number of interstate disputes. As a result, we viewed it as an appropriate starting point.

Because of the rare-event nature of militarized inter-state conflict (MID), the `gbme.glmstart` function built in `gbme.r` to estimate starting values and priors from the data often over-fits the data. As a result it generates priors and starting values that delay the MCMC convergence. If possible, it is better to find a different set of informative priors rather than fall back on those created by default. This model was run for 500,000 iterations. Once the MCMC had converged for the year 2000, we used the information drawn from these posterior distributions to set up priors and starting values for other years, that is, 1950-1995. We know that 1950-1995 is temporally prior to Y2K, but the estimates for Y2K are our (best, extant) *prior* information for the data generating process for the earlier years.

1.1 Uninformative priors and starting values for Y2K

We used the following code to set up priors and starting values for Y2K. Fast and stable convergence of the MCMC is the goal. In theory, the MCMC will converge if one runs the chain long enough. But in practice, we need to choose priors and starting values to speed up the process. A MCMC of more than 1 million iterations and taking more than a week to run obviously is not a generally feasible option. With poorly chosen priors and starting values, several million iterations may be necessary for convergence.

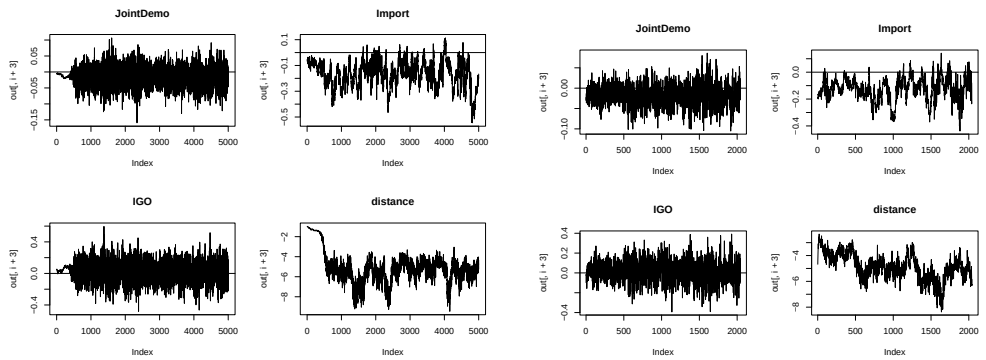
We used the following code for Y2K:

*Thanks to Brandon J. Kinne for his patient queries about our research and to Nils Weidmann for his guidance.

[1] Our former replication page has moved, along with one of the coauthors, to Essex University: http://privatewww.essex.ac.uk/∼caox/publications/ajpsarchive.htm.

```
data.2000<-dget("data2000")
Yd<-data.2000[[1]]
# read in program to analyze the data
n<-dim(Yd)[1]

Xd<-data.2000[[2]]
Xs<-Xr<-data.2000[[3]]

source("gbme.r")
n<-dim(Yd)[1]
Y<-Yd

tmp<-matrix(1,n,n)
tmp[lower.tri(tmp,diag=T)]<-NA
UT<<-tmp

n<<-dim(Y)[1]
rd<<-dim(Xd)[3]
rs<<-dim(Xs)[2]
rr<<-dim(Xr)[2]

pi.Sab<-c(100,10,100,4) # 4 is the degree of freedom for inverse-Wishart
pi.s2u=c(2,1)
pi.s2v=c(2,1)
pi.s2z=matrix(nrow=3, ncol=2, c(rep(2,3), rep(1,3)))

pim.bd<-c(0,0,0,0) # suppose we have no idea about the
relationship at all piS.bd<-matrix(0, ncol=4, nrow=4)
# let's assume that they are 0 correlated; diag(piS.bd)<-100
# the variance for each  is 100.

pim.b0sr<-c(-20, rep(0, 6)) # -20x2 is our guess about the constant;
piS.b0sr<-matrix(0, ncol=7, nrow=7)
diag(piS.b0sr)<-c(400, rep(100, 6))
pim.b0s<-NULL
piS.b0s<-NULL
pi.s2a<-NULL

#starting values - provide if estart=F:
#leave the starting values as those from the estimation.
beta.d<-rep(0,4)  #(a)$startv$beta.d
beta.u<-c(-20, rep(0, 6))
s<-rep(0, n)
r<-rep(0, n)
z<-matrix(0, ncol=3, nrow=n)

priors<-list(pi.Sab=pi.Sab,pi.s2u=pi.s2u,pi.s2v=pi.s2v,pi.s2z=pi.s2z,
                  pim.bd=pim.bd,piS.bd=piS.bd,
                  pim.b0sr=pim.b0sr,piS.b0sr=piS.b0sr,
                  pim.b0s=pim.b0s, piS.b0s=piS.b0s , pi.s2a=pi.s2a)
startv<-list(beta.d=beta.d,beta.u=beta.u,s=s,r=r,z=z)

# call the main estimation package, gbme
gbme(Y=Yd,Xd=Xd,Xs=Xs,Xr=Xr,k=3,fam="binomial",N=matrix(1,n,n),
     NS=500000,awrite=T,bwrite=T, estart=F, directed=T)
```

Figure 1: *Comparison of MCMC for 2000 data with uninformative and informative priors.*



(a) MCMC for 2000 data with uninformative priors and starting values; analysis for 2007 AJPS article.

(b) Using results from (a) as priors, these estimations were undertaken in 2008.

Figure 1 shows the estimation of the four dyadic coefficients. These results are those reported in the 2007 article. Replications may differ slightly owing to random sampling and the length of the replicated chain, as well as the seed of the random number generator–which should have been fixed for replication purposes, but unfortunately, *was not*. More detailed information about these posteriors are reported in Table 2 in Ward, Siverson, and Cao (2007).

To make sure that the MCMC converged, we used the posteriors from the estimation of Y2K (i.e., those in Figure 1(a)) as priors and starting values and started MCMC again for Y2K. Figure 1(b) illustrates these results and shows that the MCMC stays in the same general areas, verifying that the MCMC based on uninformative priors and starting values.

1.2 Re-replication

We also returned to tabla rasa, and replicated the initial chain for Y2K from *un*informative priors. Similar caveats hold for this re-replication: starting seed values and chain length may introduce some additional variability into these results. Figure 2 illustrates these results.

2 Step Two

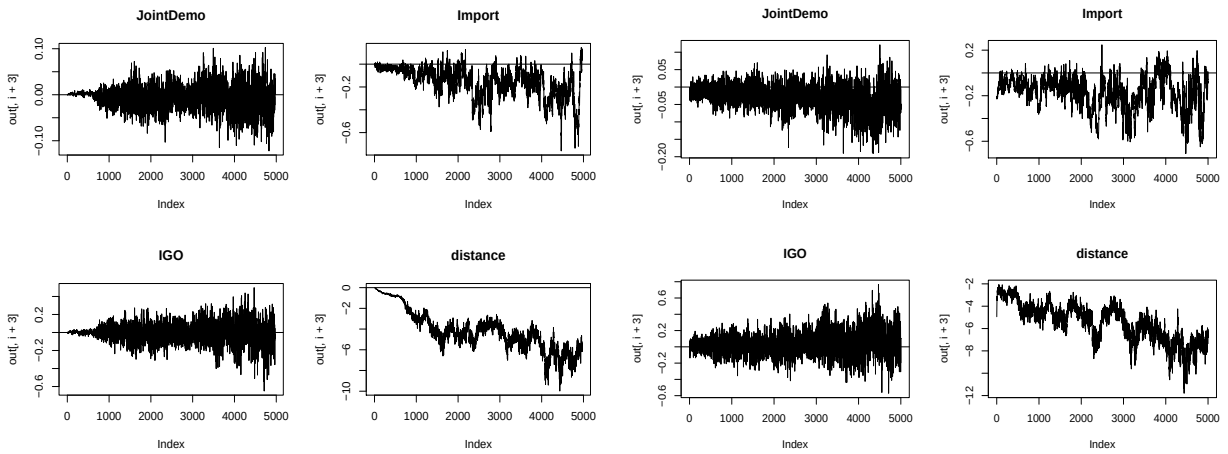
Next we used the posterior information from the run for Y2K as starting values for each other year. This is the best information we have for starting values. As an example, we present the analysis for 1950.

```
## a routine to set up priors and starting values for gbme:

year.pre<-2000 # to be used to get posterior and priors for starting values
year.cur<-1950 # the year to be estimated
burn<-2500 # choose number of burn-in
k<-3
diffuse<-100 # this parameter is the estimate of the
             # covariance of pim.bd and pim.b0sr;

out<-read.table("OUT.base", header=T)
dat<-dget(paste("data", as.character(year.pre), sep=""))
m<-dim(dat[[1]])[1] # of countries in MCMC for this year.
```

Table 1: Posterior Distributions of estimated coefficients for Y2K estimated from priors and posteriors taken from Y2K. These results are equivalent to those found in the July 2007 AJPS article.

	2.5%	Mean	97.5%
Constant	−88.68	−58.18	−34.68
$Polity_i \times Polity_j$	−0.08	−0.02	0.04
$Import_{ij}$	−0.37	−0.15	0.03
IGO_{ij}	−0.24	0.01	0.26
$Distance_{ij}$	−8.62	−5.78	−4.02
$Population_i$	−0.01	0.01	0.04
GDP_i	0.00	0.01	0.01
$Polity_i$	−0.50	0.47	1.53
$Population_j$	−0.03	−0.00	0.02
GDP_j	0.00	0.01	0.01
$Polity_j$	−0.57	0.41	1.43
σ_a^2		189.14	
σ_{ab}		141.14	
σ_b^2		153.87	
σ_ε^2		294.52	
ρ	0.99	1.00	1.00
$\ln \mathcal{L}$		−29.84	

Figure 2: Comparisons of MCMC for Y2K with uninformative and informative priors, replicated this time.



(a) MCMC for 2000 data with uninformative priors and starting values, computed in 2008.

(b) MCMC for 2000 data using results in (a) as informative priors, computed in 2008.

```
n<-dim(out)[1] # of iterations.

# mean sender and receiver effects: s and r
wg.names<-as.character(rownames(dat[[3]]))
```

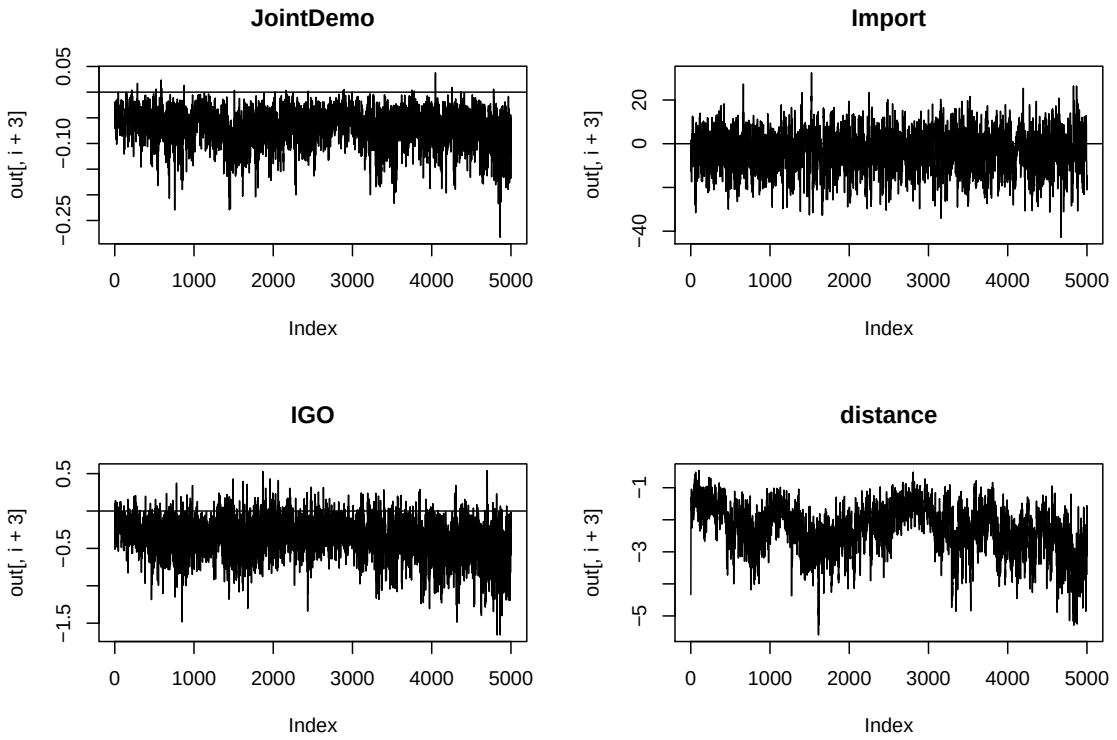
```
A<-read.table("A.base",sep=" ",col.names=wg.names)
B<-read.table("B.base",sep=" ",col.names=wg.names)
A<-A[-c(1:burn),]
B<-B[-c(1:burn),]
a<-apply(A,2,mean) # mean random effect for sender
b<-apply(B,2,mean) # mean random effect for receiver
cbind(a,b) # mean sender/receiver effects;
cor.test(a,b)
ab<-cbind(a,b)
Sab<-var(ab) # we are not going to use this for pi.Sab later on
             # we will use the mean values of S2a, Sab, and S2b;
###
Sum.post<-NULL
for (j in c(3:dim(out)[2])){
                                sum.post<-c(quantile(out[,j][-c(1:burn)],
                                c(.025, .975))[1],
                                mean(out[,j][-c(1:burn)]),
                                quantile(out[,j][-c(1:burn)], c(.025, .975))[2],
                                sd(out[,j][-c(1:burn)]))
                                Sum.post<-rbind(Sum.post, sum.post)
                                }

Sum.post<-round(Sum.post, digits=4)
colnames(Sum.post)<-c("2.5%", "Mean", "97.5%", "St.Dev")
rownames(Sum.post)<-colnames(out)[c(3:dim(out)[2])]

### priors: we get information from the posterior distributions:
pi.Sab<-c(Sum.post["s2a", "Mean"],
          Sum.post["sab", "Mean"],
          Sum.post["s2b", "Mean"],4)
# 4 is the degree of freedom for inverse-Wishart, somehow fixed for a diffuse prior.
s2u<-2*Sum.post["s2e", "Mean"]*(1+Sum.post["rho", "Mean"])
# calculate the mean of s2u and s2v by reconstructing
s2v<-2*Sum.post["s2e", "Mean"]*(1-Sum.post["rho", "Mean"])
# them from posterior mean of s2e and rho.
pi.s2u=c(2,s2u) # here is very convenient because the shape parameter is 2, then
pi.s2v=c(2,s2v) # the scale parameter is the mean this is special case.
pi.s2z<-cbind( rep(2,k), rep(1, k) ) # let's just choose some arbitrary values here.

pim.bd<-c(Sum.post["bd1", "Mean"],
          Sum.post["bd2", "Mean"],
          Sum.post["bd3", "Mean"],
          Sum.post["bd4", "Mean"])
piS.bd<-cov(cbind(out[,"bd1"][-c(1:burn)],
            out[,"bd2"][-c(1:burn)],
            out[,"bd3"][-c(1:burn)],
            out[,"bd4"][-c(1:burn)]))*diffuse*diffuse

pim.b0sr<-c(Sum.post["b0", "Mean"],
            Sum.post["bs1", "Mean"],
            Sum.post["bs2", "Mean"],
            Sum.post["bs3", "Mean"],
            Sum.post["br1", "Mean"],
            Sum.post["br2", "Mean"],
            Sum.post["br3", "Mean"])
```

```
piS.b0sr<-cov(cbind(out[,"b0"][-c(1:burn)],
                out[,"bs1"][-c(1:burn)],
                out[,"bs2"][-c(1:burn)],
                out[,"bs3"][-c(1:burn)],
                out[,"br1"][-c(1:burn)],
                out[,"br2"][-c(1:burn)],
                out[,"br3"][-c(1:burn)]))*diffuse*diffuse
pim.b0s<-NULL
piS.b0s<-NULL
pi.s2a<-NULL

### starting values
beta.d<-c(Sum.post["bd1", "Mean"],
          Sum.post["bd2", "Mean"],
          Sum.post["bd3", "Mean"],
          Sum.post["bd4", "Mean"])
beta.u<-c(Sum.post["b0", "Mean"],
            Sum.post["bs1", "Mean"],
            Sum.post["bs2", "Mean"],
            Sum.post["bs3", "Mean"],
            Sum.post["br1", "Mean"],
            Sum.post["br2", "Mean"],
            Sum.post["br3", "Mean"])
            # these are the same as the posterior means;

### move to another directory ###
## dimension stuff for this year ##
dd<-c(paste("hyb2_",as.character(year.cur), sep=""))
setwd(dd)
dat.c<-dget(paste("data", as.character(year.cur), sep=""))
n<-dim(dat.c[[1]])[1] # of countries in MCMC for this year.
wg.names.p<-as.character(rownames(dat.c[[3]]))
sr<-matrix(0, nrow=n, ncol=2)
rownames(sr)<-wg.names.p
for (i in 1:m){
    for (j in 1:n){if (rownames(ab)[i]==rownames(sr)[j]){sr[j,]<-ab[i,]}
                }}
s<-sr[,1]; r<-sr[,2]

z<-matrix(0, nrow=n, ncol=k) # makes these starting at 0s;
rownames(z)<-wg.names.p

priors<-list(pi.Sab=pi.Sab,pi.s2u=pi.s2u,pi.s2v=pi.s2v,
                  pi.s2z=pi.s2z,
                  pim.bd=pim.bd,piS.bd=piS.bd,
                  pim.b0sr=pim.b0sr,piS.b0sr=piS.b0sr,
                  pim.b0s=pim.b0s, piS.b0s=piS.b0s , pi.s2a=pi.s2a)
startv<-list(beta.d=beta.d,beta.u=beta.u,s=s,r=r,z=z)

Yd<-dat.c[[1]]
n<-dim(Yd)[1]

Xd<-dat.c[[2]]
Xs<-Xr<-dat.c[[3]]
Y<-Yd
```

```
source("gbme.r")

# call the main estimation package, gbme
gbme(Y=Yd,Xd=Xd,Xs=Xs,Xr=Xr,k=3,fam="binomial",N=matrix(1,n,n),
     NS=300000,awrite=T,bwrite=T, estart=F, directed=T)
```

The starting values for the latent space vectors were initialized to zeros. They can also be estimated from posterior distributions. The former gives much better MCMC convergence than the latter, which we learned after numerous trials. Figure 3 displays the MCMC for the dyadic covariate coefficients for 1950.

Figure 3: *MCMC for 1950 with priors and starting values drawn from posteriors of 2000.*



3 Archive

The programs and the data are all available on the replication archive (http://dvn.iq.harvard.edu/dvn/dv/mward) for the publication. The minor differences in replication are attributable to be the random aspects of the estimation. We should have set the seed used in the MCMC implementation to a known value, for replication purposes.

3.1 Notes on 2008 replication folders

3.1.1 Directory `Rep_2000_others`

Using the posteriors from the 2007 results for the year 2000 found in `OUT.base, A.base, B.base` these runs replicate analysis for 1950 (Figure 3 above), 1955, and 2000 (Figure 1 (b) above). There are two runs replicating 2000. The first is in `hyb2_2000` which uses 200,000 iterations; another MCMC of 500,000 iterations is found in `hyb_2000`.

3.1.2 Directory `uninform5k_2000`

This directory focuses on replication of the base year, 2000, sing uninformative priors and starting values/ we estimate the year 2000 for 500,000 iterations and the posteriors are reported in Figure 2 (a) (above). Using these results we also initiated another run for the same year; the results are in the directory `hyb_2000` and reported in Figure 2 (b) above.

3.2 Some Guidance on Replication

First, we estimate the year 2000 using uninformative priors and starting values. The code is in `routine_uninform.R` in directory `uninform5k_2000`.

Second, for other years, we use the output from 2000 and draw priors and starting values from `OUT`, `A`, and `B` of 2000. The code is in `Replicate_routine_hybrid.R` in `Rep_2000_others`. Notice that in this directory `OUT`, `A`, and `B` for the 2000 run are renamed `OUT.base`, `A.base`, and `B.base` to signify that these are the base inputs for all years.

To initiate MCMC for other years, say, 1950, we create a directory in `Rep_2000_others` to store the results, e.g., directory `hyb2_1950`. This should contain the necessary data (e.g., `data1950`) as well as the code (`gbme.r`). Make sure to set the directory at the level of directory `Rep_2000_others` and run $\mathcal{R}$ code `Replicate_routine_hybrid.R`. This code reads from `OUT.base`, `A.base`, and `B.base` in `Rep_2000_others` and goes automatically to `hyb2_1950`. The results for year 1950 will be stored in `hyb2_1950`. A similar strategy applies to 1955 (shown in the files) and all other years.

References

Ward, Michael D., Randolph M. Siverson and Xun Cao. 2007. "Disputes, Democracies, and Dependencies: A Re-examination of the Kantian Peace." *American Journal of Political Science* 51(3):583–601.