

Bayesian Multivariate Prior for Multiple Linear Regression

Overview

This example uses the MCMC procedure to fit a Bayesian multiple linear regression (MLR) model by using a multivariate prior on the regression parameters. It demonstrates how to use existing SAS multivariate density functions for specifying prior distributions. Finally, it demonstrates how to create new density functions by using the FCMP procedure and by using new density functions in the MCMC procedure.

The SAS source code for this example is available as an attachment in the text file. In Adobe Acrobat, right-click the icon in the margin and select **Save Embedded File to Disk**. You can also double-click to open the file immediately.

[source code](#)

Analysis

Researchers are interested in determining the relationship of gestational length and litter size on brain weight after accounting for body weight. They study the historical benefits in evolution of larger brain weight after accounting for the coinciding disadvantages of longer gestational length and smaller litter size.

The following data set contains the average brain and body weights in grams and kilograms, respectively, the average gestational length in days, and the litter size for 95 different species as given in Sacher and Staffeldt (1974).

```
data brainweight;
  input brain body gestation litter @@;
  log_brain = log(brain); log_body = log(body);
  log_gestation = log(gestation);
  datalines;
17.5 3.5 26 1
3.5 0.93 34 4.6
3.15 0.15 46 3
1.14 0.049 51 1.5
1.37 0.064 46 1.5
22 2.1 135 1
12.8 1.2 90 1.2
```

```

... more lines ...

  93    13 120   1
 200    39 180   1
 210    66 158  1.2
 125    49 150  2.4
 106    30 151   2
;

```

Figure 1 displays summary statistics for the BRAINWEIGHT data set provided by the MEANS procedure. The differing orders of magnitude for the minimum and maximum gestational length, brain weight and body weight suggest the need for a log transformation.

Figure 1 PROC MEANS Summary

The MEANS Procedure				
Variable	Mean	Std Dev	Minimum	Maximum
brain	174.124	254.097	0.450	1600.000
body	79.994	178.268	0.017	1400.000
gestation	145.958	95.261	16.000	390.000
litter	2.324	1.751	1.000	8.000

Bayesian MLR Model

Suppose you want to fit a Bayesian MLR model for the logarithm of brain weight with density as

$$\begin{aligned}
 \log(\text{BRAIN})_i &\sim \text{normal}(\mu_i, \sigma^2) \\
 \mu_i &= \mathbf{X}_i \boldsymbol{\beta}
 \end{aligned}
 \tag{1}$$

where \mathbf{X}_i is the vector of covariates listed as $\mathbf{X}_i = \{1 \log(\text{BODY})_i \log(\text{GESTATION})_i \text{LITTER}_i\}$ for $i = 1, \dots, n$ species. The African elephant has been omitted from this example because it is easily recognized as an extreme outlier.

The likelihood function for the logarithm of the brain weight and corresponding covariates is

$$p(\log(\text{BRAIN})_i | \mathbf{X}_i, \boldsymbol{\beta}) = \text{normal}(\mu_i, \sigma^2)
 \tag{2}$$

where $p(\cdot|\cdot)$ denotes a conditional probability density. The normal density is evaluated at the specified value of $\log(\text{BRAIN})_i$ and the corresponding mean parameter μ_i defined in Equation 1. The four regression parameters in the likelihood are β_0 through β_3 .

Suppose you had expert or prior knowledge that some of the covariates were correlated. You might want to use a multivariate prior to incorporate your information. Using the multivariate normal prior, you enable covariates to be independent or correlated a priori. You can also specify the a priori correlation that you believe to be positive or negative. Suppose you thought, a priori, that

body weight and gestational age were positively correlated and that body weight and litter size were negatively correlated. Calculate the prior covariance as the product of the prior correlation and standard deviations. More formally, the formula for calculating the covariance of the j and k covariate is $\sigma_{jk} = \rho_{jk}s_j s_k$ where ρ_{jk} is the prior correlation and s_j and s_k are the sample standard deviations of the j th and k th covariates, respectively.

Suppose the prior correlation between the first and second covariate was $\rho_{12} = 0.5$, and a priori you thought the standard deviations of log of body weight and gestational age were $s_1 = 4$ and $s_2 = 1.5$, respectively. The prior covariance of the log of body weight and gestational age can be calculated as $\sigma_{12} = \rho_{12}s_1s_2 = (0.5)(4)(1.5) = 3$. Similarly, suppose the prior correlation between the first and third covariate was $\rho_{13} = -0.5$ and a priori the standard deviation of the litter size was $s_3 = 1$. The prior covariance of the log of body weight and litter size can be calculated as $\sigma_{13} = \rho_{13}s_1s_3 = (-0.5)(4)(1) = -2$.

Suppose the following prior distributions are placed on the parameters:

$$\pi(\boldsymbol{\beta}^*) = \text{MVN}_4(\boldsymbol{\mu}_{pr} = \mathbf{0}, \Sigma) \quad (3)$$

$$\Sigma = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 4 & -3 \\ 0 & 4 & 4 & 0 \\ 0 & -3 & 0 & 4 \end{bmatrix}$$

$$\pi(\sigma^2) = f_{i\Gamma}(\text{shape} = 2.001, \text{scale} = 1.001) \quad (4)$$

where $\pi(\cdot)$ indicates a prior distribution, $\text{MVN}_4(\boldsymbol{\mu}_{pr}, \Sigma)$ indicates a multivariate normal prior of dimension four with mean vector $\boldsymbol{\mu}_{pr}$ and variance matrix Σ , and $f_{i\Gamma}$ is the density function for the inverse-gamma distribution.

Using Bayes' theorem, the likelihood function and prior distributions determine the posterior distribution of the parameters as follows:

$$\pi(\boldsymbol{\beta}, \sigma^2 | \log(\text{BRAIN})_i, \mathbf{X}_i) \propto p(\log(\text{BRAIN})_i | \mathbf{X}_i, \boldsymbol{\beta}) \pi(\boldsymbol{\beta}) \pi(\sigma^2)$$

PROC MCMC obtains samples from the desired posterior distribution. You do not need to specify the exact form of the posterior distribution.

The following SAS statements use the likelihood function and prior distributions to fit the Bayesian MLR model with the multivariate prior. The PROC MCMC statement invokes the procedure and specifies the input data set. The NBI= option specifies the number of burn-in iterations. The NMC= option specifies the number of posterior simulation iterations. The THIN=5 option specifies that one of every five samples is saved in the posterior sample. The SEED= option specifies a seed for the random number generator (the seed guarantees the reproducibility of the random stream). The PROPCOV=QUANEW option uses the estimated inverse Hessian matrix as the initial proposal covariance matrix.

```
ods graphics on;
proc mcmc data=brainweight nbi=5000 nmc=25000 thin=5 seed=1181
  propcov=quanew;
  array mu_pr[4] mu_pr0-mu_pr3;
  array beta[4] beta0-beta3;
  array data[4] 1 log_body log_gestation litter;
```

```

begincnst;
  call zeromatrix(mu_pr);
  rc = logmpdfset('lp',4,0,4,0,3,4,0,-2,0,4);
endcnst;

parms beta: 0;
parms sig2 1;

beginprior;
  lmvn = logmpdfnormal(of beta0-beta3, of mu_pr0-mu_pr3, 'lp');
  prior beta: ~ general(lmvn);
  prior sig2 ~ igamma(shape = 2.001, scale = 1.001);
endprior;

call mult(beta, data, mu);
model log_brain ~ n(mu, var = sig2);
run;

data _null_;
  rc = logmpdffree();
  put rc=;
run;
ods graphics off;

```

Each of the two ARRAY statements associates a name with a list of variables and constants. The first ARRAY statement is used to specify the prior mean. The second ARRAY statement specifies names for the regression coefficients. The third ARRAY statement contains all of the covariates and is used to take advantage of PROC MCMC's ability to use matrix functions that make the code succinct.

The BEGINCNST and ENDCNST statements are used in conjunction with the DATA step functions. Programming statements enclosed by the BEGINCNST and ENDCNST statements are executed once per procedure call. You use the LOGMPDFSET function within the statements to set up a symmetric positive definite matrix from its lower triangular elements listed in row-major format. The RC assignment statement is set to 0 when the numeric arguments describe a positive definite matrix; otherwise it is set to a nonzero value.

The first PARMS statement places all regression parameters in a single block and assigns them an initial value of 0. The second PARMS statement places the variance parameter in a separate block and assigns it an initial value of 1.

The BEGINPRIOR and ENDPRIOR statements reduce unnecessary observation-level computations. The statements inside the BEGINPRIOR and ENDPRIOR statements block are not executed for each observation at each iteration. This enables a quick update of the symbols enclosed in the statements. The LMVN assignment statement uses the LOGMPDFNORMAL function to calculate the log density of the multivariate normal density given in Equation 3. Next, the PRIOR statement for the β parameters uses the GENERAL function, which indicates that you are using a SAS statement to construct a nonstandard density or distribution. The argument is an expression that takes the value of the logarithm function of the prior or likelihood distribution. In this case, the nonstandard density is the multivariate normal distribution and the argument is the logarithm of its density. The second PRIOR statement assigns the inverse-gamma prior distribution to σ^2 as given in Equation 4.

The CALL statement uses the MULT matrix multiplication subroutine to calculate μ_i . The MODEL statement specifies the likelihood function as given in Equation 2.

The DATA step after the call to the MCMC procedure calls the LOGMPDFFREE function. Use the LOGMPDFFREE function to free the workspace previously allocated with the LOGMPDFSET function. When called without arguments, the LOGMPDFFREE frees all the symbols previously allocated by LOGMPDFSETSQ or LOGMPDFSET. Each freed symbol is reported in the SAS log.

Figure 2 displays convergence diagnostic graphs to assess whether the Markov chain has converged. The trace plot indicates that the chain appears to have reached a stationary distribution. The chain also has good mixing and is dense.

The autocorrelation plot indicates low autocorrelation and efficient sampling. Finally, the kernel density plot shows the smooth, unimodal shape of posterior marginal distribution for β_0 . In a similar fashion, the rest of the diagnostic plots should be examined to ensure convergence of all parameters.

Figure 2 Diagnostic Plots for β_1

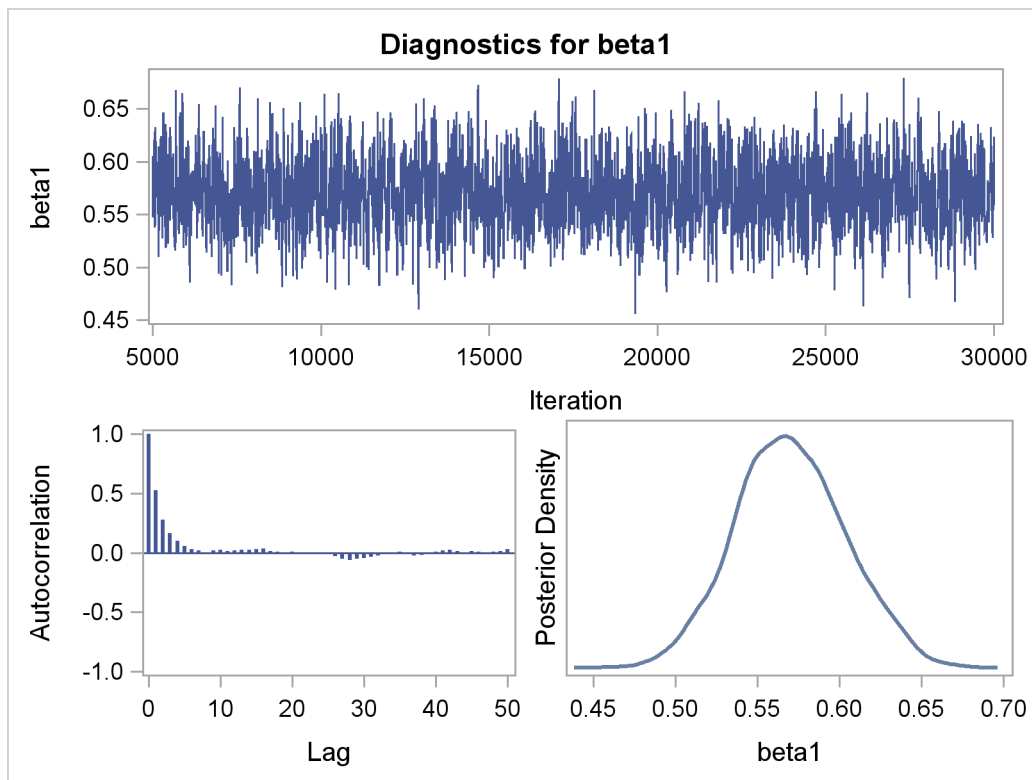


Figure 3 reports summary and interval statistics for each regression parameter and the variance parameter. The brain weight increases by a factor of $\exp(\beta_2) = \exp(0.4527) = 1.5725$ (approximately 57%) for each logarithmic change of one day of gestational age. Similarly, the brain weight decreases by $1 - \exp(\beta_3) = 1 - \exp(-0.1018) = 0.1018$ (approximately 10%) for each addition to the litter size.

Figure 3 Posterior Model Summary of Bayesian MLR

The MCMC Procedure						
Posterior Summaries						
Parameter	N	Mean	Standard Deviation	Percentiles		
				25%	50%	75%
beta0	5000	0.7583	0.6518	0.3168	0.7525	1.1771
beta1	5000	0.5703	0.0333	0.5474	0.5688	0.5924
beta2	5000	0.4527	0.1352	0.3654	0.4546	0.5445
beta3	5000	-0.1074	0.0427	-0.1363	-0.1074	-0.0780
sig2	5000	0.2434	0.0354	0.2185	0.2405	0.2649

Posterior Intervals					
Parameter	Alpha	Equal-Tail Interval		HPD Interval	
beta0	0.050	-0.4789	2.0902	-0.4861	2.0695
beta1	0.050	0.5072	0.6371	0.5064	0.6356
beta2	0.050	0.1744	0.7064	0.1890	0.7180
beta3	0.050	-0.1934	-0.0244	-0.1898	-0.0226
sig2	0.050	0.1840	0.3227	0.1793	0.3157

Create Your Own Density Function

Suppose you want to create your own density function for use as a prior distribution or density function in the likelihood. Use the FCMP procedure to create the density function and then call the function in the MCMC procedure. The FCMP procedure is part of the Base SAS package. (See “The FCMP Procedure” in the *Base SAS Procedures Guide*.)

The following SAS statements create a function called LPDFMVNORMN which uses the FCMP procedure to calculate the density of the multivariate normal distribution.

```
proc fcmp outlib=sasuser.funcs.myfun;
  function lpdfmvnormn(x[*],mu[*],sig[*,*]);
  array diff[1] / nosym;
  array sigi[1] / nosym;

  n = dim(x);
  call dynamic_array(diff,n);
  call dynamic_array(sigi,n,n);

  call det(sig,d);
  call subtractmatrix(x,mu,diff);
  call inv(sig,sigi);
  sum=0;
  do i=1 to n;
    do j=1 to n;
      sum = sum + diff[i]*diff[j]*sigi[i,j];
    end;
  end;
endfunction;
```

```

        end;
    end;

    logden = -0.5*(n*log(2*constant('pi')) + log(d) + sum);
    return(logden);
endsub;
run;

```

The PROC FCMP statement invokes the procedure. The OUTLIB= option specifies the name of an output data set to which compiled subroutines and functions are written. The FUNCTION statement specifies a subroutine declaration for a routine that returns a value. You specify the name of the function as LPDFMVNORMN, which takes three array-input arguments, among which X and MU are one-dimensional arrays and COV is a two-dimensional array.

The two ARRAY statements create two temporary arrays that are used in the calculation of the multivariate normal density. A temporary array does not have associated element variables, and its dimensions can be modified arbitrarily using the CALL to the DYNAMIC_ARRAY subroutine. The N assignment statement returns the length of the X array and the dimension of the multivariate normal density. The two DYNAMIC_ARRAY statements allocate appropriate dimension and size for arrays DIFF and SIGI. The rest of the programming statements calculate the logarithm density of the multivariate normal distribution and return the value.

To use the LPDFMVNORMN function in the MCMC procedure, you need the OPTIONS statement to designate which CMP library contains the FCMP function. Note that when you use your own defined function to calculate the density of the multivariate normal density, you do not need the LOGMPDFSET function or the extra DATA step after the MCMC procedure to free the memory.

The following statements invoke the MCMC procedure and call the LPDFMVNORMN function you defined in the FCMP procedure.

```

options cmplib=sasuser.funcs;

proc mcmc data=brainweight nbi=5000 nmc=25000 thin=5 seed=1181
    propcov=quanew;
    array data[4] 1 log_body log_gestation litter;
    array beta[4] beta0-beta3;
    array sigma[4,4];
    array pm[4];

    begincnst;
        call zeromatrix(pm);
        call identity(sigma);
        call mult(sigma, 4, sigma);
        sigma[3,2] = 3; sigma[2,3] = 3;
        sigma[4,2] = -2; sigma[2,4] = -2;
    endcnst;

    parms beta: 0;
    parms sig2 1;

    beginprior;
        lmvn = lpdfmvnormn(beta, pm, sigma);
        prior beta: ~ general(lmvn);
    endprior;
run;

```

```

    prior sig2 ~ igamma(shape = 2.001, scale = 1.001);
endprior;

call mult(beta, data, mu);
model log_brain ~ n(mu, var = sig2);
run;

```

The first two ARRAY statements stay the same as in the previous call, but now two additional ARRAY statements are added for the prior mean and variance hyperparameters. The BEGINCNST/ENDCNST statements declare constants in the program. In this call to the MCMC procedure, the CALL to the IDENTITY subroutine fills in the prior covariance matrix with the identity matrix. The CALL to the MULT subroutine designates 4 on the diagonal instead of 1. The last four statements designate the prior covariances to incorporate your prior correlation.

The PARMs statements do not change in this call to the MCMC procedure. Now the BEGINPRIOR/ENDPRIOR block contains the new LPDFMVNORMN function. The LMVN assignment statement assigns the logarithm of the multivariate normal density computed with the LPDFMVNORMN function. Now the PRIOR statement for the regression parameters uses the GENERAL function, and its argument is LMVN. The PRIOR statement for σ^2 , the CALL statement to the MULT subroutine, and MODEL statements remain unchanged from the previous call to the MCMC procedure.

References

- Sacher, G. A. and Staffeldt, E. F. (1974), "Relation of Gestation Time to Brain Weight for Placental Mammals: Implications for the Theory of Vertebrate Growth," *The American Naturalist*, 108(963), 593–615.