Bayesian Analysis for Categorical Data with Missing Traits Under a Multivariate Threshold Animal Model

D. H. Lee

Dept. of Animal Life and Resources, Hankyong National University

다형질 Threshold 개체모형에서 Missing 기록을 포함한 이산형 자료에 대한 Bayesian 분석

이 득 환

한경대학교 동물생명자원학과

ABSTRACT

Genetic variance and covariance components of the linear traits and the ordered categorical traits, that are usually observed as dichotomous or polychotomous outcomes, were simultaneously estimated in a multivariate threshold animal model with concepts of arbitrary underlying liability scales with Bayesian inference via Gibbs sampling algorithms. A multivariate threshold animal model in this study can be allowed in any combination of missing traits with assuming correlation among the traits considered. Gibbs sampling algorithms as a hierarchical Bayesian inference were used to get reliable point estimates to which marginal posterior means of parameters were assumed. Main point of this study is that the underlying values for the observations on the categorical traits sampled at previous round of iteration and the observations on the continuous traits can be considered to sample the underlying values for categorical data and continuous data with missing at current cycle (see appendix). This study also showed that the underlying variables for missing categorical data should be generated with taking into account for the correlated traits to satisfy the fully conditional posterior distributions of parameters although some of papers (Wang et al., 1997; VanTassell et al., 1998) presented that only the residual effects of missing traits were generated in same situation.

In present study, Gibbs samplers for making the fully Bayesian inferences for unknown parameters of interests are played rolls with methodologies to enable the any combinations of the linear and categorical traits with missing observations. Moreover, two kinds of constraints to guarantee identifiability for the arbitrary underlying variables are shown with keeping the fully conditional posterior distributions of those parameters. Numerical example for a threshold animal model included the maternal and permanent environmental effects on a multiple ordered categorical trait as calving ease, a binary trait as non-return rate, and the other normally distributed trait, birth weight, is provided with simulation study.
(**Key words** : Threshold animal model, Missing trait, Bayesian inference, Liability, Identifiability)

I. INTRODUCTION

The variance and covariance components for genetic evaluation have been main issues in animal breeding since quantitative genetics were adopted in this field at early 20 century.

Corresponding author : D. H. Lee, Department of animal life and Resources, Hankyong National University 67 Scokjeong-dong, Ansung, Kyonggi-do, 456-749, Korea, email:dhlee@hnu.hankyong.ac.kr)

Furthermore, these genetic variations can be used to estimate genetic merits for target to improve by selecting the superior animals in quantitative genetics. The genetic variance and covariance components or functions of these components are often estimated with assuming a linear model on which the dependent variables are distributed as the continuous linear combinations with certain relevant assumptions. However, the observations in a linear model are often measured in a discrete scale, and hence the usual methods for variance component estimation under assumption of a normal distribution are not appropriate. For example, in American Gelbvieh Association, calving ease is scored as 1 (natural calving, no assistance), 2 (easy pull), 3 (hard pull) or 4 (mechanical force or Cesarean). In addition, some of field data for animal breeding can be observed as the dichotomous outcomes like pregnancy or not after service (non-return rate). Because calving ease or non-return rate is an important trait to determine the economic loss for heifer and cow in reproduction, the various methodologies have been devised to analyzing those traits. One of the appealing methodologies for these traits is based on threshold liability concept, which was originated by Wright (1934) in studies of the number of digits in guinea pigs. In his threshold model, it is postulated that there exists a latent or underlying variable (liability), which has a continuous distribution. A set of thresholds divides this continuous variable into the discrete scores. Thus an observed value of categorical trait is a representation of the liability which is fallen in between two thresholds. Several applications of this model can be found in Gianola (1982), Foulley et al. (1987), Wang et al (1994), Sorensen et al. (1994), Jensen et al (1994) and Berger et al. (1995).

With the advent of inferential algorithms based simulations, Bayesian methods are being increasingly applied to genetic inference in animal breeding. This is partly due to the fact that the complex analytic solutions of Bayesian method are now feasible with help of the inferential algorithms. The most popular family of such algorithms is Markov Chain Monte Carlo (MCMC) including Gibbs sampling. On the other hand, a certain genetic inference requires a super population model or Bayes model. In this sense, the number of parameters to be estimated is greater than or equal to the number of observations, which is often arising in animal breeding. Gianola and Foulley (1983) described a Bayesian approach in a single trait threshold model assuming known genetic variance. Foulley et al. (1983) developed a method to deal with a binary trait and two continuous traits without allowing for missing data. This result was generalized to the situation of one multiple ordered categorical trait and several continuous traits. Sorensen et al. (1995) published an applied methodology of Bayesian analysis via Gibbs sampling in an univariate threshold model. Their methodology was further extended by Wang et al. (1997) to allow for one multiple ordered categorical trait and one continuous trait with missing. Lee et al. (2002) presented the bias problems that were partly affected by reparameterization for the dispersion parameters especially on a binary trait. However, we will note that the number of traits does not play much role in the analytic models as long as the discrete traits have the multiple ordered categories. In this note, we will also describe a general hierarchical Bayes threshold animal model via Gibbs sampling for analysis of any number and kind of the categorical traits with the continuous traits jointly. Specially, some methodologies for generating the arbitrary underlying values without violating the fully conditional posterior distributions will be shown and parameterization to identify the underlying

variables will also be shown with respect of Bayesian prospective in case of having the missing traits and a multi-trait animal model. As example, some results for estimation on one linear, one 4-categoircal trait, and one binary trait in simulated data with different proportion of missing records will be provided under a multi-trait threshold animal model with several random effects including the maternal genetic effects and permanent environmental effects as well as the direct genetic effects.

Ⅱ. METHODOLOGY

1. Model

As implementing Bayesian theorem of a multivariate threshold animal model, Gibbs sampling algorithms (Geman and Geman, 1984) as a hierarchical Bayesian approach (Box and Tiao, 1992, p. 58) were carried out. We will describe statistical notations for the models in this study to a general case for convenience. Let $y_1, y_2, \cdots, y_k$ be the $n\times 1$ vectors of the continuous or ordered categorical observations. We assumed that first $m(\geq 1)$ traits are the multiple ordered categorical observations and the expression of the underlying continuous random vectors are $u_1, \cdots, u_m$, respectively with the unknown threshold values $t_1, \cdots, t_m$ where $t_i = (t_{i1}, t_{i2}, \cdots, t_{iC-1})'$ satisfying $-\infty = t_{i0} < t_{i1} < \cdots < t_{iC-1} < t_{iC} = \infty$ for $i=1,\cdots,m$. That is, if we denote the j^{th} element of u_i as u_{ij}, then the categorical observation y_{ij} is recorded by

$y_{ij} = \lambda$, if $t_{i\lambda-1} < u_{ij} \leq t_{i\lambda}$ for $i=1,\cdots,m$ and $\lambda = 1,\cdots C$.

Some of the categorical and/or continuous observations y_{ij}'s can be missing in field data. Thus we should allow some missing values in the records. The observed random vectors will be denoted by y_i^0, $i=1,\cdots,k$. We will also use $(y_{m+1},\cdots,y_k)$ and $(u_{m+1},\cdots,u_k)$ interchangeably for notational convenience. Now we consider the following model for calving ease and non-return rate with representing the multiple categorical and binary traits, respectively, as well as birth weight as a continuous trait:

$$\mathrm{u}_i = X\beta_i + Z_h h_i + Z_d a_i + Z_m m_i + Z_p p_i + e_i,$$
$$\text{for } i=1,\cdots,k \qquad (1)$$

where β_i is the fixed effect associated with sex or age of dam; h_i is the random herd-year-season effect; $a_i\,(m_i)$ is the direct (maternal) additive genetic effect; and p_i is the maternal permanent environmental effect. It is assumed that $W = (X, Z_h, Z_d, Z_m, Z_p)$ is a known matrix associated with β_i, h_i, a_i, m_i and p_i, respectively. X has full column rank. Let $\theta_i = (\beta'_i, h'_i, a'_i, m'_i, p'_i)'$ for i=1,$\cdots$,k and $\theta = (\theta'_1, \cdots, \theta'_n)$. Then a hierarchical Bayes model for model (1) might be set up with conditional on θ and R having a certain joint prior distribution whether proper or improper

$$U = \begin{pmatrix} \mathrm{u}_1 \\ \vdots \\ \mathrm{u}_k \end{pmatrix} \sim N\left(\begin{pmatrix} W\theta_1 \\ \vdots \\ W\theta_k \end{pmatrix}, R\otimes \mathrm{I}_n \right) \qquad (2)$$

Where R is the $k\times k$ variance-covariance matrix of $\mathbf{u}$.

In what follow, we will use a minus subscript to delete an appropriate element from a matrix or a vector. For example, u_{-i} will be denoted the matrix $\mathbf{u}$ with $\mathbf{u}_i$ deleted. Similarly u_{-ij} be denoted the vector u_i with u_{ij} deleted. The joint distributions of the underlying values for categorical traits and the observations for continuous traits on (2) given the location parameters can be stated in terms of the residuals because there is one-to-one relationship with u_{ij}. This can be denoted as $e_{ij} = u_{ij} - w_{ij}\theta$ when $e_{.j} \sim N(0,\mathrm{R})$.

2. Liability without missing

The posterior distribution of the underlying scale for a categorical trait given the correlated traits without missing was well defined as:

$$\Pr[y_{ij} = \lambda \mid u_{-ij}, \theta, R, t_i] = \Pr[t_{i\lambda-1} < u_{ij} \le t_{i\lambda} \mid u_{-ij}, \theta, R, t_i]$$
$$= \Phi\left(\frac{t_{i\lambda} - \xi_{ij}}{\sigma_{i\cdot}}\right) - \Phi\left(\frac{t_{i\lambda-1} - \xi_{ij}}{\sigma_{i\cdot}}\right)$$

Now, under (2), it is easy to check that, given θ, R, t, u_{-ij}, and y^0 on a categorical trait, the conditional density of u_{ij} is given by:

$$f(u_{ij} \mid y_{ij} = \lambda, u_{-ij}, \theta, R, t) = \frac{\phi\left(\frac{u_{ij} - \xi_{ij}}{\sigma_{i\cdot}}\right)}{\Phi\left(\frac{t_{i\lambda} - \xi_{ij}}{\sigma_{i\cdot}}\right) - \Phi\left(\frac{t_{i\lambda-1} - \xi_{ij}}{\sigma_{i\cdot}}\right)} 1(u_{ij} \in [t_{i\lambda-1}, t_{i\lambda}]),$$

and $u_{ij} \mid y_{ij} = \lambda, u_{-ij}, \theta, R, t \sim TN_{t_{i\lambda-1}, t_{i\lambda}}(\xi_{ij}, \sigma_{i\cdot}^2)$ (3)

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and distribution functions of a standard normal distribution; ***TN*** is a truncated normal distribution between the truncated points $t_{i\lambda-1}$, $t_{i\lambda}$; $1(\cdot)$ is an indicator function. On the other hand, if y_i is a continuous trait, then $u_{ij} \mid \theta$, R, t, u_{-ij}, y^0 has a degenerating distribution at y_{ij} because $y_i \equiv u_{i\cdot}$; $\xi_{ij} = w_{ij}'\theta_i + R_{i(-i)}R_{(-i)(-i)}^{-1}e_{-ij}$; $\sigma_{i\cdot}^2 = r_{ii} - R_{i(-i)} \cdot R_{(-i)(-i)}^{-1}R_{(-i)i}$ where

$R = \begin{bmatrix} r_{ii} & R_{i(-i)} \\ R_{(-i)i} & R_{(-i)(-i)} \end{bmatrix}$ partitioned with the i^{th} row and column.

3. Liability with missing

However, we can draw the missing traits from a (multi) normal distribution as accounting for the correlated traits as described by Sorensen (1996). One obstacle for drawing the missing is that we could not guarantee the prior distributions of the location parameters and the residual variances because of weak information of the missing traits. Wang et al.(1997) and VanTassell et al.(1998) tried to generate the missing residuals rather than the missing observations. But their Gibbs samplers for these parameters were not consistent with a fully conditional posterior distribution as Bayesian inference. Li and Lee (2002) tried to generate the underling values for the categorical traits with missing given all the other priors (underlying values on previous round, dispersion and location parameters). In this case, the improper priors for uncertainty have a risk to lead the improper posterior distributions. We got the improper posterior distributions for the location parameters as well as the dispersion parameters as heuristic approach with algorithms by Li and Lee (2002) in a simulated data with over 50% missing. So in this study, we try to draw the missing values from the fully conditional posterior distributions with modifying equation (3) as follow:

$$u_{mi} \mid u_{oj}, \theta, R \sim N(\xi_{mj}, \sigma_{m\cdot}^2) \quad (4)$$

where u_{mj} is a vector of the missing traits on jth record; u_{oj} is a vector of the observed traits;

$\xi_{mj} = w_j'\theta_m + R_{mo}R_{oo}^{-1}e_{oj}$; $\sigma_{m\cdot}^2 = R_{mm} - R_{mo}R_{oo}^{-1}R_{om}$,

where $R = \begin{bmatrix} R_{mm} & R_{mo} \\ R_{om} & R_{oo} \end{bmatrix}$ with corresponding with the missing and observed traits, respectively. Moreover, we also need to take into account for the observed values of categorical traits. If a categorical trait were observed, the underlying scale for this would be generated from a truncated normal distribution. In this case, we can only consider the correlated traits with having the observation because of taking a proper prior distribution. If concerning this problem, the underlying values for categorical traits (whether observed or missing) and the continuous traits with missing can be generated with step by step (appendix).

In next step, we are concerning identifiability

for the arbitrary underlying variables. Sorensen et al. (1995) presented that two of constraints must be imposed to guarantee identifiability for the underlying scales of categorical traits. Residual variance of one and $t_{i,1} = const$ are usually restricted, especially, in cases of the dichotomous variables. In a multivariate threshold model, however, the marginal posterior distribution of the residual variance for categorical traits would be not one. One of reasons should be due to the correlated traits. Lee et al. (2002) presented that restriction for the residual variance of one should cause the biased estimates of the other variance components (e.g. genetic variance components), and the other way is restriction of a threshold and/or the underlying values. For this reason, we tried that the underlying scale for categorical trait given by the correlated traits was generated from a normal distribution with density as:

$$u_{ij}^{*} = \frac{u_{ij}}{\sqrt{r_{ii}}} \sim N\left(\frac{W\theta_i + r_{i(-i\omega)}R_{(-i\omega)(-i\omega)}^{-1}e_{-i\omega}}{\sqrt{r_{ii}}}, \frac{r_{ii} - r_{i(-i\omega)}R_{(-i\omega)(-i\omega)}^{-1}r_{(-i\omega)i}}{r_{ii}}\right) \quad (5)$$

For generating the other missing trait, u_{ij} would be assumed to know as described in appendix.

By equation (5), the scales of the underlying variables can be imposed approximately to the residual variance of one.

4. Thresholds

Samples for the thresholds can be modified the works by Sorensen et al. (1995) as follow:

The fully conditional posterior distribution of the each threshold is independent and is given by :

$$t_{i\lambda}^{n} \mid \theta, R, Q_h, G, Q_p, t_{-i\lambda}, U, y^{0} \sim UN\left(\max\left(\max_j\left(u_{ij}^{*} \mid y_{ij} = \lambda\right), t_{i\lambda-1}^{n}\right), \min\left(\min_j\left(u_{ij}^{*} \mid y_{ij} = \lambda+1\right), t_{i\lambda+1}^{n-1}\right)\right) \quad (6)$$

One of thresholds can be impose to guarantee identifiability of the location parameters. We impose one of thresholds for the categorical traits as follow:

$$t_{ij}^{*} = t_{ij} - t_{i1}, \text{ for } j = 1, \cdots, C_i - 1 \quad (7)$$

where C_i is the number of categories on the i^{th} categorical trait.

5. Location and dispersion parameters

As mentioned before, $\beta = \beta'_1, \cdots, \beta'_k)'$ is the vector of fixed effects while $h=(h'_1, \cdots, h'_k)'$, $\theta_g = (a'_1, \cdots, a'_k, m'_1, \cdots, m'_k)'$, and $\theta_p = (p'_1, \cdots, p'_k)'$ are random. Traditional assumptions in Bayesian inference can be achieved by putting the prior distributions appropriately. We consider following some prior distributions:

(a) $f(\beta) \propto const$

(b) $f(h \mid Q_h) \sim N(0, Q_h \otimes I)$ where Q_h is a $k \times k$ variance-covariance matrix of $h=(h'_1, \cdots, h'_k)'$.

(c) $f(a\ m \mid G) \sim N(0, G \otimes A)$ where G is a $2k \times 2k$ variance-covariance matrix of $\theta_g = (a'_1, \cdots, a'_k, m'_1, \cdots, m'_k)'$ and A is a known numerator relationship matrix.

(d) $f(p \mid Q_p) \sim N(0, Q_p \otimes I)$ where Q_p is a $k \times k$ variance-covariance matrix of $\theta_p = (p'_1, \cdots, p'_k)'$.

Note that the prior distributions belong to a conjugate family or a noninformative prior. Because Q_h, Q_p, G and R are unknown, we rely on the hierarchical Bayes procedure and put Jeffreys noninformative priors (Gelman et al., 1995) for those variance-covariance matrices.

(e) $f(R) \propto |R|^{-k/2}$

(f) $f(Q_h) \propto |Q_h|^{-k/2}$

(g) $f(G) \propto |G|^{-k}$

(h) $f(Q_p) \propto |Q_p|^{-k/2}$

According to the prior distribution of each parameter showed by (a) to (h) and the conditional distributions of the underlying

variables and the thresholds by equations (5) and (6), each unknown parameter given all the other priors can be sampled from the fully conditional posterior distribution of each parameter according to the distribution property because Bayesian ways of all expressing uncertainty are to draw density functions of all unknown (Blasco, 2001). The location parameters as assumed to the fixed effects on (a) can be sampled from the normal distributions with assuming the flat priors. The random effects given by (b) and (c) can be sampled from the multivariate normal distribution with certain means and covariances provided by VanTassell et al. (1998). The variance-covariance matrices are sampled from general inverted Wishart distribution with assuming the flat prior distributions due to noninformative priors rather than the conditional inverted Wishart distributions provided by Korsgaard et al. (1999).

6. Initial values

The initial values for the location parameters, especially, those of the missing traits, should significantly influence to the posterior distributions because of weak information. We assessed the posterior distributions of the location parameters given null initial values. In this case, Gibbs samples for the location parameters, especially, contemporary group effects in model (1) occasionally showed to blow up and heuristically the improper posterior distributions. According to this property, the dispersion parameters also blew up. One of possibility to solve this problem is that the initial values for missing traits can be replaced to overall mean and a few rounds of iteration can be cycled to adjust the location and dispersion parameters. This way can lead to a proper posterior distributions for the location parameters as well as the dispersion parameters.

III. NUMERICAL EXAMPLES

1. Simulated data

Data were simulated with the model (1). Each effect was generated from a normal distribution with density as:

$\beta_{SEX} \sim N(0, T_{SEX} \otimes I)$; $\beta_h \sim N(0, Q_h \otimes I)$; h, $(a\ m)$, and p were generated given by (b), (c) and (d), respectively; e was generated from the normal distribution of $N(0, R \otimes I)$.

Where the variance-covariance matrix for each parameter for three traits was assumed as:

$$T_{SEX} = diag[10 \quad 25 \quad 25],\ Q_h = diag[15 \quad 25 \quad 25]$$

$$Q_p = \begin{bmatrix} 6 & 2 & 2 \\ & 10 & 5 \\ symm & & 18 \end{bmatrix},\quad R = \begin{bmatrix} 45 & 27 & 20 \\ & 100 & 0 \\ symm & & 100 \end{bmatrix}$$

and the variance-covariance matrix for the additive genetic effects correlated with the maternal genetic effects was denoted as:

$$G = \begin{bmatrix} 50 & 40 & 35 & -10 & -2 & -3 \\ & 50 & 35 & -2 & -10 & -3 \\ & & 40 & -6 & -10 & -10 \\ & & & 10 & 6 & 8 \\ & symm & & & 20 & 10 \\ & & & & & 20 \end{bmatrix}$$

The base population consisted of 10 sires and 1000 dams. Mating within each generation was random, and 10% males and 90% females were randomly selected for each generation. A total of five generations was created with [200 10 10] herd groups for first trait (birth weight), second trait (calving ease), and third trait (non-return rate), respectively. We denoted final original (three continuous traits) dataset to DS1. The continuous values for calving ease and non-return rate of each record from DS1 were discretized to 4 and 2 categories using the thresholds of 2/3, 1.5 and 2 (for calving ease) and 1 (for non-return rate) on the standardized scale, respectively (NOMISS). 10% of records for each trait from dataset NOMISS were

randomly treated to the missing (MISS10). 50% of records for two categorical traits (4-categories and binary) from NOMISS were also randomly treated to the missing (MISS50).

2. Analysis

Three different data sets (NOMISS, MISS10, MISS50) were used to estimate the location and dispersion parameters by a three-trait threshold animal model using Gibbs sampling algorithms described above. Three continuous traits without the missing (DS1) were also used to estimate the same parameters by Gibbs sampling in a general three-trait animal model and the results were compared to the estimates of the variance components by REML algorithm with the same model and data set. We assumed the sex effects on the three traits to the fixed, the herd effect on first trait (birth weight) with 1000 levels to the random and the herd effects on the other two traits with 50 levels to the fixed. The permanent environmental effects on three traits were assumed to the random. Furthermore, the additive genetic effects correlated with the maternal additive genetic effects for three traits were assumed to be random as assuming on the simulation step. Finally, 11085 animals on which 10075 records were observed at least one trait were used for this study. The threshold or linear animal model was similar to equation (1) as:

$$y_L = X\beta_L + Z_h h_L + Z_d a_L + Z_m m_L + Z_{PE} p_L + e_L$$

for the linear trait,

$$y_4 = X\beta_4 + X_h h_4 + Z_d a_4 + Z_m m_4 + Z_{PE} p_4 + e_4$$

for the 4 categorical trait, and

$$y_2 = X\beta_2 + X_h h_2 + Z_d a_2 + Z_m m_2 + Z_{PE} p_2 + e_2$$

for the binary trait.

Where β, h, a, m, p and e are the sex, contemporary group, direct additive genetic, maternal additive genetic, permanent environmental, and the residual effects, respectively.

Gibbs samplers for all the unknown parameters on each data set was run with one chain of 35,000 rounds of iteration and every 10^{th} samples for the dispersion parameters and the thresholds were stored. The burn-in period was determined visually and discarded first 15,000 samples. Statistics computed for the last 20,000 samples included the posterior means, 95% confidence intervals, number of effective samples, lag length for the (co)variance components and the genetic parameters as described by Raftery and Lewis (1992). The posterior means of the breeding values as the direct and maternal genetic effects were calculated using the samples on every round after 1000 samples at starting. For justifying these methodologies heuristically as comparing the breeding values by BLUP to those by Gibbs sampling, the breeding values as frequentist approach were also estimated by the best linear unbiased prediction (BLUP) with a pre-conjugated gradient method. For estimating the breeding values by BLUP, it was assumed that the categorical observations were linear and the covariance components for all the random effects were the posterior means by Gibbs sampling in threshold model on the same data set. Correlations between the breeding values by threshold model (TM) and linear model (LM) and the breeding values by linear model with BLUP algorithm using original data were estimated.

Ⅳ. RESULTS AND DISCUSSIONS

1. Initial values

The prior distribution for each parameter was important because if the improper prior distribution especially for dispersion parameters might lead to the improper posterior distributions (Datta and Maiti, 1998). Datta and Maiti (1998) provided a set of necessary and sufficient

conditions for the posterior distributions corresponding to a certain class of improper priors on the variance components. The uniform prior density for genetic variance components was also not improper (Hobert and Casella, 1996). However, In present study with 50% missing records, when the initial values for missing categorical traits were assumed to global means, the posterior distributions for the location parameters would be proper as heuristic described before. This trend should be more sensitive for the continuous traits than the categorical traits even though it would be depended on a magnitude of scales. The user can supply initial covariance components. Gauss-Seidel iteration with the initial values for the variance components can also be used to calculate the initial values for the location parameters. However, for the categorical traits, the initial values for the location parameters and the thresholds can be used maximum *a priori* estimates (MAP) evaluated at the approximate marginal maximum likelihood estimate of the covariance components (Hoeschele and Tier, 1995) or an approximate estimate based on a linear model which was a score based on the cumulative distribution function (CDF) (Djemali et al., 1987). However, MAP estimates would dramatically increase computing demands and algorithm complexity for a general model and the other method provided by Djemali et al. (1987) was required to calculate CDF and inverted CDF to get the initial location parameters. Furthermore, these initial values do not much affect to the conditional posterior distributions. In this study, we calculate the location parameters using Gauss-Seidel iteration given the arbitrary sequential values for the thresholds and the initial covariance components. This would not cause to the improper posterior distributions for the location parameters.

2. Constraints

Constraints must be imposed to guarantee identifiability because the underlying scales for categorical traits are arbitrary. We expect that marginal posterior distribution of residual variance should not be one because of sampling error and noise with having distribution if standard parameterization referred by Sorensen et al. (1995) were imposed. Lee et al. (2002) assumed the prior of infinite threshold to 10 and sampled underlying values with a truncated normal distribution to reduce the biased estimates of variance components. However, they could not justify for fully conditional posterior distributions for the underlying variables. On this study, the results using Gibbs sampling algorithms with generating underlying variables by equation of (5) and (7) in threshold models were closed to results in linear model with original scales if concerned Monte-Carlo standard deviations (Table 1).

3. Underlying values

Even though the underlying values given the location parameters and the residual (co) variances on equation (4) were well defined, we restricted the underlying values to be generated from a normal distribution as showing on equation (5) without violating Bayesian inference. Furthermore, because the underlying variables for categorical traits should be reparameterized on equation (5) simultaneously, restriction for residual variances for adjusting scales of liabilities does not needed. The way to draw these dependent variables can be sampled by trait given the other traits that were observed or sampled at preceding step because of considering the correlated traits (Lee et al., 2001). Table 1 also showed that the fully conditional posterior distributions of underlying variables were well

Table 1. Estimates of residual (co)variances and correlations by REML using linear traits and those as posterior means by linear model (LM) using linear traits and by threshold model using categorical traits in three-trait animal models and four different simulated data sets

	REML	LM	NOMISS	MISS10	MISS50
Residual (co) variances					
$\sigma^2_{e(L)}$	34.310	33.810(0.091)	34.050(0.162)	34.090(0.128)	33.870(0.117)
$\sigma^2_{e(4)}$	88.290	88.160(0.334)	1.000(0.000)	1.000(0.000)	1.000(0.000)
$\sigma^2_{e(2)}$	90.330	89.690(0.194)	1.003(0.000)	1.003(0.000)	1.003(0.000)
$\sigma^2_{e(L-4)}$	20.030	19.730(0.138)	2.297(0.020)	2.351(0.049)	2.545(0.050)
$\sigma^2_{e(L-2)}$	17.270	16.760(0.114)	1.843(0.022)	1.779(0.038)	2.357(0.109)
Residual correlations					
Linear-4-cat.	0.36	0.36 (0.002)	0.39 (0.002)	0.40 (0.008)	0.44 (0.008)
Linear-Binary	0.31	0.30 (0.002)	0.32 (0.004)	0.30 (0.006)	0.40 (0.019)
Thresholds for 4-cat.					
t_2			1.134(0.002)	1.133(0.003)	1.170(0.009)
t_3			1.732(0.002)	1.721(0.005)	1.786(0.019)

Subscript L: linear trait, 4: 4-categorical trait, 2: binary trait.
t : thresholds for 4-categorical trait

defined although the covariance components between traits seemed to be overestimated with depending on the amount of the missing. For example, the posterior mean (PM) for residual covariance between the linear trait and the 4-categorical (binary) traits were 10%(27%) larger on data set of MISS50 than on data set of NOMISS (Table 1). Subsequently, residual correlations between linear and categorical traits seemed to be overestimated. These characteristics might be caused from loss the information of data due to missing as well as categorizing. According to these features, the fully conditional posterior means of the thresholds for 4-categorical trait were shown the difference by each data set. This result showed a trend that the larger the missing in data are included, the larger Monte-Carlo standard deviation(MCSD) are shown. The other reason we expect is due to distribution of categories. Because the frequency of each category were highly skewed, assumption of a normal distribution for the underlying values should be weak so that the threshold points going to be ambiguous. The other approach that we can use as assumption can be a model with assumption of Dirichlet distribution for parameters. Rekaya (2001) used property of Dirichlet distribution to figure out uncertainty for the dispersion parameters. But his basic model was assumed with a normal distribution. No others did assume this distribution in animal breeding.

4. Genetic parameter estimates

Table 2 showed the comparison of posterior means (PM) and Monte-Carlo standard deviation (MCSD) for heritabilities in the three-trait linear and threshold model with REML and Gibbs sampling using the simulated data. The posterior

Table 2. Heritability Estimates for three traits by REML using linear traits and those as posterior means by linear model (LM) using linear traits and by threshold model using categorical traits in three-trait animal models and four different simulated data sets

	REML	LM	NOMISS	MISS10	MISS50
Direct heritabilities					
Linear	0.23	0.24(0.001)	0.23(0.003)	0.22(0.002)	0.23(0.002)
4-cat.	0.20	0.20(0.004)	0.18(0.003)	0.18(0.006)	0.19(0.016)
Binary	0.09	0.10(0.002)	0.10(0.002)	0.11(0.009)	0.17(0.039)
Maternal heritabilities					
Linear	0.37	0.37(0.001)	0.39(0.002)	0.39(0.002)	0.40(0.002)
4-cat.	0.16	0.17(0.004)	0.16(0.004)	0.14(0.007)	0.18(0.009)
Binary	0.34	0.33(0.005)	0.26(0.019)	0.26(0.014)	0.30(0.013)

means of heritabilities using Gibbs sampling algorithms were similar to estimates using restricted maximum likelihood in three-trait linear model on original scales of observations. The posterior means of heritabilities for maternal effects on 4-categorical trait and binary trait would be close to the estimates for same effects by REML even though, on binary trait, the estimates for the direct genetic effects would be overestimated and contrariwise those for maternal genetic effects would be underestimated. These features would be due to relatively small Gibbs samples with considering relatively large number of the missing records. For example, the distribution of heritibility for the direct genetic effects on binary trait in simulated dataset of MISS50 showed the over-dispersion (h^2= 0.17) with comparing estimate by REML (0.09). The posterior distribution for this heritability was also shown relatively large MCSD (0.039). Some of reasons for these features might be caused from loss of information due to missing and categorized as describing before. The posterior means for genetic correlations between traits and effects in threshold models were similar to estimates by REML in linear models with original scales(Table 3). However, lag-length, which stands for the autocorrelation between adjacent samples, would increase according to the proportion of the missing records of each corresponding trait.

5. Location parameter estimates

Table 4 showed the correlation coefficients for the direct and maternal genetic effects between by the linear animal model (LM) on the original scales without the missing traits and by the threshold animal models (TM) with three different simulated data sets according to the missing observations. As shown in Table 4, if the estimates for the breeding values using BLUP in LM on the original scale would be represented for or close to the true values, the direct genetic effects in TM were higher correlated to the true values than those in LM on all data sets. For example, the direct genetic effects for the 4-categorical trait in TM were shown 4% higher correlated to true values than in LM. Furthermore, those for the binary trait were also shown 12% higher correlated to true values than in LM when using no missing traits.

Table 3. Genetic correlation Estimates for three traits by REML using linear traits and those as posterior means by linear model (LM) using linear traits and by threshold model using categorical traits in three-trait animal models and four different simulated data sets

	REML	LM	NOMISS	MISS10	MISS50
$r_{g(d)}^{(1)}$					
Linear-4-cat.	0.64	0.65(0.005)	0.53(0.012)	0.55(0.030)	0.43(0.022)
Linear-Binary	0.40	0.43(0.010)	0.28(0.033)	0.30(0.026)	0.34(0.038)
4-cat.-Binary	0.13	0.48(0.009)	0.61(0.022)	0.73(0.033)	0.59(0.082)
$r_{g(m)}^{(2)}$					
Linear-4-cat.	0.07	-0.24(0.014)	-0.23(0.009)	-0.25(0.033)	-0.13(0.011)
Linear-Binary	0.11	-0.02(0.008)	-0.07(0.016)	-0.07(0.013)	0.06(0.043)
4-cat.-Binary	-0.23	0.13(0.015)	0.02(0.018)	-0.04(0.027)	0.20(0.038)
$r_{g(t)}^{(3)}$					
Linear	0.48	0.19(0.005)	0.19(0.008)	0.22(0.006)	0.17(0.005)
4-cat.	-0.10	-0.12(0.011)	-0.06(0.020)	0.05(0.038)	-0.12(0.029)
Binary	0.10	0.24(0.020)	0.55(0.135)	0.43(0.038)	0.15(0.065)

(1) Correlations for direct genetic effects between traits. (2) Correlations for maternal genetic effects between traits. (3) Correlations between direct and maternal genetic effects.

Table 4. Correlation coefficients between estimates for breeding values of three linear traits (original scales) without missing in the linear animal model and estimates for same effects by the threshold animal model (TM) using Gibbs sampling or the linear animal model (LM) using best linear unbiased prediction at three different simulated data sets

	LM			TM		
	NOMISS	MISS10	MISS50	NOMISS	MISS10	MISS50
Direct BV						
Linear	0.99(0.99)	0.97(0.97)	0.99(0.99)	0.99(0.99)	0.98(0.98)	0.99(0.99)
4-cat.	0.90(0.90)	0.87(0.87)	0.84(0.83)	0.94(0.93)	0.91(0.91)	0.86(0.85)
Binary	0.75(0.74)	0.75(0.74)	0.65(0.65)	0.87(0.87)	0.86(0.85)	0.80(0.79)
Maternal BV						
Linear	0.99(0.98)	0.97(0.97)	0.98(0.98)	0.99(0.99)	0.98(0.97)	0.98(0.98)
4-cat.	0.81(0.79)	0.79(0.77)	0.69(0.66)	0.87(0.85)	0.84(0.82)	0.75(0.72)
Binary	0.68(0.68)	0.68(0.69)	0.50(0.51)	0.80(0.82)	0.79(0.80)	0.72(0.73)

( ) : Rank correlation.

In evidence, the estimates on the 4-categorical and the binary trait were shown in data sets of MISS10 as 0.91:0.87 and 0.86:0.75, respectively. There were also shown the same trends regardless of the moment or rank correlations on data set of MISS50. The maternal genetic effects also were shown the same trends, which were higher correlation in TM rather than LM. We can conclude that a threshold model should be better than a linear model for estimating the breeding values as well as the genetic parameters for any categorical traits, especially in a sort of multi-trait animal model, at using the structure of data with the missing.

V. 요 약

한우의 근내지방도 또는 임신 여부 등과 같이 이산형 분포의 성질을 갖는 다수의 형질들에 대한 유전모수 및 종축의 유전능력을 평가하기 위한 방법으로써 Threshold 모형하에서 Bayesian 추론방법의 일종인 Gibbs sampling방법을 모의실험을 통하여 알아보았으며 기록이 누락된 다수의 형질을 포함하는 다형질 Threshold 개체모형에서의 종축평가 방법론을 제시하였다. 이산형 형질의 관측치에 대응하는 임의의 잠재변수는 기록을 갖고 있는 형질들에 대한 사전정보를 고려한 사후조건확률분포에서 Gibbs sampling을 할 때 모수에 근접하는 확률분포를 얻을 수 있었으며 이러한 이산형 기록들에 대한 육종가 추정치는 선형모형에서 보다 Threshold 모형에서의 추정치가 실제 모수에 더욱 근접하는 것을 알 수 있었다. 따라서 기록이 누락된 개체들에 대한 이산형 분포를 갖는 형질들에 대하여 선형분포를 갖는 형질들과 함께 동시 유전분석할 때 Threshod 모형이 일반 선형모형 보다 적합함을 알 수 있었다.

VI. ACKNOWLEDGEMENTS

Dr. I. Misztal's instruction for Bayesian inference was very helpful and some documentation for generating the residual effects by Dr. R. Rekaya and Dr. R. Tempelman were very appreciated. I also appreciate some comments about Bayesian inference to Dr. R. L. Quaas. I thank to Dr. S. C. Li who had worked for some theorem of Bayesian inference with me when he visited at University of Georgia.

VII. REFERENCES

1. Blasco, A. 2001. The Bayesian controversy in animal breeding. J. Anim. Sci. 79:2023-2046.
2. Box, G. E. P. and Tiao, G. C. 1992. *Bayesian inference in statistical analysis*. New York. Wiley.
3. Berger, P. J., Lin, E. C., Van Arendonk, J. and Janss, L. 1995. Properties of genetic parameter estimates from selection experiments by Gibbs sampling. J. Dairy Sci. 78 (Suppl 1):246.
4. Datta, G. S. and Maiti, T. 1998. Multivariate Bayesian small area estimation: An application to survey and satellite data. The Indian J. of Statistics. 60:344-362.
5. Djemali, M., Berger, P. J. and Freeman, A. E. 1987. Ordered categorical sire evaluation for dystocia in Holsteins. J. Dairy Sci. 70:2374-2384.
6. Foulley, J. L., Gianola, D. and Hoeschele, I. 1987. Empirical Bayes estimation of parameters for n polygenic binary traits. Genet. Sel. Evol. 15:407-424.
7. Foulley J. L., Gianola, D. and Thompson, R. 1983. Prediction of genetic merit from data on binary and quantitative variates with an application to calving difficulty, birth weight and pelvic opening. Genet. Sel. Evol. 15:407-424.
8. Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. 1995. *Bayesian data analysis*. Chapman & Hall.
9. Geman, S. and Geman, D. 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intel. 6:721-741.
10. Gianola, D. 1982. Theory and analysis of threshold characters. J. Anim. Sci. 54:1079-1096.
11. Gianola, D. and Fouley, J. L. 1983. Sire evaluation for ordered categorical data with a

threshold model. Genet. Sel. Evol. 15:201-244.
12. Hobert, J. and Casella, G. 1996. Effect of improper prior on Gibbs sampling in hierarchical linear mixed model. J. Am. Stat. Assoc. 91:1461-1473.
13. Hoeschele, I. B. and Tier, B. 1995. Estimation of variance components of threshold characters by marginal posterior mode and means via Gibbs sampling. Genet. Sel. Evol. 27:519-540.
14. Jensen, J., Wang, C. S., Sorensen, D. and Gianola, D. 1994. Bayesian inference on variance and covariance components for traits influenced by maternal and direct genetic effects using the Gibbs sampler. Acta Agric Scand, Sect A, Anim. Sci. 44:193-201.
15. Korsgaard, I. G., Andersen, A. H. and Sorensen, D. 1999. A usefull reparameterisation to obtain samples from conditional inverse Wishart distributions. Genet. Sel. Evol. 31:177-181.
16. Lee, D. H., Rekaya, R. and Misztal, I. 2002. Analysis of Binary Data: Effect of Different Parameterizations on the Bias of Genetic Parameters. Asian J. Anim. Sci. (Submitted).
17. Lee, D. H., Misztal, I., Bertrand, J. K. and Rekaya, R. 2001. Bayesian analysis of multiple-linear and categorical traits with varying number of categories. J. Anim. Sci. 79(suppl. 1):342.
18. Li, S. C. and Lee, D. H. 2002. Bayesian analysis of threshold animal models with Gibbs sampling. Korean J. Stat. Sci. (accepted).
19. Raftery, A. E. and Lewis, S. M. 1992. How many iterations in the Gibbs sampler? In: Bayesian Statistics IV. Oxford University Press. UK. 763-773.
20. Rekaya, R. 2001. Bayesian inference in mixed linear model using Dirichlet process prior. J. Anim. Sci. 79(suppl. 1):110.
21. Sorensen, D. A., Andersen, S., Gianola, D. and Korsgaard, I. 1995. Bayesian inference in threshold models using Gibbs sampling. Genet. Sel. Evol. 27:229-249.
22. Sorensen, D. 1996. *Gibbs sampling in quantitative genetics*. Internal reports no. 82. from the Danish Institute of Animal Science.
23. Sorensen, D., Wang, C. S., Jensen, J. and Gianola, D. 1994. Bayesian analysis of genetic trend due to selection to beef cattle breeding. Genet. Sel. Evol. 25:3-30.
24. VanTassell, C. P., VanVleck, L. D. and Gregory, K. E. 1998. Bayesian analysis of twinning and ovulation rates using a multiple-trait threshold model and Gibbs sampling. J. Anim. Aci. 76: 2048-2061.
25. Wang, C. S., Rutledge, J. J. and Gianola, D. 1994. Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. Genet. Sel. Evol. 26:91-115.
26. Wang, C. S., Quaas, R. L. and Pollak, E. J. 1997. Bayesian analysis of calving ease scores and birth weights. Genet. Sel. Evol. 20:117-143.
27. Wright, S. 1934. An analysis of variability in number of digits in an inbred strain of guinea pigs. Genetics 19:506-536.

Ⅷ. APPENDIX

According to equation (5), the underlying liabilities with considering missing traits can be generated by trait from a univariate (truncated) normal distribution with considering correlated traits because of computational convenience as:

Let $U_{i(o)}$ be observation (or liability scale) for the i^{th} linear (or categorical) trait and if missing, $U_{i(m)}$ can be denoted. Likewise, let $U_{-i(o)}$ be observation (or liability scale) for linear (or categorical) traits except the i^{th} trait.

Step 1 : calculate the residual effects for the i^{th} observed categorical or linear trait $e_{i(o)} = U^{n-1}_{i(o)} - W\theta_i$ where $U^{n-1}_{i(o)}$ is liability scale for categorical trait generated previous round or observed linear trait and θ_i is location parameter. This step can be performed for all traits.

Step 2 : generate i^{th} liability scale for categorical trait regardless of missing or observed and linear trait with missing

1. if the i^{th} trait were a categorical trait with no missing

$E(U_{i(o)} \mid U_{-i(o)}, \theta, R, t, y_{i(o)} = k) = W\theta_i + R_{i(o),-i(o)} R^{-1}_{-i(o),-i(o)} e_{-i(o)}$

$\mathrm{var}(U_{i(o)} \mid U_{-i(o)}, \theta, R, t, y_{i(o)} = k) = R_{i(o),-i(o)} R^{-1}_{-i(o),-i(o)} R_{-i(o),i(o)}$

$U_{i(o)} \mid U_{-i(o)}, \theta, R, t, y_{i(o)} = k) \sim TN_{t_{k-1}, t_k}(E(U_{i(o)} \mid \cdots), \mathrm{var}(U_{i(o)} \mid \cdots))$

$U^*_{i(o)} = U_{i(o)} / \sqrt{r_{i(o),i(o)}}$

2. else if the i^{th} trait were a categorical trait with missing

$E(U_{i(m)} \mid U_{-i(o)}, \theta, R) = W\theta_i + R_{i(m),-i(o)} R^{-1}_{-i(o),-i(o)} e_{-i(o)}$

$\mathrm{var}(U_{i(m)} \mid U_{-i(o)}, \theta, R) = R_{i(m),-i(o)} R^{-1}_{-i(o),-i(o)} R_{-i(o),i(m)}$

$U_{i(m)} \mid U_{-i(o)}, \theta, R) \sim N(E(U_{i(m)} \mid \cdots), \mathrm{var}(U_{i(m)} \mid \cdots))$

$U^*_{i(m)} = U_{i(m)} / \sqrt{r_{i(m),i(m)}}$

3. else if the i^{th} trait were a linear trait with missing

$E(U_{i(m)} \mid U_{-i(o)}, \theta, R) = W\theta_i + R_{i(m),-i(o)} R^{-1}_{-i(o),-i(o)} e_{-i(o)}$

$\mathrm{var}(U_{i(m)} \mid U_{-i(o)}, \theta, R) = R_{i(m),-i(o)} R^{-1}_{-i(o),-i(o)} R_{-i(o),i(m)}$

$U_{i(m)} \mid U_{-i(o)}, \theta, R) \sim N(E(U_{i(m)} \mid \cdots), \mathrm{var}(U_{i(m)} \mid \cdots))$

4. else continue

Liability scale for categorical trait or missing linear trait generated a previous step should be denoted to observed value to consider for generating the other variables.

Step 2 should be performed from first trait to last trait sequentially.

(접수일자 : 2001. 9. 28 / 채택일자 : 2002. 1. 24)