X Simpósio Brasileiro de Melhoramento Animal
Uberaba, MG – 18 a 23 de agosto de 2013

S B M A 2013

Data augmentation and complex traits

Luis Varona[1]

[1]Unidad de Genética Cuantitativa y Mejora Genética. Facultad de Veterinaria. Universidad de Zaragoza. 50013. ZARAGOZA (SPAIN)

ABSTRACT

The aim of livestock breeding plans is to improve an objective of selection by acting on several criteria of selection. The criteria of selection are composed by a subset of traits selected by its easiness of measure and its correlation with the objective of selection. In general, traits included on the criteria of selection are direct measures of the performance of candidates to selection. From an alternative point of view, the Bayesian paradigm allows to model the performance of livestock under a hierarchical Bayesian scheme. This approach can use the statistical technique of Data Augmentation, which allows to generate "pseudo-phenotypes" to mimic some complex traits. Later on, these "pseudo-phenotypes" may be used as new criteria of selection. The description of the procedure is illustrated with several examples of animal growth, ranking traits, subjective scores, food efficiency and disease resistance in several livestock populations.

BAYESIAN ANALYSIS AND DATA AUGMENTATION

The main objective of livestock breeding plans is to improve the performance of the livestock populations by acting on several criteria of selection. The criteria of selection are composed by a subset of traits selected by its easiness of measure and its correlation with the objective of selection. In general, traits included on the criteria of selection are direct measures of the performance of candidates to selection. Once traits included in the criteria of selection are identified, the Best Linear Unbiased Predictior (BLUP) usually obtains predictions of the breeding values after assuming the following linear model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$$

where **y** is the vector of phenotypic records, **X** and **Z** are the incidence matrices that links with systematic (**b**) and additive genetic (**u**) effects and **e** is the vector of residuals.

Linear models for genetic evaluation can be understood also under the Bayesian paradigm. Then, the Bayesian likelihood of data is:

$$p\left(\mathbf{y}\middle|\mathbf{b},\mathbf{u},\sigma_e^2\right) = \prod_{i=1}^{n} \frac{1}{\left(2\pi\sigma_e^2\right)^{0.5}} e^{-\frac{\left(y_i-\mathbf{x}_i\mathbf{b}-\mathbf{z}_i\mathbf{u}\right)^2}{2\sigma_e^2}}$$

where σ_e^2 is the residual variance and $\mathbf{x}_i$ and $\mathbf{z}_i$ are the vectors of the incidence matrices for the phenotypic record y_i. Further, the prior distribution for the additive genetic effects is:

$$p\left(\mathbf{u}\middle|\mathbf{G}\right) = \frac{1}{\left(2\pi\right)^{\frac{k}{2}}\left|\mathbf{G}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}\mathbf{u}'\mathbf{G}^{-1}\mathbf{u}}$$

where $\mathbf{G}=\mathbf{A}\,\sigma_a^2$. In addition, prior distributions for the variance components and systematic effects are usually assumed to be uniform within some appropriate bounds.

Thus, the final inference is achieved by the generation of random samples from the posterior distribution of parameters given the data:

$$p\left(\mathbf{b},\mathbf{u},\sigma_a^2,\sigma_e^2\middle|\mathbf{y}\right)$$

The usual approach to obtain samples for the joint posterior distribution, and, at the same time, for the marginal posterior distributions of the parameters of interest are McMC algorithms, such as the Gibbs Sampler.

Under the Bayesian approach, the probability structure can be defined by sequential probability distributions:

$$p(AB|C) = p(A|B)\,p(B|C)$$

Thus, it is possible to define auxiliary variables that can be augmented under the McMC structure.

$$p(\mathbf{y},\omega|\mathbf{b},\mathbf{u},\mathbf{G},\mathbf{R}) = p(\mathbf{y}|\omega)\,p(\omega|\mathbf{b},\mathbf{u},\mathbf{G},\mathbf{R})$$

where ω is the vector of augmented variables. The usual objective of augmented variables is to clarify the inference by the definition of known statistical distributions. However, in some cases, these augmented variables may have a biological sense, and, as a consequence, they can be used as criteria of selection. The procedure is illustrated with serveral examples.

EXAMPLES

Example 1. Categorical data

One of the simplest applications of the data augmentation approach is the Bayesian approach of the threshold models (Sorensen et al., 1995) for categorical data. Under the probit approach, it is assumed the existence of an auxiliary variable or augmented data (ω) called liability. Thus:

$$p(\mathbf{y}|\omega) = \prod_{i=1}^{n}\left(1(\omega_i > 0)1(y_i = 1) + 1(\omega_i \le 0)1(y_i = 0)\right)$$

Further, the prior distribution for the liability included the usual systematic and breeding values, with the assumption of a residual variance set to 1 (or any fixed value).

$$p(\omega|\mathbf{b},\mathbf{u}) = \prod_{i=1}^{n}\frac{1}{(2\pi)^{0.5}}e^{-\frac{(y_i-\mathbf{x}_i\mathbf{b}-\mathbf{z}_i\mathbf{u})^2}{2}} \quad \text{and} \quad p(\mathbf{u}|\sigma_a^2) = \frac{1}{(2\pi)^{\frac{k}{2}}|\mathbf{A}|^{\frac{1}{2}}\sigma_a}e^{-\frac{\mathbf{u}'\mathbf{A}^{-1}\mathbf{u}}{2\sigma_a^2}}$$

Finally, the inference of the additive variance component, the systematic effects or the breeding values is performed for the liability or augmented variable.

Example 2. Subjective data

Interesting generalizations of the categorical data analysis under the threshold model are the cases when data comes from a subjective evaluation by several expert panelists (Varona and Hernández, 2006; Varona et al., 2009). One suitable approximation comes from the generalization of the threshold model to a multi-threshold model that defines a different set of thresholds for each panelist.

$$p(\mathbf{y}|\omega,\mathbf{t}) = \prod_{i=1}^{S}\prod_{j=1}^{n_i}\left(1(\omega_{ij} < t_{i1})1(y_{ij} = 1) + 1(t_{i1} < \omega_{ij} < t_{i2})1(y_{ij} = 2) + ... + 1(\omega_{ij} > t_{ik})1(y_{ij} = k+1)\right)$$

where t is the vector of specific thresholds for each panelist. As before, the prior distribution of liability corresponds to the standard mixed model equations, and the variances components and the breeding values are referenced to this liability.

Example 3. Ranking Data

Another interesting application of the data augmentation is the Thurstonian model (Gianola and Simianer, 2006). The Thurstonian model allows to model ordered categorical response such as the one obtained in horse competitions. As an example, lets think in one competition with four competitors (A, B, C and D). If the results of the competition was that C is the first, A the second, D the third and B the last one, the Thurstonian model will assume a liability that ensure ordered liabilities as:

$\omega_C > \omega_A, \omega_C > \omega_D, \omega_C > \omega_B, \omega_A > \omega_D, \omega_A > \omega_B, \omega_D > \omega_C$

Later on, and as in the previous analysis, the prior distributions assume that the liability is regulated by systematic effects and breeding values. Moreover, this liability can be treated as a potential criteria of selection.

Example 4. Production Functions.

The biological cycle of livestock individual may be sometimes modeled using production functions, such as growth or lactation curves. Another interesting application of data augmentation is the use of a hierarchical Bayesian approach to model these production functions (Varona et al., 1997). This approach assumes a conditional probability for the available data along time given several parameters of the production function, which also can have some biological sense. As an example, Varona et al. (1998) applied the hierarchical Bayesian approach to the Wood lactation curve. Thus the Bayesian likelihood was defined as:

$$p\left(\mathbf{y}|\mathbf{a},\mathbf{k},\mathbf{c},\sigma_t^2\right)=\prod_{i=1}^{m}\prod_{j=1}^{n_i}\frac{1}{\left(2\pi\sigma_t^2\right)^{0.5}}e^{-\frac{\left(y_{ij}-a_i t_{ij}^{k_i}\exp\left(-c_i t_{ij}\right)\right)}{2\sigma_t^2}}$$

where **a**, **k** and **c** are the vector of the three parameters of the Wood function, that can be interpreted as the base milk production, ascent and descent to peek. Further, the prior distributions assume a multivariate mixed model for the augmented variables.

Example 5. Food Intake.

A similar approach using hierarchical Bayesian schemes can be used to model food intake data (Piles et al., 2006). There, in the first stage, food intake records (**y**) are modeled with a linear dependency on metabolic weight ($\mathbf{w}^{0.75}$) and the increase of weight (**Δw**).

$$p\left(\mathbf{y}|\mathbf{a},\mathbf{b},\sigma_t^2\right)=\prod_{i=1}^{m}\prod_{j=1}^{n_i}\frac{1}{\left(2\pi\sigma_t^2\right)^{0.5}}e^{-\frac{\left(y_{ij}-a_i w_{ij}^{0.75}-b_i\Delta w_{ij}\right)}{2\sigma_t^2}}$$

Later, and in a second a stage, **a** and **b** are modeled with the standard multivariate mixed models.

Example 6. Residual Asymmetry

The standard assumption of the Bayesian analysis of phenotypic data in animal breeding is that the residuals follow a Gaussian distribution. However, in some cases, it is possible to assume an asymmetric Gaussian distribution (Sahu et al., 2003), that describe the variation with two parameters, the variance (σ^2) and the degree of asymmetry (λ). This degree of asymmetry can be understood as a measure of sensitivity to negative (or positive) environmental influences on phenotypes (Varona et al., 2008). Thus, the likelihood of data can be expressed as:

$$p\left(\mathbf{y}|\mathbf{b},\mathbf{u},\sigma_e^2,\lambda\right)=\prod_{i=1}^{n}\frac{2}{\sqrt{\sigma_e^2+\lambda_i}}\phi\left(\frac{y_i-\mathbf{x}_i\mathbf{b}-\mathbf{z}_i\mathbf{u}}{\sqrt{\sigma_e^2+\lambda_i}}\right)\Phi\left(\frac{\lambda_i}{\sigma_e^2}\frac{y_i-\mathbf{x}_i\mathbf{b}-\mathbf{z}_i\mathbf{u}}{\sqrt{\sigma_e^2+\lambda_i}}\right)$$

where λi is defined as the individual degree of asymmetry and it is here the augmented phenotype. Thus, a priori, it can be explained by the standard mixed model analysis.

$$\lambda=\mathbf{Xb}+\mathbf{Zu}+\mathbf{e}$$

Example 7. Binomial Models

Another interesting approach of the hierarchical Bayesian schemes and the augmentation of phenotypes is its potential application to discrete phenotypes that follow alternative distributions. In recent paper, Varona et al. (2010) propose the use of Poisson, binomial and negative binomial models for pig mortality. As an example, the Bayesian likelihood for the binomial model is:

$$p\left(\mathbf{y}|\theta,\mathbf{t}\right)=\prod_{i=1}^{N}\binom{t_i}{y_i}\theta_i^{y_i}\left(1-\theta_i\right)^{t_i-y_i}$$

where θ_i is the individual probability of born alive. Further, a priori, a linear model for the logit transformation is assumed:

$$\log it(\theta) = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$$

Conclusions

The power of the Bayesian analysis under a hierarchical Bayesian approach allows defining new variables that can be augmented and modeled. In some cases, these variables may have a biological sense and the development of linear models provide new breeding values that can be used to create new criteria of selection.

REFERENCES

GIANOLA, D.; SIMIANER, H.A. A Thrustonian model for quantitative genetic analysis of ranks: a Bayesian approach. **Genetics**, v.174, p.1613-1624, 2006.

GOMEZ, M.D.; VARONA, L.; MOLINA, A.; et al.M. Genetic evaluation of racing performance in trotter horses by competitive models. **Livestock Science**, v.140, p.155-160, .

PILES, M.; GARCÍA-TOMÁS, M.; RAFEL, O.; et al. Individual efficiency for the use of feed resources in rabbits. **Journal of Animal Science**, v. 85, p.2846-2853, 2007.

SAHU, S.K.; DEY, D.K.; BRANCO, M.D. A new class of multivariate skew distributions with applications to Bayesian regression models. **The Canadian Journal of Statistics**, v.31, p.129-150, 2003.

SORENSEN, D.A.; ANDERSEN, S.; GIANOLA, D.; et al. Bayesian inference in threshold models using Gibbs sampling. **Genetics Selection Evolution**, v.27, p.229-249,1995.

VARONA, L.; HERNÁNDEZ, P. A Mulithreshold model for Sensory Analysis. **Journal of Food Science**, v.71, p.333-336, 2006.

VARONA, L.; IBAÑEZ-ESCRICHE, N.; QUINTANILLA, R.; et al. Bayesian analysis of quantitative traits using skewed distributions. **Genetics Research**, v.90, p.179-190, 2008.

VARONA, L.; MORENO, C.; ALTARRIBA, J. A model with heterogeneous for subjective traits: fat cover and conformation score in the Pirenaica beef cattle. **Journal of Animal Science**, v.87, p.1210-1217, 2009.

VARONA, L.; MORENO, C.; GARCÍA-CORTÉS, L.A.; et al. Multiple trait genetic analysis of underlying biological variables of production functions. **Livestock Science**, v.47, p.201-209, 1997.

VARONA, L.; MORENO, C.; GARCÍA-CORTÉS, L.A.; et al. Bayesian Analysis of Wood's Lactation Curve for Spanish Dairy cows. **Journal of Dairy Science**, v.81, p.1469-1478, 1998.

VARONA, L.; SORENSEN, D.A. A Genetic Analysis of Mortality in Pigs. **Genetics**, v.184, p.277-284, 2010.