Joint Analysis of Binomial and Continuous Traits with a Recursive Model: A Case Study Using Mortality and Litter Size of Pigs

Luis Varona* and Daniel Sorensen†,[1]

*Departamento de Anatomía, Embriología y Genética Animal, Facultad de Veterinaria, Universidad de Zaragoza, E-50013, Spain,
and †Department of Molecular Biology and Genetics, Aarhus University, PB 50, DK-8830 Tjele, Denmark

ABSTRACT This work presents a model for the joint analysis of a binomial and a Gaussian trait using a recursive parametrization that leads to a computationally efficient implementation. The model is illustrated in an analysis of mortality and litter size in two breeds of Danish pigs, Landrace and Yorkshire. Available evidence suggests that mortality of piglets increased partly as a result of successful selection for total number of piglets born. In recent years there has been a need to decrease the incidence of mortality in pig-breeding programs. We report estimates of genetic variation at the level of the logit of the probability of mortality and quantify how it is affected by the size of the litter. Several models for mortality are considered and the best fits are obtained by postulating linear and cubic relationships between the logit of the probability of mortality and litter size, for Landrace and Yorkshire, respectively. An interpretation of how the presence of genetic variation affects the probability of mortality in the population is provided and we discuss and quantify the prospects of selecting for reduced mortality, without affecting litter size.

MIXED linear models (Henderson 1984) are broadly used in livestock and plant breeding and play an important role in evolutionary and theoretical quantitative genetics (Lande 1979; Cheverud 1984; Walsh 2003). The classical approach for a multiple-trait analysis is to use models posing that the nature of the correlation between response variables (phenotypes) is due to linear associations between unobservables, such as additive genetic values or nongenetic sources, like permanent or temporary environmental effects.

Structural equation models represent an extension of the standard linear model to account for links (feedback and/or recursiveness) involving either the phenotypes directly or latent variables; they are well established in econometrics and sociology (Goldberger 1972; Jöreskog 1973; Duncan 1975). These models were discussed in the early genetics literature by Wright (1921) but this work has not received much attention in quantitative genetics. Xiong *et al.* (2004) proposed the use of structural equation models for modeling and identifying genetic networks. In a quantitative genetics context, Gianola and Sorensen (2004) studied the consequences of the existence of simultaneous and recursive relationships between phenotypes on genetic parameters and presented statistical methods for inference. An application to study the relationship between somatic cell score and milk yield in goats is in de los Campos *et al.* (2006). Varona *et al.* (2007) present a recursive model for the joint analysis of litter size and average litter weight in Danish pigs. These studies were concerned with normally distributed traits. Here the methodology is developed further for the joint analysis of a binomial and a continuous trait and it is shown that a computationally simple implementation can be arrived at by appropriate choice of the recursive specification. The method is illustrated using mortality and litter size in two breeds of Danish pigs.

Litter size is basically determined by ovulation rate and embryo mortality (Blasco *et al.* 1995); these processes take place mainly at the early stages of gestation. Piglet weight at birth is determined mostly by growth in late gestation and is importantly related to piglet survival. It is then reasonable to postulate a one-way causal path establishing an effect of litter size on piglet mortality. This specification defines a recursive two-trait system. On the other hand, simultaneity occurs when trait 1 affects trait 2 and vice versa.

doi: 10.1534/genetics.113.159475

Manuscript received December 16, 2013; accepted for publication January 7, 2014; published Early Online January 10, 2014.

Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.159475/-/DC1.

[1]Corresponding author: Department of Molecular Biology and Genetics, Aarhus University, E-mail: daniel.alberto.sorensen@gmail.com

Litter size has been under selection in the Danish pig-breeding program since the early nineties and resulted in considerable increase in total number born and also in the proportion of stillborn piglets (Sorensen *et al.* 2000; Su *et al.* 2007). Sorensen *et al.* (2000) report an increase in the observed proportion of piglets born dead at higher-litter-size values. This has raised a number ethical and economic concerns and has led to measures designed to reduce mortality. A recently implemented approach in the Danish pig-breeding program is based on changing the emphasis of selection from total number born to total number of piglets alive 5 days after farrowing (Su *et al.* 2007). Despite the fact that this selection strategy is not addressing the problem of mortality directly, it seems to have had a beneficial effect on both litter size and mortality (Nielsen *et al.* 2013).

A number of studies have reported genetic variation for mortality with heritabilities ranging from 0.03 to 0.17. These studies have assumed normality of the sampling model for mortality (*e.g.*, Van Arendonk *et al.* 1996), based inferences on a variety of threshold models (*e.g.*, Roehe and Kalm 2000; Arango *et al.* 2006), or implemented mixed models for count data (Varona and Sorensen 2010). Mortality data, regarded as a trait of the mother, show typically a large proportion of "zeros" (many litters do not have stillborn piglets). The study of Varona and Sorensen (2010) included a variety of models that accounted for this feature of the data and concluded that the best fit was achieved with a hierarchical binomial logit mixed model. In this work we extend this model in two directions. First, the probability of mortality is assumed to be a function of the total number of piglets born in the litter. This is achieved by assigning a recurrent relationship between the logit of the probability of mortality and litter size. Linear and higher-order functions of litter size are investigated, and the quality of fit of the models is studied. The second extension allows for a joint analysis of mortality and litter size. The recursive parameterization implemented has the attractive feature that the joint posterior distribution of the two traits factorizes into two independent posterior distributions, one for each trait, whereby the computational burden of implementation is reduced.

The article is organized as follows. *Material and Methods* introduces the models, including the prior and posterior distributions, the method used to compare the models, and a brief description of the data. This is followed by *Results*, where the focus is on mortality but results for litter size are briefly reported. A *Discussion* comprises the final section of the paper and the *Appendix* sketches technical details regarding the model and the Markov chain Monte Carlo algorithm.

Material and Methods

Model and prior distributions

Total number of dead piglets at birth (called mortality hereinafter, treated as a trait of the mother) and total number of piglets born (called litter size hereinafter, treated as a trait of the mother) are analyzed jointly using a model that exploits the factorization of their joint distribution. The conditional binomial model for mortality in litter i, Y_i, given litter size in litter i, t_i is

$$f(Y_i = y_i|t_i, \phi_i) = \binom{t_i}{y_i}\phi_i^{y_i}(1-\phi_i)^{t_i-y_i},\ Y_i = 0, 1, \ldots, t_i, \tag{1}$$

where ϕ_i is the probability that a piglet dies (referred to as the probability of mortality hereinafter) in litter i, which is assumed to vary over the observations as an inverse logistic deterministic function of unknown parameters. Thus, the linear structure of the logit of ϕ_i is assumed to be equal to

$$\text{logit}\ \phi_i = x_i'\alpha_y + z_i'\tilde{u}_y + w_i'\tilde{p}_y + g_j(t_i), \tag{2}$$

where x_i', z_i', and w_i' are vectors of observed incidence matrices X, Z, and W, α_y is a vector of systematic effect parameters affecting mortality (herd-year and parity), $\tilde{u}_y$ is a vector of residual additive genetic values affecting mortality defined in the *Appendix*, $\tilde{p}_y$ is a vector of residual permanent environmental effects affecting mortality (see also the *Appendix* for an explanation), and $g_1(t_1) = \lambda_1 t_i$ ($j = 1$ for Model 1), $g_2(t_i) = \lambda_1 t_i + \lambda_2 t_i^2$ ($j = 2$ for Model 2), $g_3(t_i) = \lambda_1 t_i + \lambda_2 t_i^2 + \lambda_3 t_i^3$ ($j = 3$ for Model 3), where the λ's are recurrent parameters. In the case of Model 1 it is easy to see that two possible partitions are

$$\begin{aligned}\text{logit}\ \phi_i &= x_i'\alpha_y + z_i'\tilde{u}_y + w_i'\tilde{p}_y + \lambda_1 t_i\\ &= x_i'\alpha_y + z_i'u_y + w_i'p_y + \lambda_1 x_i'\alpha_t + \lambda_1 e_{t_i},\end{aligned}$$

where $u_{y_i} = E(u_{y_i}|u_{t_i}) + \tilde{u}_{y_i}$ and $p_{y_i} = E(p_{y_i}|p_{t_i}) + \tilde{p}_{y_i}$ (see *Appendix*) represent draws from the marginal distributions of additive genetic values and permanent environmental effects affecting mortality.

Given vectors of systematic effects α_t (herd-year and parity), of additive genetic values u_t and of permanent environmental effects p_t, litter size records are conditionally independent and assumed to follow the Gaussian process

$$t_i|\alpha_t, u_t, p_t \sim N\left(x_i'\alpha_t + z_i'u_t + w_i'p_t, \sigma^2_{e_t}\right). \tag{3}$$

In the *Appendix* it is shown that the structure of $(\tilde{p}_{y_i}, p_{t_i})$ is

$$\left(\tilde{p}_{y_i}, p_{t_i}\right) \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma^2_{\tilde{p}_y} & 0\\ 0 & \sigma^2_{p_t}\end{pmatrix}\right]$$

and also in the *Appendix* it is shown that under the recursive parameterization employed,

$$\left(\tilde{u}_{y_i}, u_{t_i}\right) \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma^2_{\tilde{u}_y} & 0\\ 0 & \sigma^2_{u_t}\end{pmatrix}\right].$$

The covariance matrix of the joint distribution of the vectors $\tilde{u}_y$ and u_t is $G \otimes A$, where $G = \text{diag}(\sigma^2_{\tilde{u}_y}, \sigma^2_{u_t})$. Therefore the joint distribution factors into the product of the marginal distributions; that is,

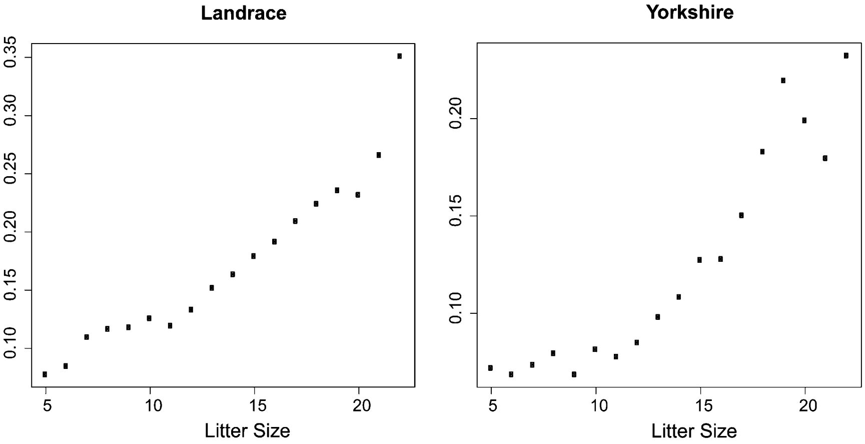

Figure 1 Raw phenotypic averages of the proportion of dead-born piglets as a function of litter size, for Landrace and for Yorkshire.

$$\widetilde{p}\left(u_y,u_t|A,\sigma^2_{\tilde{u}_y},\sigma^2_{u_t}\right)=p\left(\tilde{u}_y\big|A,\sigma^2_{\tilde{u}_y}\right)p\left(u_t|A,\sigma^2_{u_t}\right)$$

and similarly,

$$p\left(\tilde{p}_y,p_t|\sigma^2_{\tilde{p}_y},\sigma^2_{p_t}\right)=p\left(\tilde{p}_y|\sigma^2_{\tilde{p}_y}\right)p\left(p_t|\sigma^2_{p_t}\right).$$

In these expressions, A is the additive genetic relationship matrix, $\sigma^2_{\tilde{u}_y}$ is the residual additive genetic variance for mortality (conditional variance of the additive genetic value for mortality, given the additive genetic value of liter size), $\sigma^2_{u_t}$ is the additive genetic variance for litter size, $\sigma^2_{p_y}$ is the variance of permanent environmental effects for mortality, and $\sigma^2_{p_t}$ is the variance of permanent environmental effects for litter size. The phenotypic variance for litter size is $\sigma^2_t = \sigma^2_{u_t} + \sigma^2_{p_t} + \sigma^2_{e_t}$, where $\sigma^2_{e_t}$ is the variance of the conditional distribution of litter size.

The variance component parameters, the recurrent parameters, and the vectors α_y and α_t are assigned independent improper uniform distributions *a priori*.

Posterior distributions

Given the likelihood models (1) and (3) and the prior distributions of the parameters, the joint posterior distributions corresponding to Model 1, say, are

$$\begin{aligned} &p\left(\alpha_y,\tilde{u}_y,\tilde{p}_y,\alpha_t,u_t,p_t,\lambda_1,\sigma^2_{\tilde{u}_y},\sigma^2_{\tilde{p}_y},\sigma^2_{u_t},\sigma^2_{p_t},\sigma^2_{e_t}|y,t\right)\\ &\quad\propto f\left(y|t,\alpha_y,\tilde{u}_y,\tilde{p}_y,\lambda_1\right)p(t|\alpha_t,u_t,p_t)\\ &\quad\times p\left(\tilde{u}_y|\sigma^2_{\tilde{u}_y}\right)p\left(\tilde{p}_y|\sigma^2_{\tilde{p}_y}\right)p\left(u_t|\sigma^2_{u_t}\right)p\left(p_t|\sigma^2_{p_t}\right), \end{aligned} \tag{4}$$

where y and t are vectors with elements y_i and t_i, respectively and

$$\begin{aligned} f\left(y|t,\alpha_y,\tilde{u}_y,\tilde{p}_y,\lambda_1,\lambda_2,\lambda_3\right) &= \prod_{i=1}^{n} f(Y_i=y_i|t_i,\phi_i)\\ &= \prod_{i=1}^{n}\begin{pmatrix} t_i\\ y_i\end{pmatrix}\phi_i^{y_i}(1-\phi_i)^{t_i-y_i}, \end{aligned}$$

$$\begin{aligned} p(t|\alpha_t,u_t,p_t)=\left(2\pi\sigma^2_{e_t}\right)^{-n/2}&\exp\Big[-\tfrac{1}{2\sigma^2_{e_t}}(t-X\alpha_t+Zu_t+Wp_t)'\\ &\times(t-X\alpha_t+Zu_t+Wp_t)\Big]. \end{aligned}$$

Note that in (4), the joint posterior distribution factorizes into two independent posterior distributions, one for each trait.

Model comparison

The models are compared using the pseudo-log-marginal probability of the data. The pseudo-log-marginal probability of the data is a standard measure of model comparison (Gelfand 1996) and is defined and computed as follows. Consider data vector $\mathbf{y}' = (y_i,\mathbf{y}'_{-i})$, where y_i is the ith datum, and $\mathbf{y}_{-i}$ is the vector of data with the ith datum deleted. The conditional predictive distribution can be interpreted as the probability of each data point given the remainder of the data and has probability density

$$\begin{aligned} p\left(y_i\,|\,\mathbf{y}_{-i}\right) = \int p\left(y_i\,|\,\boldsymbol{\theta},\mathbf{y}_{-i}\right)f\left(\boldsymbol{\theta}\,|\,\mathbf{y}_{-i}\right)d\boldsymbol{\theta},\\ \boldsymbol{\theta}=\{\theta_i\}^n_{i=1}, \end{aligned} \tag{5}$$

where $\boldsymbol{\theta}$ is the vector of parameters. The actual value of $p(y_i|\mathbf{y}_{-i})$ is known as the *conditional predictive ordinate* (CPO) for the ith observation. The pseudo-log-marginal probability of the data are given by

$$\sum_i \ln p(y_i|\mathbf{y}_{-i}). \tag{6}$$

A Monte Carlo approximation of the CPO (5) for observation i is given by (Gelfand 1996)

$$\widehat{p}(y_i|\mathbf{y}_{-i},M_k)=N\left[\sum_{j=1}^{N}\frac{1}{p\left(y_i|\theta_i^{(j)},M_k\right)}\right]^{-1}, \tag{7}$$

where N is the number of MCMC draws, M_k is a label for model k, and $\theta_i^{(j)}$ is the jth draw from the posterior of θ_i under model k corresponding to the ith observation. The so-called Log CPOs reported below are based on

$$\sum_i \ln \hat{p}(y_i|\mathbf{y}_{-i},M_k).$$

MCMC algorithm

The fully conditional posterior distributions associated with mortality do not all have closed forms, except for the variance

Table 1 Posterior means and standard deviations (in brackets) of variance components, recursive parameters and of LogCPO in Landrace, for mortality

	Model 0[a]	Model 1	Model 2	Model 3
$\sigma^2_{\tilde{u}_y}$	0.168 (0.035)	0.162 (0.032)	0.162 (0.031)	0.157 (0.031)
$\sigma^2_{\tilde{p}_y}$	0.344 (0.032)	0.281 (0.028)	0.283 (0.028)	0.288 (0.029)
λ_1		0.094 (0.005)	0.087 (0.017)	0.075 (0.016)
$\lambda_2 \times 10^{-3}$			0.278 (0.570)	1.530 (1.010)
$\lambda_3 \times 10^{-4}$				−0.314 (0.248)
Log CPO	−10164	−9930	−9934	−9949

[a] For Model 0, $\sigma^2_{\tilde{u}_y} = \sigma^2_{u_y}$ and $\sigma^2_{\tilde{p}_y} = \sigma^2_{p_y}$.

components $\sigma^2_{\tilde{u}_y}$ and $\sigma^2_{\tilde{p}_y}$, which are scaled inverted chi-square distributions and can therefore be easily updated. For the remaining parameters of the model, $\alpha_y, \tilde{u}_y$, and $\tilde{p}_y$, a random walk single-site Metropolis–Hastings algorithm was chosen as updating strategy. This required a little preliminary experimentation to tune the input parameters. For $\tilde{u}_y$ and $\tilde{p}_y$, the random walk proposal consisted of a draw from a uniform distribution centered at the current value, and with lower and upper bounds given by plus and minus the updated draw from $\sigma_{\tilde{u}_y}$ and $\sigma_{\tilde{p}_y}$, respectively. For α_y the uniform was also centered at the current value and the bounds were given by ± 0.15. In the case of the recursion parameters, the bounds were as follows: for λ_1, ± 0.015, for λ_2, ± 0.0015, and for λ_2, ± 0.00015. The final inference was based on single chains of length 5 million (several were run with different starting values as checks and the convergence of the chains to their posterior distributions was studied by visual inspection of trace plots of chosen parameters). The effective chain sizes for the dispersion parameters for mortality varied from ∼750 to 3600 in Landrace and from 400 to 4600 in Yorkshire. For the best fitting models, the effective chain sizes associated with the regression parameter(s) were 3700 in Landrace (Model 1) and varied from 40 to 90 in Yorkshire (Model 3).

Data

Data were obtained from an existing database of performance records collected from nucleus farms of Danish Landrace and Danish Yorkshire during the period from May 2002 until December 2004. Pedigrees were traced back five generations or more. For Landrace, the data comprised records from 5178 litters and a pedigree file of 8800 individuals. The Yorkshire data consisted of records from 3938 litters and a pedigree file of 7143 individuals. Sows were kept under commercial conditions and all matings took place using artificial insemination. More details can be found in Su *et al.* (2007).

Results

The raw means for litter size for parities 1, 2, 3, and >4 are as follows: 13.4, 15.3, 16.1, and 16.3 for Landrace and 12.3, 14.1, 14.5, and 14.6 for Yorkshire. The average observed proportion of dead-born piglets in parities 1, 2, 3, and >4 are 0.17, 0.17, 0.20, and 0.23 in Landrace and 0.11, 0.09, 0.12, and 0.17 in Yorkshire. Figure 1 shows the raw mortality proportions for a given litter size, across the range of values of litter size of the data sets, in Landrace and Yorkshire. The figures provide a rough illustration for the phenotypic relationship between mortality and litter size, especially within the range defined by litter sizes between 7 and 20 in Landrace and between 6 and 19 in Yorkshire. Within this range each point is represented with a minimum of 100 observations, and outside this range, especially at litter sizes <3 and >23 in Landrace, and 4 and 21 in Yorkshire, with <20 observations. The figures indicate that the proportions increase nonlinearly with the size of the litters, but the relationships are a little different in the two breeds (this is more clearly visualized in Figure 4, which displays the probability of mortality as a function of litter size, based on the best-fitting models—Model 1 for Landrace and Model 3 for Yorkshire).

The Monte Carlo estimates of Log CPO (best model has the largest value) indicate that in both breeds, the poorest fit is obtained with Model 0, which assumes that the probability of mortality does not depend on litter size. The differences in the quality of fit are not very marked among the remaining models. For Landrace, the results in Table 1 indicate that a linear relationship (Model 1) gives the best overall fit. The regression parameters differ in Yorkshire (Table 2) and the Log CPO indicates that for this breed a cubic relationship (Model 3) between the logit of mortality and litter size gives the best overall fit.

Shown in Table 1 are Monte Carlo estimates of posterior means and posterior standard deviations of various parameters in Landrace. In the case of mortality, the figures in the table indicate that ∼37% of the total variance of the logit of mortality (total variance is equal to $\sigma^2_{\tilde{u}_y} + \sigma^2_{\tilde{p}_y} = 0.443$) is accounted for by the residual additive genetic variance (for the best-fitting model, Model 1 for Landrace). These results imply that at the level of the logit for mortality, the additive genetic correlation between mortality and litter size based on Model 1 is ∼0.20 in both breeds [calculated from (9), using estimates of posterior means of the additive genetic variance for litter size, $\hat{\sigma}^2_{u_t} \approx 0.8$, retrieves an estimate of the genetic correlation $\lambda_1(\hat{\sigma}_{u_t}/\hat{\sigma}_{u_y}) \approx 0.09(0.80/0.16)^{0.5} = 0.20$]. Similar calculations show that the estimate of the correlation between permanent environmental effects is ∼0.13 in both breeds. Nielsen *et al.* (2013) report estimates of the genetic correlation between mortality and litter size ranging between 0.22 and 0.28. However, their analysis treats mortality as a Gaussian trait and the figures are not directly comparable with the results reported here.

Table 2 Posterior means and standard deviations (in brackets) of variance components, recursive parameters and of Log CPO in Yorkshire, for mortality

	Model 0[a]	Model 1	Model 2	Model 3
$\sigma^2_{u_y}$	0.170 (0.044)	0.165 (0.046)	0.163 (0.045)	0.168 (0.053)
$\sigma^2_{p_y}$	0.585 (0.051)	0.494 (0.049)	0.484 (0.049)	0.487 (0.053)
λ_1		0.121 (0.007)	−0.0094 (0.0362)	−0.191 (0.046)
$\lambda_2 \times 10^{-2}$			0.473 (0.129)	1.943 (0.306)
$\lambda_3 \times 10^{-3}$				−0.371 (0.076)
Log CPO	−6363	−6185	−6181	−6176

[a] For Model 0, $\sigma^2_{\tilde{u}_y} = \sigma^2_{u_y}$ and $\sigma^2_{\tilde{p}_y} = \sigma^2_{p_y}$.

For Yorkshire (see Table 2), estimates of variances for mortality are broadly similar although the additive genetic variance comprises a somewhat smaller proportion of the total variance of the logit of mortality (26% for Model 3, the best-fitting model in Yorkshire).

The estimates of heritability for litter size in both breeds are very similar. The posterior mean is equal to 0.077 with a posterior standard deviation of 0.020. These estimates are similar to those reported by Su *et al.* (2007).

We have also studied how predictions of residual additive genetic values for mortality (based on posterior means) differ among the four models. In both breeds, the three product moment correlations between the predictions based on Model 0 and those based on each of the other three models are in the vicinity of 0.96. The three product moment correlations between predictions derived from Models 1, 2, and 3 are all >0.99.

Discussion

In previous work we performed genetic analyses of count data using a number of discrete models (Varona and Sorensen 2010) with an illustration using mortality in pigs. Mortality data show overdispersion, due to a high proportion of litter records with an absence of mortality and heterogeneity induced by covariation among observations. The models accounted for both sources of overdispersion and the study confirmed the presence of genetic variation for mortality. The model that showed the best global fit was a hierarchical binomial logit model and was therefore chosen in this work. In contrast with the models implemented by Varona and Sorensen (2010), in this work the logit of the probability of mortality is assumed to be functionally related to litter size. Both linear and nonlinear functions at the level of the logit were studied and the results indicate that in Landrace, the linear relationship leads to the best global fit. In Yorkshire, quadratic and cubic relationships produce better global fits and all the models translate into nonlinear relationships between the probability of mortality and litter size.

In mixed linear models the interpretation of variance components is straightforward because the random effects operate on the same scale as the values of the response variable. This is not the case in generalized mixed models. One way of studying the direct impact of the variances on the probability of mortality is as follows. Consider first the simplified version (excluding "random effects") of the model defined in (2), with $g_1(t_i) = \lambda_1 t_i$,

$$\ln\left(\frac{\phi_i}{1-\phi_i}\right) = \mu_i + \lambda t_i,$$

where μ_i is the mean of the ith record (that includes the sum of the effects herd-year and parity) and t_i is the effect of litter size. For example, in Landrace, replacing the values of the parameters for parity 1 and herd-year 1 by their posterior means (resulting in a value of $\mu_i \approx -2.5$), and using the posterior mean of λ (0.094), translates into a value of the probability

$$\hat{\phi}_i = \frac{\exp\left(\hat{\mu}_i + \hat{\lambda} t_i\right)}{1 + \exp\left(\hat{\mu}_i + \hat{\lambda} t_i\right)}$$

equal to 0.17 for $t_i = 10$ and 0.21 for $t_i = 17$. The extended "mixed model" version of the logit for the ith record is

$$\ln\left(\frac{\phi_i}{1-\phi_i}\right) = \mu_i + \lambda t_i + q_i,$$

where the random effect $q_i \sim N(0, \sigma^2_q)$ is the sum of the residual additive genetic effect and of the residual permanent environmental effect, with $\sigma^2_q = \sigma^2_{\tilde{u}_y} + \sigma^2_{\tilde{p}_y}$. Given μ_i and t_i the probability ϕ_i is a function of q_i,

$$\phi_i = f(q_i) = \frac{\exp(\mu_i + \lambda t_i + q_i)}{1 + \exp(\mu_i + \lambda t_i + q_i)}.$$

The inverse function is $f^{-1}(\phi_i) = \ln[\phi_i/(1-\phi_i)] - \mu_i + \lambda t_i$ and the Jacobian is equal to $(\phi_i(1-\phi_i))^{-1}$. Therefore the probability density of ϕ_i is

$$p(\phi_i) = \frac{1}{\sqrt{2\pi\sigma^2_q}} \exp\left[-\frac{(\ln[\phi_i/(1-\phi_i)] - \mu_i + \lambda t_i)^2}{2\sigma^2_q}\right] \frac{1}{\phi_i(1-\phi_i)}. \qquad (8)$$

The effect of the variance component on the probability of mortality can be studied using the density (8).

Figure 2 displays the distribution of ϕ_i for $t_i = 10$, where σ^2_q is replaced by its posterior mean $\hat{\sigma}^2_q = 0.443$ (computed as 0.162 + 0.281; see Table 1). The variance parameter $\hat{\sigma}^2_q$ defines the range and shape of the distribution in a manner

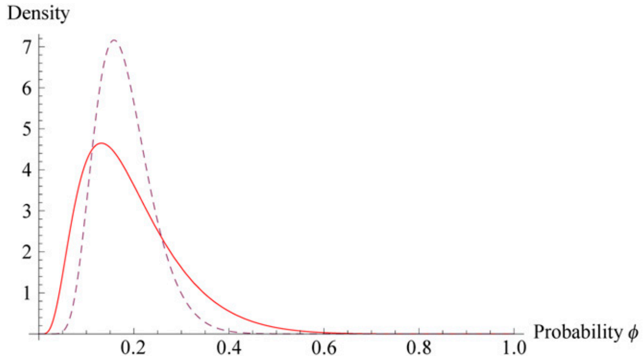

Figure 2 Probability distribution of the probability of mortality for $t_i = 10$. Solid line, $\sigma_q^2 = 0.443$; dashed line $\sigma_{\tilde{u}_y}^2 = 0.162$.

that depends on the value of $\mu_i + \lambda t_i$. For example, for $t_i = 10$, the 0.1, 0.5, and the 0.9 quantiles are $\phi_i = 0.082, 0.173$, and 0.330, respectively, with a mode at $\phi_i = 0.131$. As σ_q^2 tends to zero, the distribution of ϕ_i becomes a point mass at a value of ϕ_i given by the solution to $\ln(\phi_i/[1-\phi_i]) = \mu_i + \lambda t_i$ (resulting in $\phi_i \approx 0.17$). The figure shows also the distribution of the probability of mortality evaluated at the posterior mean of the residual additive genetic variance for mortality $\hat{\sigma}_{\tilde{u}_y}^2 = 0.162$ (dashed lines).

The model specified by Equations 1 and 2 leads to a simple strategy to reduce mortality, without affecting litter size. Given the model, the residual additive genetic values of mortality are independent of the additive genetic values of litter size. The model predicts, therefore, that selecting on the basis of residual additive genetic values for mortality should not lead to correlated changes in litter size. This lack of association is supported by Figure 3, which discloses the posterior distribution of the product moment correlation between the residual additive genetic values of mortality and the additive genetic values of litter size. The value of zero is in a region of very high-density mass.

To give an idea of the likely response to selection to reduce mortality that is expected under the model, we plotted in Figure 4 the range and mean of values of the probability of mortality based on the top and bottom 20% of the distribution of the posterior means of the residual additive genetic values of mortality, for a given value of litter size. The range is governed by the selection pressure and the variability of the posterior means among individuals. For example, Figure 4 indicates that in the Yorkshire population, for a value of litter size of 14 piglets, the average probability of mortality is ~9.14%, and selecting for reduced mortality from the lowest 20% of the distribution changes this probability to 8.16% [a relative change in the proportion of mortality equal to $(9.14\% - 8.16\%)/9.14\% \approx 11\%$]. At a value of litter size of 17, the average probability of mortality is ~13.24% and selecting from the lowest 20% changes the probability to 11.88% [a relative change in the proportion of mortality equal to $(13.24\% - 11.88\%)/13.24\% \approx 10\%$]. It is of course possible to retrieve from the MCMC output draws from the marginal distribution of the additive genetic values for mortality u_{y_i} using Equation 10 and base selection on these instead.

The classical analysis involving two or more traits is based on a description of their correlation structure at the level of additive genetic and environmental correlations. In this work a recursive parameterization as in Varona *et al.* (2007) was chosen instead. In this parameterization, a one-way causal path establishes a direct effect of the size of the litter on mortality, omitting the details of the underlying nature of this relationship. A graphical representation of this relationship is in Figure 4. The models that condition the logit of mortality on litter size retrieve estimates of residual (additive genetic and permanent environmental) variances, in contrast to Model 0, which provides estimates of marginal variances. The figures in Table 1 and Table 2 illustrate this, especially for the total variance of the logit of mortality (given by $\sigma_{\tilde{u}_y}^2 + \sigma_{\tilde{p}_y}^2$). Although the signal is not strong for each of the terms taken separately, their sum is clearly larger for Model 0 (which yields instead estimates of $\sigma_{u_y}^2 + \sigma_{p_y}^2$) than for the remaining models, in both breeds.

From a practical point of view it is relevant to compare the predictive ability of this model with the one currently in operation in the Danish pig-breeding program. In the latter, the traits analyzed are total number born and total number born alive, from which parameters associated with number of piglets dead are derived. The model is based on multivariate normality and ignores the truncated nature of one of the traits (number born alive is smaller or equal to total number born). However, it has the appeal of ease of implementation,

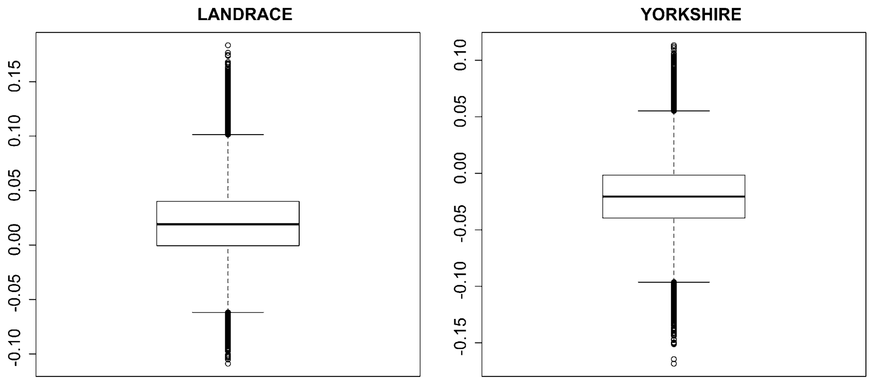

Figure 3 Monte Carlo posterior distribution (in form of boxplots) of the product moment correlation between residual additive genetic values for mortality and additive genetic values for litter size in Landrace and Yorkshire.

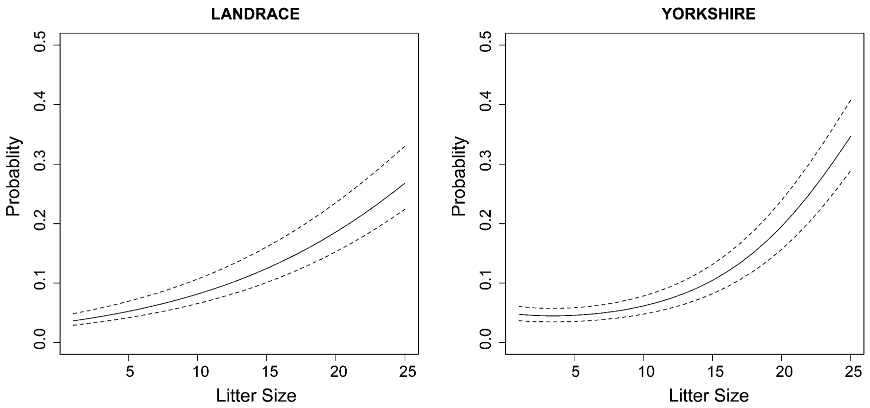

Figure 4 Posterior means of the average probability of mortality (solid lines) *vs.* litter size. For a given litter size, the range (dashed lines) is defined by the average posterior means of the probability of mortality of the top and bottom 20% selected.

from the point of view of both computing requirements and data collection. Details of this model can be found in Su *et al.* (2007) and Nielsen *et al.* (2013). The comparison presented here for each breed is based on a 10-fold cross-validation (for example, Hastie *et al.* 2009), whereby the total number of sows with records were divided into 10 groups of equal size. Phenotypic records of each fold were excluded in the training phase, predicted in the validating data set, and the correlation between observed and predicted number of dead piglets in the validating data set was computed. The predictions of the number of dead piglets using both models are obtained conditional on observed total number of piglets born. We label the model currently in operation by MVN (multivariate normal), and the binomial-normal model by BN. The average correlations (over the 10 folds) in Landrace based on MVN and on BN are 0.52 and 0.57, respectively. In Yorkshire, the correlations are 0.46 (MVN) and 0.50 (BN). The figures indicate that the BN model has a small advantage in terms of predictive ability. However, a breeding program involves a large number of traits and further work is needed, including refinements of the model to account for possible effects of the sire on mortality (Strange *et al.* 2013), to study the feasibility of incorporating the BN model into a system that can yield routine predictions of aggregate genotypes in a computationally efficient manner.

We have investigated the properties of the logit model for mortality and the Gaussian model for litter size using a modestly sized data set. In this investigation no attempt was made at exploring efficient MCMC algorithms. An implementation on a larger scale requires more attention to algorithmic details, especially if the model is extended to include dense genetic marker information. This should be the subject of future studies.

The Landrace and Yorkshire data and pedigree files as well as the FORTRAN code used for fitting the models can be found in supporting information, File S1.

Acknowledgments

The final manuscript benefited from criticism and suggestions from two reviewers and the editor. The work was funded by the Green Development and Demonstration Programme (grant no. 3405-11-0279) of the Danish Ministry of Food, Agriculture and Fisheries, the Pig Research Center, and Aarhus University.

Literature Cited

Arango, J., I. Misztal, S. Tsuruta, M. Culbertson, J. W. Holl *et al.*, 2006 Genetic study of individual preweaning mortality and birth weight in large white piglets using threshold linear models. Livest. Sci. 101: 208–218.

Blasco, A., J. P. Bidanel, and C. Haley, 1995 Genetics and neonatal survival, pp. 17–38 in *The Neonatal Pig: Development and Survival*, edited by M. A. Varley. CAB International, Wallingford, UK.

Cheverud, J. M., 1984 Quantitative genetics and developmental constraints on evolution by selection. J. Theor. Biol. 110: 155–171.

de los Campos, G., D. Gianola, P. Boettcher, and P. Moroni, 2006 A structural equation model for describing relationships between somatic cell score and milk yield in dairy goats. J. Anim. Sci. 84: 2934–2941.

Duncan, O. D., 1975 *Introduction to Structural Equation Models*. Academic Press, San Diego, CA.

Gelfand, A. E., 1996 Model determination using sampling-based methods, pp. 145–161 in *Markov Chain Monte Carlo in Practice*, edited by W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. Chapman & Hall, London.

Gianola, D., and D. Sorensen, 2004 Quantitative genetic models describing simultaneous and recursive relatiosnhips between phenotypes. Genetics 167: 1407–1424.

Goldberger, A. S., 1972 Structural equation methods in the social sciences. Econometrica 40: 979–1001.

Hastie, T., R. Tibshirani, and J. Friedman, 2009 *The Elements of Statistical Learning*. Springer, New York.

Henderson, C. R., 1984 *Applications of Linear Models in Animal Breeding*. University of Guelph, Guelph, Ontario, Canada.

Jöreskog, K. G., 1973 A general method for estimating a linear structural equation system, pp. 85–112 in *Structural Equation Models in the Social Sciences*, edited by A. S. Goldberger, and O. D. Duncan. Seminar, New York.

Lande, R., 1979 Quantitative genetic analysis of multivariate evolution, applied to brain:body allometry. Evolution 33: 402–416.

Nielsen, B., G. Su, M. S. Lund, and P. Madsen, 2013 Selection for increased number of piglets at day five after farrowing has increased litter size and reduced piglet mortality. J. Anim. Sci. 91: 2575–2582.

Roehe, R., and E. Kalm, 2000 Estimation of genetic and environmental risk factors associated with pre-weaning mortality in piglets using generalized linear mixed models. Anim. Sci. 70: 227–240.

Sorensen, D., A. Vernersen, and S. Andersen, 2000 Bayesian analysis of response to selection: a case study using litter size in Danish Yorkshire pigs. Genetics 156: 283–295.

Strange, T., B. Ask, and B. Nielsen, 2013 Genetic parameters of the piglet mortality traits stillborn, weak at birth, starvation, crushing, and miscellaneous in crossbred pigs. J. Anim. Sci. 91: 1562–1569.

Su, G., M. S. Lund, and D. Sorensen, 2007 Selection for litter size at day five to improve litter size at weaning and piglet survival rate. J. Anim. Sci. 85: 1385–1392.

Van Arendonk, J. A. M., C. Van Rosmeulen, L. L. G. Janss, and E. F. Knol, 1996 Estimation of direct and maternal genetic (co)variances for survival within litters of piglet. Livest. Prod. Sci. 46: 163–171.

Varona, L., and D. Sorensen, 2010 A genetic analysis of mortality in pigs. Genetics 184: 277–284.

Varona, L., D. Sorensen, and R. Thompson, 2007 Analysis of litter size and average litter weight in pigs using a recursive model. Genetics 177: 1791–1799.

Walsh, B., 2003 Evolutionary quantitative genetics, pp. 380–442 in *Handbook of Statistical Genetics*, Vol. 1, edited by D. J. Balding, M. Bishop, and C. Cannings. John Wiley, Chichester, UK.

Wright, S., 1921 Correlation and causation. J. Agric. Res. 210: 557–585.

Xiong, M., J. Li, and X. Fang, 2004 Identification of genetic networks. Genetics 166: 1037–1052.

Communicating editor: I. Hoeschele

Appendix

The model for the joint distribution of additive genetic and permanent environmental values for mortality and litter size

The independence of the residual additive genetic values for mortality and additive genetic values for litter size is based on the following result. A recurrent formulation for the additive genetic values for mortality and litter size for an individual is (Varona *et al.* 2007)

$$\left(u_{y_i}, u_{t_i}\right) \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{u_y} & \lambda\sigma^2_{u_t} \\ \lambda\sigma^2_{u_t} & \sigma^2_{u_t} \end{pmatrix}\right], \tag{A1}$$

where λ is a recurrent parameter that describes the linear relationship between mortality and litter size. Then,

$$\left(u_{y_i} \middle| u_{t_i}\right) \sim N\left[\lambda u_{t_i}, \sigma^2_{u_y} - \lambda^2 \sigma^2_{u_t}\right].$$

We can write

$$u_{y_i} = E\left(u_{y_i} \middle| u_{t_i}\right) + \tilde{u}_{y_i}, \tag{A2}$$

where $\tilde{u}_{y_i}$ is the residual term in the additive genetic regression of mortality on litter size and is referred to as the residual additive genetic value of mortality. Then,

$$\left(\tilde{u}_{y_i}, u_{t_i}\right) \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\tilde{u}_y} & 0 \\ 0 & \sigma^2_{u_t} \end{pmatrix}\right], \tag{A3}$$

where $\sigma^2_{\tilde{u}_y} = \sigma^2_{u_y} - \lambda^2 \sigma^2_{u_t}$. The covariance matrix of the joint distribution of the vectors $\tilde{u}_y$ and u_t is $G \otimes A$, where $G = \mathrm{diag}(\sigma^2_{\tilde{u}_y}, \sigma^2_{u_t})$ is the diagonal covariance matrix of (A3). Therefore the joint distribution factors into the product of the marginal distributions; that is,

$$p\left(\tilde{u}_y, u_t \middle| A, \sigma^2_{\tilde{u}_y}, \sigma^2_{u_t}\right) = p\left(\tilde{u}_y \middle| A, \sigma^2_{\tilde{u}_y}\right) p\left(u_t \middle| A, \sigma^2_{u_t}\right).$$

In the case of the permanent environmental values, the starting point is

$$\left(p_{y_i}, p_{t_i}\right) \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{p_y} & \lambda\sigma^2_{p_t} \\ \lambda\sigma^2_{p_t} & \sigma^2_{p_t} \end{pmatrix}\right], \tag{A4}$$

which in a similar manner leads to

$$\left(\tilde{p}_{y_i}, p_{t_i}\right) \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\tilde{p}_y} & 0 \\ 0 & \sigma^2_{p_t} \end{pmatrix}\right]. \tag{A5}$$

Sketch of the MCMC algorithm

The fully posterior distributions of the parameters, except for the variance components, do not have closed forms. For example, updating the *j*th element of α_y requires the computation of its fully conditional posterior distribution

$$\begin{aligned} p\left(\alpha_{y,j} \middle| \mathrm{all}, y, t\right) \propto \prod_{i=1}^{n} & \left[\frac{\exp\left(x_i'\alpha_y + z_i'\tilde{u}_y + w_i'p_y + g(t_i)\right)}{1+\exp\left(x'\alpha_y + z'_i\tilde{u}_y + w_i'p_y + g(t_i)\right)}\right]^{y_i I\left(x_i'\alpha_y = \alpha_{y,j}\right)} \\ & \times \left[1 - \frac{\exp\left(x_i'\alpha_y + z_i'\tilde{u}_y + w_i'p_y + g(t_i)\right)}{1+\exp\left(x_i'\alpha_y + z_i'\tilde{u}_y + w_i'p_y + g(t_i)\right)}\right]^{(t_i - y_i) I\left(x_i'\alpha_y = \alpha_{y,j}\right)}, \end{aligned} \tag{A6}$$

where $I(\cdot)$ is the indicator function that takes the value 1 if the argument is satisfied, and zero otherwise. This distribution is not of standard form and an updating strategy based on a uniform random walk Metropolis–Hastings algorithm was chosen. A similar strategy was adopted to update the residual additive genetic and permanent environmental effects.

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.159475/-/DC1

Joint Analysis of Binomial and Continuous Traits with a Recursive Model: A Case Study Using Mortality and Litter Size of Pigs

Luis Varona and Daniel Sorensen

DOI: 10.1534/genetics.113.159475

File S1

Datasets and computer code

Available for download as a .zip file at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.159475/-/DC1