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

Luis Varona¹ and Daniel Sorensen²

¹Departamento de Anatomía, Embriología y Genética Animal, Facultad de Veterinaria, Universidad de Zaragoza, E-50013, Spain; ²Department of Molecular Biology and Genetics, Aarhus University, PB 50, DK-8830 Tjele, Denmark

Abstract

The present 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 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.

INTRODUCTION

Mixed linear models (Henderson, 1984) are broadly used in livestock and plant breeding and play an important role in evolutionary and theoretical quantitative genetics (Cheverud, 1984; Lande, 1979; 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 non-genetic 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,
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 mostly determined 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.

Litter size has been under selection in the Danish pig breeding programme 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 programme 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 on 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 either normality of the sampling model for mortality (eg Van Arendonk et al., 1996), based inferences on a variety of threshold models (eg 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 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 parameterisation implemented has the attractive feature that the joint posterior distribution of the two traits factorises into two independent posterior distributions, one for each trait, whereby the computational burden of implementation is reduced.

The paper is organised as follows. Section 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 presentation of 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 (mortality hereinafter, treated as a trait of the mother) and total number of piglets born (litter size hereinafter, treated as a trait of the mother) are analysed jointly using a model that exploits the factorisation of their joint distribution. The conditional binomial model for mortality in litter \( i \), \( Y_i \), given litter size in litter \( i \), \( t_i \) is

\[
  f (Y_i = y_i|t_i, \varphi_i) = \left( \frac{t_i}{y_i} \right)^{y_i} (1 - \varphi_i)^{t_i - y_i} \quad Y_i = 0, 1, \ldots, t_i, \tag{1}
\]

where \( \varphi_i \) 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 \( \varphi_i \) is assumed to be equal to

\[
  \text{logit } \varphi_i = x'_i \alpha_y + z'_i \tilde{u}_y + w'_i \tilde{p}_y + g_j (t_i) \tag{2}
\]

where \( x'_i \), \( z'_i \), and \( w'_i \) are vectors of observed incidence matrices \( X \), \( Z \), and \( W \), \( \alpha_y \) 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 \( g_j (t_i) = \lambda_1 t_i \) \((j = 1 \text{ for MODEL 1})\), \( g_2 (t_i) = \lambda_1 t_i + \lambda_2 t_i^2 \) \((j = 2 \text{ for MODEL 2})\), \( g_3 (t_i) = \lambda_1 t_i + \lambda_2 t_i^2 + \lambda_3 t_i^3 \) \((j = 3 \text{ for MODEL 3})\), where the \( \lambda \)'s are recurrent parameters. In the case of MODEL 1 it is easy to see that two possible partitions are

\[
  \text{logit } \varphi_i = x'_i \alpha_y + z'_i \tilde{u}_y + w'_i \tilde{p}_y + \lambda_1 t_i = x'_i \alpha_y + z'_i u_y + w'_i p_y + \lambda_1 x'_i \alpha_t + \lambda_1 e_t_i,
\]

where \( u_y_i = E (u_{yi}|u_{ti}) + \tilde{u}_y \) and \( p_y_i = E (p_{yi}|p_{ti}) + \tilde{p}_y \) (see Appendix) represent draws from the marginal distributions of additive genetic values and permanent environmental effects affecting mortality.
Given vectors of systematic effects $\alpha_t$ (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

$$t_i|\alpha_t, u_t, p_t \sim N(x_i'\alpha_t + z_i'u_t + w_i'p_t, \sigma_t^2)$$

(3)

In the Appendix it is shown that the structure of $(\tilde{p}_{yi}, p_t)$ is

$$(\tilde{p}_{yi}, p_t) \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{\tilde{p}_y}^2 & 0 \\ 0 & \sigma_{p_t}^2 \end{pmatrix}\right)$$

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

$$(\tilde{u}_{yi}, u_t) \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{\tilde{u}_y}^2 & 0 \\ 0 & \sigma_{u_t}^2 \end{pmatrix}\right).$$

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

$$p(\tilde{u}_y, u_t|A, \sigma_{\tilde{u}_y}^2, \sigma_{u_t}^2) = p(\tilde{u}_y|A, \sigma_{\tilde{u}_y}^2)p(u_t|A, \sigma_{u_t}^2)$$

and similarly,

$$p(\tilde{p}_y, p_t|\sigma_{\tilde{p}_y}^2, \sigma_{p_t}^2) = p(\tilde{p}_y|\sigma_{\tilde{p}_y}^2)p(p_t|\sigma_{p_t}^2).$$

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

The variance component parameters, the recurrent parameters and the vectors $\alpha_y$ and $\alpha_t$ 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, is

$$p\left(\alpha_y, \tilde{u}_y, \tilde{p}_y, \alpha_t, u_t, p_t, \lambda_1, \sigma_{\tilde{u}_y}^2, \sigma_{\tilde{p}_y}^2, \sigma_{u_t}^2, \sigma_{p_t}^2, \sigma_{e_t}^2|y,t\right) \propto f\left(y|t, \alpha_y, \tilde{u}_y, \tilde{p}_y, \lambda_1\right)$$

$$p\left(t|\alpha_t, u_t, p_t\right)p(\tilde{u}_y|\sigma_{\tilde{u}_y}^2)p(\tilde{p}_y|\sigma_{\tilde{p}_y}^2)p(\tilde{p}_y|\sigma_{\tilde{p}_y}^2)$$

(4)
where \( y \) and \( t \) are vectors with elements \( y_i \) and \( t_i \), respectively and

\[
f(y|t, \alpha, \tilde{u}_y, \tilde{p}_y, \lambda_1, \lambda_2, \lambda_3) = \prod_{i=1}^{n} f(Y_i = y_i | t_i, \phi_i) = \prod_{i=1}^{n} \left( t_i \right)^{\phi_i} (1 - \phi_i)^{t_i - y_i},
\]

\[
p(t|\alpha_t, u_t, p_t) = (2\pi \sigma_{\alpha_t}^2)^{-\frac{n}{2}} \exp \left[ -\frac{1}{2\sigma_{\alpha_t}^2} (t - X\alpha_t + Zu_t + WP_t)'(t - X\alpha_t + Zu_t + WP_t) \right].
\]

Notice that in (4), the joint posterior distribution factorises 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 \( y' = (y, y'_{-i}) \), where \( y_i \) is the \( i \)th datum, and \( y_{-i} \) is the vector of data with the \( i \)th datum deleted. The conditional predictive distribution has probability density

\[
p(y_i | y_{-i}) = \int p(y_i | \theta_i, y_{-i}) f(\theta | y_{-i}) d\theta,
\]

\[
\theta = \{\theta_i\}_{i=1}^{n},
\]

where \( \theta \) is the vector of parameters and can be interpreted as the probability of each data point given the remainder of the data. The actual value of \( p(y_i | y_{-i}) \) is known as the *conditional predictive ordinate* (CPO) for the \( i \)th observation. The pseudo log-marginal probability of the data is given by

\[
\sum_i \ln p(y_i | y_{-i}),
\]

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

\[
\hat{p}(y_i | y_{-i}, M_k) = N \left[ \sum_{j=1}^{N} \frac{1}{p(y_i | \theta^{(j)}_i, M_k)} \right]^{-1},
\]

where \( N \) is the number of McMC draws, \( M_k \) is a label for model \( k \), and \( \theta^{(j)}_i \) is the \( j \)th draw from the posterior of \( \theta_i \) under model \( k \) corresponding to the \( i \)th observation. The so called LogCPO’s reported below are based on

\[
\sum_i \ln \hat{p}(y_i | y_{-i}, M_k).
\]

**McMC algorithm:**

The fully conditional posterior distributions associated with mortality do not all have closed forms, except for the variance components \( \sigma_{\tilde{u}_v}^2 \) and \( \sigma_{\tilde{p}_v}^2 \), which are scaled inverted chi-squared 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 \( \alpha_y \) the uniform was also centered at the current value and the bounds were given by \( \pm 0.15 \). In the case of the recursion parameters, the bounds were as follows: for \( \lambda_1 \), \( \pm 0.015 \), for \( \lambda_2 \), \( \pm 0.0015 \), and for \( \lambda_3 \), \( \pm 0.00015 \). 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 approximately 750 to 3,600 in Landrace and from 400 to 4,600 in Yorkshire. For the best fitting models, the effective chain sizes associated with the regression parameter(s) were 3,700 in Landrace (MODEL 1), and in Yorkshire, they varied from 40 to 90 (MODEL 3).

**Data:**

[Table 1 about here.]

Data were obtained from an existing data base 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 5 generations or more. For Landrace, the data comprised records from 5,178 litters and a pedigree file of 8,800 individuals. The Yorkshire data consisted of records from 3,938 litters and a pedigree file of 7,143 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**

[Figure 1 about here.]

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 below 3 and beyond 23 in Landrace, and 4 and 21 in Yorkshire, with less than 20 observations. The figures indicate that the proportions
increase non-linearly with the size of the litters, but the relationships are a little different in the two breeds (this is more clearly visualised 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 LogCPO (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 LogCPO indicate 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 approximately 37% of the total variance of the logit of mortality (total variance is equal to \( \sigma^2_u + \sigma^2_p = 0.443 \)) 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 approximately 0.20 in both breeds (calculated from (9), using estimates of posterior means of the additive genetic variance for litter size, \( \hat{\sigma}^2_u \approx 0.8 \), retrieves an estimate of the genetic correlation \( \lambda_1 (\hat{\sigma}_u / \hat{\sigma}_y) \approx 0.09 (0.80/0.16)^{0.5} = 0.20 \)). Similar calculations show that the estimate of the correlation between permanent environmental effects is approximately 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.

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 4 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 MODEL 1, 2 and 3 are all larger than 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 both, 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 the present work. In contrast with the models implemented by Varona and Sorensen (2010), in the present work the logit of the probability of mortality is assumed to be functionally related to litter size. Both linear and non-linear 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 non linear 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 generalised 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 \( g_1(t_i) = \lambda_1 t_i \),

\[
\ln \left( \frac{\varphi_i}{1 - \varphi_i} \right) = \mu_i + \lambda t_i
\]

where \( \mu_i \) is the mean of the \( i \)th 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 \( \lambda \) (0.094), translates into a value of the probability

\[
\hat{\varphi}_i = \frac{\exp(\hat{\mu}_i + \hat{\lambda} t_i)}{1 + \exp(\hat{\mu}_i + \hat{\lambda} t_i)}
\]

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 \( i \)th record is

\[
\ln \left( \frac{\varphi_i}{1 - \varphi_i} \right) = \mu_i + \lambda t_i + q_i
\]

where the random effect \( q_i \sim N(0, \sigma_q^2) \) is the sum of the residual additive genetic effect and of the residual permanent environmental effect, with \( \sigma_q^2 = \sigma_{uq}^2 + \sigma_{p_q}^2 \). Given \( \mu_i \) and \( t_i \) the probability \( \varphi_i \) is a function of \( q_i \),

\[
\varphi_i = f(q_i) = \frac{\exp(\mu_i + \lambda t_i + q_i)}{1 + \exp(\mu_i + \lambda t_i + q_i)}.
\]
The inverse function is \( f^{-1}(\varphi_i) = \ln\left(\frac{\varphi_i}{1-\varphi_i}\right) - \mu_i + \lambda t_i \) and the Jacobian is equal to \((\varphi_i(1-\varphi_i))^{-1}\). Therefore the probability density of \( \varphi_i \) is

\[
p(\varphi_i) = \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left[-\frac{\left(\ln\left(\frac{\varphi_i}{1-\varphi_i}\right) - \mu_i + \lambda t_i\right)^2}{2\sigma_q^2}\right] \frac{1}{\varphi_i(1-\varphi_i)}.
\] (8)

Figure 2 displays the distribution of \( \varphi_i \) for \( t_i = 10 \) where \( \sigma_q^2 \) 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 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 \( \varphi_i = 0.082, 0.173, \) and 0.330, respectively, with a mode at \( \varphi_i = 0.131 \). As \( \sigma_q^2 \) tends to zero the distribution of \( \varphi_i \) becomes a point mass at a value of \( \varphi_i \) given by the solution to \( \ln\left(\frac{\varphi_i}{1-\varphi_i}\right) = \mu_i + \lambda t_i \) (resulting in \( \varphi_i \approx 0.17 \)). The figure shows also the distribution of the probability of mortality evaluated at the posterior mean of the residual additive genetic variance for mortality \( \hat{\sigma}_{\hat{u}}^2 = 0.162 \) (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 that 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 equal to 14 piglets, the average probability of mortality is approximately equal to 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 equal to 17, the average probability of mortality is approximately equal to 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 $u_y$, using equation (10) and base selection on these instead.

[Figure 4 about here.]

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 parameterisation as in VARONA et al. (2007) was chosen instead. In this parameterisation, 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 that provides estimates of marginal variances. The figures in Tables 1 and 2 illustrate this, especially for the total variance of the logit of mortality (given by $\sigma^2_u + \sigma^2_p$). Although the signal is not strong for each of the terms taken separately, their sum is clearly larger for MODEL 0 (that yields instead estimates of $\sigma^2_u + \sigma^2_p$) than for the remaining models, in both breeds.

From a practical point of view it is relevant to compare the predictive ability of the present model with the one currently in operation in the Danish pig breeding program. In the latter, the traits analysed 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, both from the point of view of 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, 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 dataset. 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 the Supplementary files.
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)

$$(u_y, u_t) \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{u_y} & \lambda \sigma^2_{u_t} \\ \lambda \sigma^2_{u_t} & \sigma^2_{u_t} \end{pmatrix} \right), \tag{9}$$

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

$$(u_y | u_t) \sim N \left( \lambda u_t, \sigma^2_{u_y} - \lambda^2 \sigma^2_{u_t} \right). \tag{10}$$

We can write

$$u_y = E(u_y | u_t) + \tilde{u}_y,$$

where $\tilde{u}_y$ 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,

$$(\tilde{u}_y, u_t) \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\tilde{u}_y} & 0 \\ 0 & \sigma^2_{u_t} \end{pmatrix} \right) \tag{11}$$

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 $G = diag \left( \sigma^2_{\tilde{u}_y}, \sigma^2_{u_t} \right)$ is the diagonal covariance matrix of (11). Therefore the joint distribution factors into the product of the marginal distributions; that is

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

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

$$(p_y, p_t) \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{p_y} & \lambda \sigma^2_{p_t} \\ \lambda \sigma^2_{p_t} & \sigma^2_{p_t} \end{pmatrix} \right), \tag{12}$$

which in a similar manner leads to

$$(\tilde{p}_y, p_t) \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\tilde{p}_y} & 0 \\ 0 & \sigma^2_{p_t} \end{pmatrix} \right). \tag{13}$$

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 $\alpha_y$ requires the
computation of its fully conditional posterior distribution

\[
p(\alpha_{y,j}|\text{all}, y, t) \propto \prod_{i=1}^{n} \left[ \frac{\exp(x_i' \alpha_y + z_i' \tilde{u}_y + w_i' p_y + g(t_i))}{1 + \exp(x_i' \alpha_y + z_i' \tilde{u}_y + w_i' p_y + g(t_i))} \right]^{y_i I(x_i' \alpha_y = \alpha_{y,j})} \\
\left[ 1 - \frac{\exp(x_i' \alpha_y + z_i' \tilde{u}_y + w_i' p_y + g(t_i))}{1 + \exp(x_i' \alpha_y + z_i' \tilde{u}_y + w_i' p_y + g(t_i))} \right]^{(t_i - y_i) I(x_i' \alpha_y = \alpha_{y,j})}
\]

(14)

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.

**Acknowledgements** 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. The final manuscript benefited from criticism and suggestions from two reviewers and the editor.
LITERATURE CITED


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. Journal of Animal Science Accepted.


LIST OF FIGURES

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

2 Probability distribution of the probability of mortality for $t_i = 10$. Continuous lines, for $\sigma^2_q = 0.443$; dashed lines for $\sigma^2_{\tilde{u}} = 0.162$).  

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.  

4 Posterior means of the average probability of mortality (full line) versus 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.
Figure 1: Raw phenotypic averages of the proportion of dead born piglets as a function of litter size, for Landrace and for Yorkshire.
Figure 2: Probability distribution of the probability of mortality for \( t_i = 10 \). Continuous lines, for \( \sigma^2_q = 0.443 \); dashed lines for \( \sigma^2_{u_y} = 0.162 \).
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 (full line) versus 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.
LIST OF TABLES

1  Posterior means and standard deviations (in brackets) of variance components, recursive parameters and of LogCPO in Landrace, for mortality. . . 22
2  Posterior means and standard deviations (in brackets) of variance components, recursive parameters and of LogCPO in Yorkshire, for mortality. . . 23
Table 1: Posterior means and standard deviations (in brackets) of variance components, recursive parameters and of LogCPO in Landrace, for mortality.

<table>
<thead>
<tr>
<th></th>
<th>MODEL 0$^\dagger$</th>
<th>MODEL 1</th>
<th>MODEL 2</th>
<th>MODEL 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2_{u_y}$</td>
<td>0.168(0.035)</td>
<td>0.162(0.032)</td>
<td>0.162(0.031)</td>
<td>0.157(0.031)</td>
</tr>
<tr>
<td>$\sigma^2_{p_y}$</td>
<td>0.344(0.032)</td>
<td>0.281(0.028)</td>
<td>0.283(0.028)</td>
<td>0.288(0.029)</td>
</tr>
<tr>
<td>$\lambda_1$</td>
<td>0.094(0.005)</td>
<td>0.087(0.017)</td>
<td>0.075(0.016)</td>
<td></td>
</tr>
<tr>
<td>$\lambda_2 \times 10^{-3}$</td>
<td></td>
<td>0.278(0.570)</td>
<td>1.530(1.010)</td>
<td></td>
</tr>
<tr>
<td>$\lambda_3 \times 10^{-4}$</td>
<td></td>
<td></td>
<td>-0.314(0.248)</td>
<td></td>
</tr>
<tr>
<td>LogCPO</td>
<td>-10164</td>
<td>-9930</td>
<td>-9934</td>
<td>-9949</td>
</tr>
</tbody>
</table>

$^\dagger$ For MODEL 0 $\sigma^2_{u_y}=\sigma^2_{u_y}$ and $\sigma^2_{p_y}=\sigma^2_{p_y}$
Table 2: Posterior means and standard deviations (in brackets) of variance components, recursive parameters and of LogCPO in Yorkshire, for mortality.

<table>
<thead>
<tr>
<th></th>
<th>MODEL 0</th>
<th>MODEL 1</th>
<th>MODEL 2</th>
<th>MODEL 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_{u_y}^2$</td>
<td>0.170(0.044)</td>
<td>0.165(0.046)</td>
<td>0.163(0.045)</td>
<td>0.168(0.053)</td>
</tr>
<tr>
<td>$\sigma_{p_y}^2$</td>
<td>0.585(0.051)</td>
<td>0.494(0.049)</td>
<td>0.484(0.049)</td>
<td>0.487(0.053)</td>
</tr>
<tr>
<td>$\lambda_1$</td>
<td>0.121(0.007)</td>
<td>-0.0094(0.0362)</td>
<td>-0.191(0.046)</td>
<td></td>
</tr>
<tr>
<td>$\lambda_2 \times 10^{-2}$</td>
<td>0.473(0.129)</td>
<td>1.943(0.306)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda_3 \times 10^{-3}$</td>
<td>-0.371(0.076)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LogCPO</td>
<td>-6363</td>
<td>-6185</td>
<td>-6181</td>
<td>-6176</td>
</tr>
</tbody>
</table>

1 For MODEL 0 $\sigma_{u_y}^2 = \sigma_{u_y}^2$ and $\sigma_{p_y}^2 = \sigma_{p_y}^2$.