A zero-inflated Poisson model with correlated parameters and application to animal breeding

Mariana Rodrigues-Motta[1,*], Daniel Gianola[2,3], and Bjørg Heringstad[3]

[1]Department of Statistics, University of Campinas,

Rua Sergio Buarque de Holanda 651, Campinas, São Paulo $13083 - 859$, Brazil

[2]Department of Animal Sciences, University of Wisconsin,

1675 Observatory Drive, Madison, Wisconsin 53706, U.S.A.

[3]Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences,

P.O. Box 5003, N-1432 Ås, Norway

*email: marianar@ime.unicamp.br

SUMMARY: Lambert (1992, *Technometrics* **34**, $1 - 14$) described the zero-inflated Poisson (ZIP) regression, a class of models for count data with an excess of zeros. In this paper, Lambert's methodology is extended to accommodate correlated genetic effects in the regression structure of the Poisson and mixture parameters. In addition, an inter correlation structure between these random genetic effects is introduced, and used to infer pleiotropy, an expression of the extent to which the mixture and Poisson parameters are influenced by common genes. The methods described here are implemented and illustrated with data on number of mastitis cases from Norwegian Red cows. Bayesian analysis yields posterior distributions useful for studying environmental and genetic variability, as well as genetic correlation. The model is assessed using posterior predictive checks.

KEY WORDS: Animal breeding; Genetic effects; Pleiotropy; Variance components; Zero-inflated Poisson models.

1. Introduction

Some important traits in animal breeding are recorded as count data, for example, litter size in pigs, embryo yield produced after superovulation, and number of mastitis (a mammary gland disease) cases in dairy cattle. Sometimes, the number of zeros exceeds the amount expected under a certain sampling model, as for example, the Poisson distribution. A possibility for modeling extra-zeros is the zero-inflated Poisson (ZIP) model as proposed by Lambert (1992). If a ZIP model is to be used in genetic animal breeding studies, it is necessary to accommodate genetic covariances between effects. This study proposes to model the mixture and Poisson parameters of a ZIP model hierarchically, each as a function of two random effects, representing genetic and environmental sources of variability, respectively. The random genetic effects are assumed correlated, to incorporate resemblance between relatives, while the environmental effects are treated as independent. As pleiotropy is a main cause of genetic correlation between traits (Falconer, 1989), a correlation structure is introduced between the genetic effects affecting the mixture and Poisson parameters, similar to the correlation between direct genetic and maternal genetic effects (Willham, 1963). This genetic correlation (measure of pleiotropy) expresses the extent to which the mixture and Poisson parameters may be influenced by common genes. In a ZIP model, the mixture parameter p is interpreted as the "perfect state" probability, and a negative correlation between genetic effects affecting p and the Poisson parameter would be expected, meaning that genes in favor of the perfect state decrease the expected number of counts. A correlation close to zero would indicate that there are no common genes affecting these parameters simultaneously.

A hierarchical ZIP model with correlated parameters is developed and applied to data on number of mastitis cases in Norwegian Red cows in first lactation. First, a hierarchical Bayes structure is presented. Second, a Markov chain Monte Carlo method for drawing samples

from posterior distributions is described. Third, a model checking is conducted via an analysis of residuals and of predictive ability.

2. The ZIP model with correlated parameters

Let $\mathbf{y} = (y_1, y_2, ..., y_n)'$ be a vector containing the number of cases of an event per animal (e.g., number of mastitis cases in dairy cows). It is assumed that, given some parameters λ_i and p_i, the distribution of observation y_i on animal i follows a zero-inflated Poisson distribution, and that all such observations are conditionally independent. The density is then

$$p(y_i = k|\lambda_i, p_i) = [p_i + (1 - p_i)e^{-\lambda_i}]^{I(k=0)}[(1 - p_i)e^{-\lambda_i}\lambda_i^k/k!]^{I(k>0)}, \tag{1}$$

for $i = 1, ..., n$, $0 < \lambda_i < \infty$ and $0 < p_i < 1$. Here, p_i is the probability that a 0 is from the "perfect" state, and λ_i is the parameter of the "imperfect" state (Poisson distribution). This model tends to a conditional Poisson model as $p_i \to 0$.

Let $\boldsymbol{\lambda}^* = (\log(\lambda_1), \ldots, \log(\lambda_n))'$ and $\mathbf{p}^* = (\text{logit}(p_1), \ldots, \text{logit}(p_n))'$ be vectors of unobservable parameters. Further, suppose that $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$ jointly satisfy the linear mixed model

$$\begin{bmatrix} \boldsymbol{\lambda}^* \\ \mathbf{p}^* \end{bmatrix} = \begin{bmatrix} \mathbf{X}_\lambda & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_p \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_\lambda \\ \boldsymbol{\beta}_p \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_\lambda & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_p \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_\lambda \\ \boldsymbol{\varepsilon}_p \end{bmatrix}, \tag{2}$$

where $\boldsymbol{\beta}_\lambda$ and $\boldsymbol{\beta}_p$ are vectors of fixed effects; $\mathbf{u}_1$ and $\mathbf{u}_2$ are vectors of random effects, and $\mathbf{X}_\lambda$, $\mathbf{X}_p$, $\mathbf{Z}_\lambda$ and $\mathbf{Z}_p$ are known incidence matrices. Factors included in $\boldsymbol{\beta}_\lambda$ may or may not be the same as those in $\boldsymbol{\beta}_p$. In (2), $\boldsymbol{\varepsilon}_p$ and $\boldsymbol{\varepsilon}_\lambda$ are vectors of residuals which are assumed to follow a multivariate normal distribution on R^{2n}, that is,

$$\begin{pmatrix} \boldsymbol{\varepsilon}_\lambda \\ \boldsymbol{\varepsilon}_p \end{pmatrix} |\boldsymbol{\Sigma} \sim \text{Normal}\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \boldsymbol{\Sigma} \otimes \mathrm{I}_n\right], \tag{3}$$

where $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2_{\varepsilon_\lambda} & 0 \\ 0 & \sigma^2_{\varepsilon_p} \end{pmatrix}$, $\mathbf{I}_n$ is an identity matrix of order n, and $\sigma^2_{\varepsilon_\lambda}$ and $\sigma^2_{\varepsilon_p}$ are residual variances reflecting overdispersion over and above that caused by extra zeros. The distribution of $\boldsymbol{\varepsilon}_\lambda$ and $\boldsymbol{\varepsilon}_p$ induces the distribution

$$\begin{pmatrix} \boldsymbol{\lambda}^*|\boldsymbol{\beta}_\lambda, \mathbf{u}_1, \sigma^2_{\varepsilon_\lambda} \\ \mathbf{p}^*|\boldsymbol{\beta}_p, \mathbf{u}_2, \sigma^2_{\varepsilon_p} \end{pmatrix} \sim \text{Normal} \left[\begin{pmatrix} \mathbf{X}_\lambda\boldsymbol{\beta}_\lambda + \mathbf{Z}_\lambda\mathbf{u}_1 \\ \mathbf{X}_p\boldsymbol{\beta}_p + \mathbf{Z}_p\mathbf{u}_2 \end{pmatrix}, \boldsymbol{\Sigma} \otimes \mathbf{I}_n \right], \tag{4}$$

where $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$ are as in (2).

Here, $\mathbf{u} = (\mathbf{u}'_1 \mathbf{u}'_2)' \sim N(\mathbf{0}, \boldsymbol{\Sigma}_u)$. In this study, we partition the $\mathbf{Z}$ incidence matrices as $\mathbf{Z}_\lambda = (\ \mathbf{Z}_{1,\lambda}\ \ \mathbf{0}\ \ \mathbf{Z}_{2,\lambda}\ \ \mathbf{0}\ )$ and $\mathbf{Z}_p = (\ \mathbf{0}\ \ \mathbf{Z}_{1,p}\ \ \mathbf{0}\ \ \mathbf{Z}_{2,p}\ )$, respectively; these matrices relate the random effects $\mathbf{u}_1 = (\mathbf{h}'_\lambda, \mathbf{h}'_p)'$ and $\mathbf{u}_2 = (\mathbf{a}'_\lambda, \mathbf{a}'_p)'$ to $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$. The vectors $\mathbf{h}_\lambda$ and $\mathbf{h}_p$, each of dimension n_h, are non-genetic effects (e.g., herd effects); the vectors $\mathbf{a}_\lambda$ and $\mathbf{a}_p$, each of dimension n_a, are additive genetic effects. Additionally, $\boldsymbol{\Sigma}_\mathbf{u} = \begin{pmatrix} \mathbf{H} \otimes \mathbf{I}_{n_h} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} \otimes \mathbf{A} \end{pmatrix}$, where $\mathbf{I}_{n_h}$ is an identity matrix of order n_h;

$$\mathbf{H} = \begin{pmatrix} \sigma^2_{h_\lambda} & 0 \\ 0 & \sigma^2_{h_p} \end{pmatrix} \tag{5}$$

and

$$\mathbf{G} = \begin{pmatrix} \sigma^2_{a_\lambda} & \sigma_{a_{\lambda,p}} \\ \sigma_{a_{\lambda,p}} & \sigma^2_{a_p} \end{pmatrix}. \tag{6}$$

Further, $\mathbf{A}$ is a known additive genetic relationship matrix; $\sigma^2_{h_\lambda}$ and $\sigma^2_{h_p}$ are variances of non-genetic (say herd) effects affecting $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$, respectively; $\sigma^2_{a_\lambda}$ and $\sigma^2_{a_p}$ are the additive genetic variances, and $\sigma_{a_{\lambda,p}}$ is the additive genetic covariance. The genetic correlation is $\rho = \sigma_{a_{\lambda,p}}/\sqrt{\sigma^2_{a_\lambda}\sigma^2_{a_p}}$.

2.1 *Parameter inference*

2.1.1 *The likelihood and priors.* Let $\boldsymbol{\tau} = \{\boldsymbol{\lambda}^*, \mathbf{p}^*, \boldsymbol{\beta}_\lambda, \boldsymbol{\beta}_p, \mathbf{h}_\lambda, \mathbf{h}_p, \mathbf{a}_\lambda, \mathbf{a}_p, \mathbf{H}, \mathbf{G}, \boldsymbol{\Sigma}\}$ be a set of unknown quantities. The conditional (given $\boldsymbol{\tau}$) likelihood for the ZIP regression model is

$$l(\boldsymbol{\tau}; \mathbf{y}) \;\; = \;\; \prod_{y_i=0} [p_i + (1-p_i)e^{-\lambda_i}] \prod_{y_i>0} [(1-p_i)e^{-\lambda_i}\lambda_i^{y_i}/y_i!]. \tag{7}$$

Replacing λ_i and p_i in (7) by the models in (2), the likelihood can be written as

$$
\begin{aligned}
l(\boldsymbol{\tau}; \mathbf{y}) \;\; = \;\; & \prod_{y_i=0} \left\{ \frac{e^{(\mathbf{x}'_{i;p}\boldsymbol{\beta}_p + \mathbf{z}'_{i;1,p}\mathbf{h}_p + \mathbf{z}'_{i;2,p}\mathbf{a}_p)}}{1 + e^{(\mathbf{x}'_{i;p}\boldsymbol{\beta}_p + \mathbf{z}'_{i;1,p}\mathbf{h}_p + \mathbf{z}'_{i;2,p}\mathbf{a}_p)}} \right. \\
& + \left. \frac{e^{-\exp(\mathbf{x}'_{i;\lambda}\boldsymbol{\beta}_\lambda + \mathbf{z}'_{i;1,\lambda}\mathbf{h}_\lambda + \mathbf{z}'_{i;2,\lambda}\mathbf{a}_\lambda)}}{1 + e^{(\mathbf{x}'_{i;p}\boldsymbol{\beta}_p + \mathbf{z}'_{i;1,p}\mathbf{h}_p + \mathbf{z}'_{i;2,p}\mathbf{a}_p)}} \right\} \\
& \times \prod_{y_i>0} [1 + e^{(\mathbf{x}'_{i;p}\boldsymbol{\beta}_p + \mathbf{z}'_{i;1,p}\mathbf{h}_p + \mathbf{z}'_{i;2,p}\mathbf{a}_p)}]^{-1} \\
& \times \left\{ e^{-\exp(\mathbf{x}'_{i;\lambda}\boldsymbol{\beta}_\lambda + \mathbf{z}'_{i;1,\lambda}\mathbf{h}_\lambda + \mathbf{z}'_{i;2,\lambda}\mathbf{a}_\lambda)} \right. \\
& \times \left. e^{y_i(\mathbf{x}'_{i;\lambda}\boldsymbol{\beta}_\lambda + \mathbf{z}'_{i;1,\lambda}\mathbf{h}_\lambda + \mathbf{z}'_{i;2,\lambda}\mathbf{a}_\lambda)}/y_i! \right\},
\end{aligned}
\tag{8}
$$

where $\mathbf{x}'_{i;\lambda}$, $\mathbf{z}'_{i;1,\lambda}$, $\mathbf{z}'_{i;2,\lambda}$, $\mathbf{x}'_{i;p}$, $\mathbf{z}'_{i;1,p}$ and $\mathbf{z}'_{i;2,p}$ and are the ith rows of matrices $\mathbf{X}_\lambda$, $\mathbf{Z}_{1,\lambda}$, $\mathbf{Z}_{2,\lambda}$, $\mathbf{X}_p$, $\mathbf{Z}_{1,p}$ and $\mathbf{Z}_{2,p}$, respectively. To achieve a reasonably vague prior, an uniform distribution is assigned to each element of $\boldsymbol{\beta}$, with large absolute values for the bounds $\boldsymbol{\beta}_{min,\lambda}$, $\boldsymbol{\beta}_{max,\lambda}$, $\boldsymbol{\beta}_{min,p}$, and $\boldsymbol{\beta}_{max,p}$. Independent scale inverse chi-square distributions with degrees of freedom (scale parameter) ν_ε (δ_ε) are assigned to $\sigma^2_{\varepsilon_\lambda}$ and $\sigma^2_{\varepsilon_p}$, respectively, and independent inverse chi-square distributions with degrees of freedom (scale parameter) ν_h (δ_h) are assigned to $\sigma^2_{h_\lambda}$ and $\sigma^2_{h_p}$, respectively. An inverse Wishart distribution with parameter matrix (degrees of freedom) V_G (ν_G) is assigned to $\mathbf{G}$.

2.1.2 *The joint posterior distribution.* Considering (8) and the priors described above, the joint posterior density can be written as

$$
\begin{aligned}
p(\boldsymbol{\tau}|\mathbf{y}) \;\; \propto \;\; & (8) \times (\sigma^2_{\varepsilon_\lambda})^{-\left(\frac{n+\nu_\varepsilon+2}{2}\right)} (\sigma^2_{\varepsilon_p})^{-\left(\frac{n+\nu_\varepsilon+2}{2}\right)} \\
& \times \exp\left\{ -\frac{1}{2\sigma^2_{\varepsilon_\lambda}} \left[\sum_{i=1}^{n} (\lambda^*_i - \mathbf{x}'_{i;\lambda}\boldsymbol{\beta}_\lambda - \mathbf{z}'_{i;1,\lambda}\mathbf{h}_\lambda - \mathbf{z}'_{i;2,\lambda}\mathbf{a}_\lambda)^2 \right. \right.
\end{aligned}
$$

$$
\begin{aligned}
&+\quad \nu_\varepsilon \delta_\varepsilon^2 \big]\big\} \\
&\times\quad \exp\left\{-\frac{1}{2\sigma_{\varepsilon_p}^2}\left[\sum_{i=1}^{n}(p_i^* - \mathbf{x}_{i;p}'\boldsymbol{\beta}_p - \mathbf{z}_{i;1,p}'\mathbf{h}_p - \mathbf{z}_{i;2,p}'\mathbf{a}_p)^2\right.\right. \\
&+\quad \nu_\varepsilon \delta_\varepsilon^2 \big]\big\} \\
&\times\quad (\sigma_{h_\lambda}^2)^{-\left(\frac{n_h+\nu_h+2}{2}\right)}\exp\left[-\frac{1}{2\sigma_{h_\lambda}^2}(\mathbf{h}_\lambda'\mathbf{h}_\lambda + \nu_h\delta_h^2)\right] \\
&\times\quad (\sigma_{h_p}^2)^{-\left(\frac{n_h+\nu_h+2}{2}\right)}\exp\left[-\frac{1}{2\sigma_{h_p}^2}(\mathbf{h_p}'\mathbf{h_p} + \nu_h\delta_h^2)\right] \\
&\times\quad |\mathbf{G}|^{-\left(\frac{n_a+\nu_G+3}{2}\right)}\exp[-\frac{1}{2}\mathrm{tr}(\mathbf{G}^{-1}\mathbf{V}_G^*)],
\end{aligned}
\tag{9}
$$

where

$$
\mathbf{V}_G^* \;=\; \begin{bmatrix} \mathbf{a}_\lambda'\mathbf{A}^{-1}\mathbf{a}_\lambda & \mathbf{a}_\lambda'\mathbf{A}^{-1}\mathbf{a}_p \\[2mm] \mathbf{a}_\lambda'\mathbf{A}^{-1}\mathbf{a}_p & \mathbf{a}_p'\mathbf{A}^{-1}\mathbf{a}_p \end{bmatrix} + \mathbf{V}_G^{-1}.
\tag{10}
$$

The joint posterior distribution with density as in (9) is not recognizable, and can be written only up to a proportionality constant. Also, marginal posterior distributions cannot be obtained analytically. Therefore, a Metropolis-Gibbs sampling scheme was tailored to sample from the marginal posterior distributions.

2.1.3 *Sampling the Poisson parameters* λ_i^*. From (9), the conditional posterior density of the vector of log-Poisson parameters $\boldsymbol{\lambda}^*$ given all other parameters ("ELSE") is $p(\boldsymbol{\lambda}^*|\text{ELSE},\mathbf{y}) \propto l(\boldsymbol{\tau};\mathbf{y})p(\boldsymbol{\lambda}^*|\boldsymbol{\beta}_\lambda, \mathbf{h}_\lambda, \mathbf{a}_\lambda, \sigma_{\varepsilon_\lambda}^2)$. The λ_i^*'s are assumed independent, a priori, and their prior densities are normal with parameters according to (4). Hence,

$$
\begin{aligned}
p(\lambda_i^*|\text{ELSE},\mathbf{y_i}) &\propto [e^{p^*} + e^{-\exp(\lambda_i^*)}]^{\mathbf{1}(y_i=0)}[e^{-\exp(\lambda_i^*)+\lambda_i^* y_i}]^{\mathbf{1}(y_i>0)} \\
&\times\quad \exp\left[-\frac{1}{2\sigma_{\varepsilon_\lambda}^2}(\lambda_i^* - \mathbf{x}_{i,\lambda}'\boldsymbol{\beta}_\lambda - \mathbf{z}_{i;1,\lambda}'\mathbf{h}_\lambda - \mathbf{z}_{i;2,\lambda}'\mathbf{a}_\lambda)^2\right],
\end{aligned}
\tag{11}
$$

$i = 1, 2, ..., n$. It can be seem that these fully conditional distributions are independent. The density in (11) does not have any obviously recognizable form. A Metropolis-Hastings algorithm was therefore tailored for drawing the $\lambda_i^* = \log(\lambda_i)$ parameters, one at a time.

2.1.4 *Sampling the mixture parameters* p_i^*.   From (9), the conditional posterior density of the vector of logits $\mathbf{p}^*$ is $p(\mathbf{p}^*|\text{ELSE}, \mathbf{y}) \propto l(\boldsymbol{\tau}; \mathbf{y})p(\mathbf{p}^*|\boldsymbol{\beta}_p, \mathbf{h}_p, \mathbf{a}_p, \sigma_{\varepsilon_p}^2)$. The p_i^*'s are assumed independent, a priori, and their prior densities are normal according to (4). The fully conditional distribution of these parameters are independent, i.e.,

$$
\begin{aligned}
p(p_i^*|\text{ELSE}, \mathbf{y}) \quad \propto \quad & [\mathrm{e}^{p_i^*} + \mathrm{e}^{-\exp(\lambda_i^*)}]^{\mathbf{1}(y_1=0)}[1 + \mathrm{e}^{p_i^*}]^{-1} \\
\times \quad & \exp\left[-\frac{1}{2\sigma_{\varepsilon_p}^2}(p_i^* - \mathbf{x}_{i,p}'\boldsymbol{\beta}_p - \mathbf{z}_{i;1,p}'\mathbf{h}_p - \mathbf{z}_{i;2,p}'\mathbf{a}_p)^2\right].
\end{aligned}
\tag{12}
$$

This distribution is not recognizable either. A Metropolis-Hastings algorithm was developed for drawing the p_i^*'s. Once samples of p_i^* are drawn, samples for p_i can be obtained as $p_i = \mathrm{e}^{p_i^*}/(1 + \mathrm{e}^{p_i^*})$.

2.1.5 *Sampling location effects affecting* $\boldsymbol{\lambda}^*$ *and* $\boldsymbol{p}^*$.   From (9), the fully conditional posterior distribution of the $\boldsymbol{\beta}$, $\mathbf{h}$ and $\mathbf{a}$ location parameters is

$$
p(\boldsymbol{\beta}, \mathbf{h}, \mathbf{a}|\text{ELSE}, \mathbf{y}) \quad \propto \quad p(\boldsymbol{\lambda}^*, \mathbf{p}^*|\boldsymbol{\beta}, \mathbf{h}, \mathbf{a})p(\boldsymbol{\beta})p(\boldsymbol{h}|\mathbf{H})p(\boldsymbol{a}|\mathbf{G}).
\tag{13}
$$

This is the conditional posterior density of location parameters in a bivariate Gaussian model with known dispersion structure, in which $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$ play the role of "traits", and the only source of correlation is through genetic effects. The derivation of the fully posterior distribution of the location parameters is given in Sorensen and Gianola (2002). Let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{h}', \boldsymbol{a}')'$ and

$$
\mathbf{M} = \begin{bmatrix} \mathbf{X}_\lambda & \mathbf{0} & \mathbf{Z}_{1,\lambda} & \mathbf{0} & \mathbf{Z}_{2,\lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_p & \mathbf{0} & \mathbf{Z}_{1,p} & \mathbf{0} & \mathbf{Z}_{2,p} \end{bmatrix}.
\tag{14}
$$

Note that this implies sorting of individuals within $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$, respectively. Then, the standard mixed model equations of animal breeders are given by

$$
\mathbf{C}\hat{\boldsymbol{\theta}} = \mathbf{t},
\tag{15}
$$

where $\mathbf{C} = \mathbf{M}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n)\mathbf{M} + \boldsymbol{\Omega}$, with

$$\boldsymbol{\Omega} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}^{-1} \otimes \mathbf{I}_{n_h} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{G}^{-1} \otimes \mathbf{A}^{-1} \end{pmatrix} \tag{16}$$

and $\mathbf{t} = \mathbf{M}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n) \begin{bmatrix} \boldsymbol{\lambda}^* \\ \boldsymbol{p}^* \end{bmatrix}$. The fully conditional posterior distribution of $\boldsymbol{\theta}$ is the multivariate normal process $\boldsymbol{\theta} | \text{ELSE}, \mathbf{y} \sim \text{Normal}(\hat{\boldsymbol{\theta}}, \mathbf{C}^{-1})$ and, for any sub-vector $\boldsymbol{\theta}_i$ of $\boldsymbol{\theta}$, $\boldsymbol{\theta}_i | \text{ELSE}, \mathbf{y} \sim \text{Normal}(\tilde{\boldsymbol{\theta}}_i, \mathbf{C}_{i,i}^{-1})$, where $\tilde{\boldsymbol{\theta}}_i$ satisfies $\mathbf{C}_{i,i}\tilde{\boldsymbol{\theta}}_i = \mathbf{t}_i - \mathbf{C}_{i,-i}\boldsymbol{\theta}_i$. Here, $\mathbf{C}_{i,i}$ is an appropriate sub-matrix of $\mathbf{C}$; $\mathbf{t}_i$ is the corresponding sub vector of $\mathbf{t}$; $\mathbf{C}_{i,-i}$ is a block of $\mathbf{C}$ linking the "$\boldsymbol{\theta}_i$ equations" to the "$\boldsymbol{\theta}_{-i}$ equations", and $\boldsymbol{\theta}_{-i}$ is $\boldsymbol{\theta}$ with $\boldsymbol{\theta}_i$ removed. The Gibbs sampler is implemented in a scalar mode, drawing from the appropriated fully conditional posterior distribution one element of $\boldsymbol{\theta}$ at a time. In this case, $\tilde{\boldsymbol{\theta}}_i$ and $\mathbf{C}_{i,i}$ are scalars and $\mathbf{C}_{i,-i}$ is a row vector.

2.1.6 *Conditional posterior distribution of the residual variances.* From (9), and the fact that $\boldsymbol{\Sigma}$ is a diagonal matrix, it follows that

$$\begin{aligned} p(\sigma_{\varepsilon_\lambda}^2 | \text{ELSE}, \mathbf{y}) &\propto p(\boldsymbol{\lambda}^* | \boldsymbol{\beta}_\lambda, \mathbf{h}_\lambda, \mathbf{a}_\lambda, \sigma_{\varepsilon_\lambda}^2) p(\sigma_{\varepsilon_\lambda}^2 | \nu_\varepsilon, \delta_\varepsilon) \\ &\propto (\sigma_{\varepsilon_\lambda}^2)^{-\left(\frac{n+\nu_\varepsilon+2}{2}\right)} \\ &\times \exp\left\{ -\frac{1}{2\sigma_{\varepsilon_\lambda}^2} \left[\sum_{i=1}^n (\lambda_i^* - \mathbf{x}_{i;\lambda}'\boldsymbol{\beta}_\lambda - \mathbf{z}_{i;1,\lambda}'\mathbf{h}_\lambda - \mathbf{z}_{i;2,\lambda}'\mathbf{a}_\lambda)^2 \right. \right. \\ &\quad + \left. \left. \nu_\varepsilon\delta_\varepsilon^2 \right] \right\} \end{aligned} \tag{17}$$

and

$$\begin{aligned} p(\sigma_{\varepsilon_p}^2 | \text{ELSE}, \mathbf{y}) &\propto p(\boldsymbol{p}^* | \boldsymbol{\beta}_p, \mathbf{h}_p, \mathbf{a}_p, \sigma_{\varepsilon_p}^2) p(\sigma_{\varepsilon_p}^2 | \nu_\varepsilon, \delta_\varepsilon) \\ &\propto (\sigma_{\varepsilon_p}^2)^{-\left(\frac{n+\nu_\varepsilon+2}{2}\right)} \\ &\times \exp\left\{ \frac{1}{2\sigma_{\varepsilon_p}^2} \left[\sum_{i=1}^n (p_i^* - \mathbf{x}_{i;p}'\boldsymbol{\beta}_p - \mathbf{z}_{i;1,p}'\mathbf{h}_p - \mathbf{z}_{i;2,p}'\mathbf{a}_p)^2 \right. \right. \\ &\quad + \left. \left. \nu_\varepsilon\delta_\varepsilon^2 \right] \right\}. \end{aligned} \tag{18}$$

These are the densities of two independent scaled inverse chi-square random variables. Sampling is straightforward.

2.1.7 *Conditional posterior distribution of the non-genetic variances.* From (9), it follows directly that the two non-genetic variances are conditionally independent. In particular,

$$p(\sigma_{h_\lambda}^2|\text{ELSE}, \mathbf{y}) \quad \propto \quad (\sigma_{h_\lambda}^2)^{-\left(\frac{n_h+\nu_h+2}{2}\right)}\exp\left\{-\frac{1}{2\sigma_{h_\lambda}^2}\mathbf{h}_\lambda{}'\mathbf{h}_\lambda + \nu_h\delta_h^2\right\} \tag{19}$$

and

$$p(\sigma_{h_p}^2|\text{ELSE}, \mathbf{y}) \quad \propto \quad (\sigma_{h_p}^2)^{-\left(\frac{n_h+\nu_h+2}{2}\right)}\exp\left\{-\frac{1}{2\sigma_{h_p}^2}\mathbf{h_p}{}'\mathbf{h_p} + \nu_h\delta_h^2\right\}. \tag{20}$$

These are densities of two independent scale inverse chi-square random variables.

2.1.8 *Conditional posterior distribution of the genetic covariance matrix* $\mathbf{G}$. From (9) it follows directly that

$$p(\mathbf{G}|\text{ELSE}, \mathbf{y}) \quad \propto \quad |\mathbf{G}|^{-\left(\frac{n_a+\nu_G+3}{2}\right)}\exp\left\{-\frac{1}{2}\text{tr}(\mathbf{G}^{-1}\mathbf{V}_G^*)\right\},$$

where $\mathbf{V}_G^*$ is given in (10). Hence, the conditional posterior distribution of $\mathbf{G}$ is the 2-dimensional inverse Wishart process $\mathbf{G}|\text{ELSE}, \mathbf{y} \sim \text{IW}_2(n_a + \nu_G, \mathbf{V}_G^{*-1})$.

2.2 *The Gibbs-Metropolis algorithm and convergence criteria*

Our Gibbs-Metropolis sampling algorithm consisted of cyclic sampling through all components of $\boldsymbol{\tau}$, drawing each parameter or subset of parameters, conditionally on the realized value of all other parameters, at each iteration of the algorithm. At iteration t, an ordering of the components of $\boldsymbol{\tau}$ was chosen and elements of $\boldsymbol{\tau}$ were sampled sequentially from their conditional distribution, given the current value of all other elements of $\boldsymbol{\tau}$. A normal distribution with mean equal to the value of $\lambda_i^{*(t)}$ at iteration (t) and appropriate variance was used as proposal distribution for sampling from (11). Similarly, a normal distribution with mean equal to the value of $p_i^{*(t)}$ at iteration (t) and some appropriate variance was used

as proposal distribution for sampling from (12). The variances of the proposal distributions were chosen to attain acceptance rates between 30% and 50% (Gelman et al., 2004).

Visual inspection of trace plots of the MCMC run and the scale reduction factor diagnostic suggested by Gelman and Rubin (1992) were used to determine the length of the burn-in period and the total number of iterations for the Gibbs-Metropolis procedure. Two chains with overdispersed starting points were used in the Gelman and Rubin (1992) method. This monitors convergence of the iterative simulation by estimating the factor by which the scale of the current distribution for a parameter under study, say τ, might be reduced if simulations were continued for an infinite amount of time. The potential scale reduction is given by $\hat{R} = \sqrt{\frac{\hat{\text{var}}(\tau|\mathbf{y})}{W}}$, which declines towards 1 as the number of iterations J goes to infinity. Here, $\hat{\text{var}}(\tau|\mathbf{y}) = \frac{J-1}{J}W + \frac{1}{J}B$, where W and B are estimates of the within and between-chain variances. Discarding early draws as burn-in, such that the starting value is "forgotten", samples are drawn as needed to attain a sufficiently small Monte Carlo error of estimation of features of the posterior distribution, such as the posterior mean.

2.3 *Discrepancy statistic for model adequacy*

The adequacy of the ZIP model fitted to the data was assessed by comparing the observed value of a statistic $T_k(\mathbf{y}, \boldsymbol{\tau})$ with its predictive distribution under the ZIP model. As a measure of "discrepancy", the statistic $D_k = T_k(\mathbf{y}|\boldsymbol{\tau}) - T_k(\mathbf{y}_{rep}|\boldsymbol{\tau})$ was used. Here,

$$T_k(\mathbf{y}, \boldsymbol{\tau}) = \frac{\sum_{i=1}^n I(y_i = k)}{n}, \tag{21}$$

where y_i is the ith component of the observed vector $\mathbf{y}$ and $k = 0, 1, \ldots$. Additionally,

$$T_{k,l}(\mathbf{y}_{rep,l}, \boldsymbol{\tau}) = \frac{\sum_{i=1}^n I(y_{i;rep,l} = k)}{n}, \tag{22}$$

where $y_{i;rep,l}$ is the ith component of the replicated vector $\mathbf{y}_{rep}$ of size n in sample $l = 1, \ldots, L$. In (21) and (22), $I(.)$ is an indicator function. A distribution of D_k values was generated for each value of k; if the model holds, D_k should be centered at 0. Computations were as follows:

1) For $L = 100$, vectors $\mathbf{v} = (v_1, ..., v_n)$ of size n = number of individuals in the data set were drawn. Each element v_i of $\mathbf{v}$ followed a ZIP density evaluated at the posterior mean of λ_i^* and p_i^*, respectively, as inputs for the Poisson and mixture parameters. 2) For each realization of $\mathbf{v}$ at sample l, $\frac{\sum_{i=1}^n I(y_{i;rep,l=k})}{n}$ was calculated at each k, leading to 100 "future" relative frequencies. 3) D_k was calculated for each k, with values far from zero being interpreted as evidence against the model.

2.4 *Residual analysis*

A residual for the ith record was defined as $r_i = y_i - E(y_i|\tau)$. The posterior mean of the standardized residual was estimated as

$$\hat{r}_i = \frac{1}{J} \sum_{j=1}^J \frac{(y_i - E(y_i|\tau^{(j)}))}{\sqrt{Var(y_i|\tau^{(j)})}}, \tag{23}$$

$i = 1, ..., n$, and J being the number of posterior samples. The expectation and variance used were $E(y_i|\tau^{(j)}) = (1 - p_i)\lambda_i$ and $Var(y_i|\tau^{(j)}) = E(y_i|\tau^{(j)})(1 + \lambda_i p_i)$, respectively. An observation would be unusual if the posterior distribution of r_i is concentrated away from zero.

3. An animal breeding application

Quantitative genetic analysis of mastitis data has been carried out mainly with linear models (e.g., Carlén, Schneider and Strandberg, 2005) and with threshold models (Gianola and Foulley, 1983, Heringstad et al., 2001, Heringstad et al., 2004, Chang et al., 2004), with the latter ones accounting for the binary structure of the data, at least when mastitis is categorized as "absent" or "present". Longitudinal binary response models have been used as well (Heringstad et al., 2003). However, the number of episodes of a disease is a random count, and a more appropriate sampling model would be the Poisson distribution. Further, the "zero count" (e.g., no disease) may have higher frequency than what is expected under Poisson sampling, so that a ZIP model might be suitable. If so, an extension for quantitative genetics

analysis of counts with an excess of zeros is of interest. This application was investigate alternative specifications for modeling number of incidences of mastitis via a ZIP model, and for making inferences about genetic (co)variation between the Poisson and mixture parameters involved in modeling.

The hierarchical ZIP model was fitted to a data set consisting of number of mastitis cases in $36,175$ Norwegian first-lactation Red cows. The data is described in detail in Rodrigues-Motta et al. (2007). The ZIP model in terms of (2) included 4 "fixed" factors affecting both $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$. Here, $\boldsymbol{\beta}_\lambda = (\boldsymbol{\beta}_{1,\lambda}, \boldsymbol{\beta}_{2,\lambda}, \boldsymbol{\beta}_{3,\lambda}, \beta_{4,\lambda})'$ and $\boldsymbol{\beta}_p = (\boldsymbol{\beta}_{1,p}, \boldsymbol{\beta}_{2,p}, \boldsymbol{\beta}_{3,p}, \beta_{4,p})'$, respectively. The vector $\boldsymbol{\beta}_{1,j}$ included effects of 15 ages at first calving ($< 20, 20, 21, 22, ..., 32, > 32$ months), the vector $\boldsymbol{\beta}_{2,j}$ included effects of 12 calendar months at first calving, $\boldsymbol{\beta}_{3,j}$ consisted of effects of 3 years at first calving (from 1990 to 1992), and $\beta_{4,j}$ was a regression on the logarithm of the number of days from first calving to the defined end of first lactation (culling, second calving, or 300 days after calving, whichever occured first), with $j = \lambda, p$. To achieve a reasonably vague prior, each element of $\boldsymbol{\beta}_\lambda$ and $\boldsymbol{\beta}_p$ was sampled from an uniform distribution spanning from -999 to 999. Herd effects represented the non-genetic random factor contained in $\mathbf{h}_\lambda$ and $\mathbf{h}_p$, respectively. The non-genetic herd effects, represented by $\mathbf{h} = (\mathbf{h}'_\lambda \mathbf{h}'_p)'$, were assumed to follow a priori a multivariate normal distribution with mean zero and covariance matrix as in (5), where I_{n_h} is an identity matrix of dimension $n_h = 5,286$. A ZIP "sire" model was fitted, thus $\mathbf{a}_\lambda$ and $\mathbf{a}_p$ are vectors containing half of the breeding values affecting λ_i^* and p_i^*, respectively; each of these vectors was of order 437×1. The sire (genetic) effects, represented by $\mathbf{a} = (\mathbf{a}'_\lambda \mathbf{a}'_p)'$, were assumed to follow a multivariate normal distribution with mean zero and covariance matrix $\mathbf{G} \otimes \mathbf{A}$, where $\mathbf{G}$ is as in (6), and $\mathbf{A}$ is a known matrix of additive genetic relationships of dimension 437. The $\mathbf{A}$ matrix was built from a sire pedigree file with a total of 437 males, where the pedigree of the 245 sires with daughters in the data set were traced back, through sires and maternal grandsires, as far back as possible. In quantitative

genetics theory, variation between sires accounts only for 1/4 of the total additive genetic variation (Falconer, 1989); the rest of the variation (genetic and environmental) is captured by the residual terms $\boldsymbol{\varepsilon}_\lambda$ and $\boldsymbol{\varepsilon}_p$. Therefore, the residual term in the model accounts for more overdispersion, and for environmental effects beyond those due to herds. The degrees of belief parameters of the scale inverse chi-squared and inverse Wishart distributions assigned as priors for variance components and matrix $\mathbf{G}$ were $\nu_\varepsilon = \nu_h = \nu_G = 5$. Further,

$$V_G = \begin{pmatrix} 0.45 & -0.025 \\ & 0.45 \end{pmatrix}, \tag{24}$$

and $\delta_\varepsilon = \delta_h = 1$. A Fortran program was written to sample the unknowns, following the scheme proposed in Section 2.2. The Gelman and Rubin (1992) convergence criterion used 2 chains starting from overdispersed values, with 10^6 iterations and a burn-in period of 5×10^5 samples. The scale reduction factors for the residual, herd and sire variances affecting $\boldsymbol{\lambda}^*$ were 1.05, 1 and 1, respectively; the scale reduction factors for the residual, herd and sire variances affecting $\mathbf{p}^*$ were 1.04, 1.05 and 1, respectively. The scale reduction factor for the sire covariance was 1. These values suggest convergence to the equilibrium distribution. However, the trace plots (results not shown) indicated that additional iterations would produce more accurate posterior estimates of the residual variance associated to $\boldsymbol{\lambda}^*$, and of all variance components associated to $\mathbf{p}^*$. A total of $500,000$ after-burn-in samples (without thinning) from one of the two chains was used to calculate Monte Carlo errors associated to the posterior mean of the variance components. The Monte Carlo error variances of residual, herd and sire variances affecting $\boldsymbol{\lambda}^*$ were 2.1×10^{-8}, 5.8×10^{-9} and 8×10^{-10}, respectively; the Monte Carlo error of variances of residual, herd and sire variances affecting $\mathbf{p}^*$ were 1.7×10^{-7}, 2.3×10^{-7} and 1.7×10^{-7}, respectively. The Monte Carlo error variance of the covariance between sire effects was 5.7×10^{-9}. These small Monte Carlo errors indicated that posterior mean estimates were precise enough.

The posterior distributions of the dispersion components affecting $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$ are given

in Figures 1 and Figure 2, respectively. The posterior means and standard deviation (SD) of the residual, herd and sire variances affecting $\boldsymbol{\lambda}^*$ were 0.76 (0.04), 0.36 (0.02) and 0.07 (0.01), respectively, with the most important source of variation being that due to residual effects, followed by herds and then by sires. In Figure 1, the posterior distributions of the residual and herd variances were nearly symmetric, while the posterior distribution of the sire variance was slightly asymmetric with a longer tail to the right. The posterior means (SD) of the residual, herd and sire variances affecting $\boldsymbol{p}^*$ were 0.4 (0.11), 0.45 (0.13) and 0.27 (0.11), respectively, with the largest source of variation being herd effects, followed by residuals and then sires. As shown in Figure 2, the posterior distributions of the residual, herd and sire variances affecting $\mathbf{p}^*$ were all skewed, with long tails to the right. These results suggest that it is more difficult to infer components of variance precisely for $\mathbf{p}^*$ than for $\boldsymbol{\lambda}^*$.

[Figure 1 about here.]

[Figure 2 about here.]

The posterior distribution of the covariance (correlation) between sire effects on $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$ is shown in Figure 1, with the mean (SD) of the sire covariance being 0.01 (0.02). The 90% credible interval is given by $(-0.02; 0.05)$, suggesting that genetic effects affecting $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$ are uncorrelated. There was large uncertainty about the sire correlation (which is equal to the genetic correlation, under additive inheritance at the $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$ levels). The posterior distribution assigned high density to values of the correlation varying from -0.4 to 0.4.

As shown in Figure 2, the distribution (over observations) of the posterior means of the perfect state probabilities $\mathbf{p}$ was skewed, and its mean (SD) was 0.07 (0.11); the 5% and 95% percentiles yielded the 90% interval $(0.02; 0.22)$. This leads to the inference that, on average, about 7% of first-lactation cows would not get mastitis, either due to being totally resistant to the disease or for never being exposed to mastitis.

The posterior predictive densities of the discrepancy statistic D_k displayed in Figure 3

(left panel) indicated that the predictive ability of the ZIP model is not entirely satisfactory. However, for number of cases of mastitis less or equal to 2, the agreement between observed and replicate data was close. The plot of posterior means of residuals displayed in Figure 3 (right panel) suggested that the model fitted the zero counts reasonably, but that it was less successful for fitting number of mastitis cases larger than 0 because the mean residual values were far from zero.

[Figure 3 about here.]

For comparative purposes, a Poisson model having the same exploratory structure for $\boldsymbol{\lambda}^*$ was fitted to the data. In the Poisson model, the posterior means (SD) of the residual, herd and sire variances affecting $\boldsymbol{\lambda}^*$ were 0.9 (0.03), 0.35 (0.02) and 0.05 (0.01), respectively. The residual variance was larger in the Poisson than in the ZIP model, suggesting that the overdispersion due to zeros was absorbed by the residual term in the Poisson model. The ZIP model captured more genetic variation, since the variation between sires was larger than in the Poisson model (posterior means were 0.07 and 0.05, respectively). Since the genetic correlation between $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$ was near zero, another ZIP model with $\sigma_{a_{\lambda,p}} = 0$ was fitted. The mean (SD) of the posterior distribution of $\boldsymbol{p}$ in the uncorrelated and correlated ZIP models were 0.1 (0.11) and 0.07 (0.11), respectively. Estimates of the variance components affecting $\boldsymbol{p}^*$ and $\boldsymbol{\lambda}^*$ are summarized in Table 1. In general, estimates were similar, except for $\sigma_{a_p}^2$, where the mean was 37% larger (0.37 vs 0.27) in the uncorrelated ZIP model. Since the credible interval for the covariance between sire effects included zero, the principle of parsimony favors the uncorrelated ZIP model.

[Table 1 about here.]

4. Discussion

Count data models have been developed for animal breeding applications, which pose either a Poisson mixed effects model (Foulley et al., 1987) or accommodate "extra-Poisson" residual variation explicitly (Tempelman and Gianola, 1996). However, part of this extra variation may be due to extra zeros relative to Poisson sampling. In this case, a ZIP model may provide a better fit to the data (Lambert, 1992). From an animal breeding perspective, quantities of main interest are the genetic values of candidates for selection and associated variance components. Here, the ZIP model was extended to accommodate genetic effects by introducing correlated random effects in the structure of the log-Poisson parameter ($\boldsymbol{\lambda}^*$) and of the logit of the mixture probability ($\mathbf{p}^*$). This structure is analogous to that of a multiple-trait linear model described, for example, in Sorensen and Gianola (2002). Moreover, a correlation between these two genetic effects would account for pleiotropic genes affecting the Poisson and the mixture probability, as in models in which a correlation between direct and maternal effects is fitted (Willham, 1963). The hierarchical structures posed for $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$ would permit to discriminate between individuals being resistance to a certain disease and those that are mildly liable.

In an application of this model to number of mastitis cases in first-lactation Norwegian Red cows, it seemed that a Poisson regression model absorbed overdispersion due to zeros in the residual term. If this is so, the Poisson mixed model would produce poorer estimates of the variance components. In the ZIP model, the components of variance affecting $\mathbf{p}^*$ were inferred less precisely than those affecting $\boldsymbol{\lambda}^*$.

Although the scale reduction factor value proposed by Gelman and Rubin (1992) was satisfied as a convergence criterion, trace plots (not shown here) suggested that, in the case of variance components affecting $\mathbf{p}^*$, additional iterations were needed for convergence. However, this can be computationally very intensive. Mixing might be improved by switching

sampling order of the unknowns at each iteration of the Gibbs-Metropolis algorithm as this would reduce the serial correlation between successively sampled quantities. The posterior correlation structure indicated a high correlation between the genetic variance affecting $\boldsymbol{\lambda}^*$ and the genetic covariance (0.93); between the genetic variance affecting $\mathbf{p}^*$ and genetic covariance (0.86), and between the genetic variances affecting $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$ (0.63). Another form of improving mixing would be via a blockwise Gibbs-Metropolis sampler, as proposed by Gelman et al. (2004).

The fully conditional distributions of $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$ are not recognizable, so a Metropolis-Hastings scheme was needed to sample the appropriate unknowns. We found that a normal proposal distribution had a better performance than a random-walk proposal. It would be of interest to examine the performance of the Metropolis-Hastings scheme under different proposal distributions. Besag, York and Mollie (1991), working in a different problem, suggested that although a proper flat prior leads to a proper joint posterior distribution, it may produce a singularity invalidating the Gibbs sampler. This problem was not detected here, but a zero-mean normal prior with a large variance may be a better option.

References

Besag, J., York, J. and Mollie, A. (1991). Bayesian image restoration with two applications in spacial statistics. *Annals of the Institute of Statistics and Mathematics* **43,** 1–59.

Carlén, E., Schneider, M. del P. and Strandberg, E. (2005). Comparison between linear models and survival analysis for genetic evaluation of clinical mastitis in dairy cattle. *Journal of Dairy Science* **88,** 797–803.

Chang, Y. M., Gianola, D., Heringstad, B. and Klemetsdal, G. (2004). Effects of trait definition on genetic parameter estimates and sire evaluation for clinical mastitis threshold models. *Animal Science* **79,** 355–363.

Falconer, D. S. (1989). *Introduction to Qunatitative Genetics*, 3rd. edition. New York: Willey.

Foulley, J. L., Gianola, D. and Im, S. (1987). Genetic evaluation of traits distributed as Poisson-binomial with reference to reproductive characters. *Theoretical and Applied Genetics* **73,** 870–877.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7,** 457–511.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis*. London: Chapman and Hall.

Gianola, D. and Foulley, J. L. (1983). Sire evaluation for ordered categorical data with a threshold model. *Genetic, Selection, Evolution* **15,** 201–224.

Heringstad, B., Rekaya, R., Gianola, D., Klemetsdal, G. and Weigel, K. A. (2001). Bayesian analysis of liability of clinical mastitis in Norwegian cattle with a threshold model: effects of data sampling method and model specification. *Journal of Dairy Science* **84,** 2337–2346.

Heringstad, B., Chang, Y. M., Gianola, D. and Klemetsdal, G. (2003). Genetic analysis of longitudinal trajectory of clinical mastitis in first-lactation Norwegian cattle. *Journal of Dairy Science* **86,** 2676–2683.

Heringstad, B., Chang, Y. M., Gianola, D. and Klemetsdal, G. (2004). Multivariate threshold model analysis of clinical mastitis in multiparous Norwegian dairy cattle. *Journal of*

Dairy Science **87,** 3038–3046.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34,** 1–14.

Rodrigues-Motta, M., Gianola, D. , Heringstad, B., Rosa, G. J. M. and Chang. Y. M. (2007). A Zero-Inflated Poisson Model for Genetic Analysis of the Number of Mastitis Cases in Norwegian Red Cows. *Journal of Dairy Science* **90,** 5306–5315.

Sorensen, D. and Gianola, D. (2002). *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics.* Springer.

Tempelman, R. J. and Gianola, D. (1996). A mixed effects model for overdispersed count data in animal breeding. *Biometrics* **52,** 265–279.

Willham, R. L. (1963). The covariance between relatives for characters composed of components contributed by related individuals. *Biometrics* **19,** 18–27.
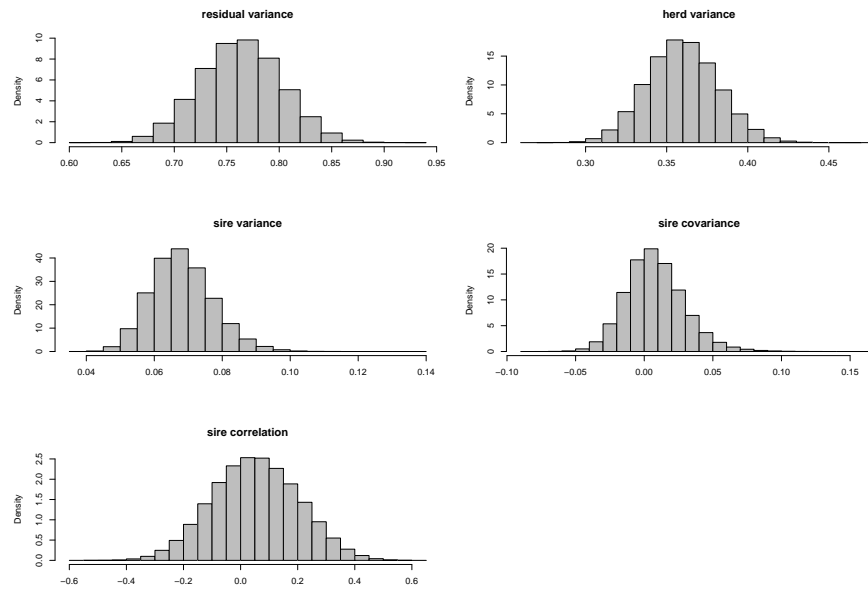
Figure 1. Posterior distributions of the dispersion components affecting $\boldsymbol{\lambda}^*$ and of sire covariances and correlations between $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$.
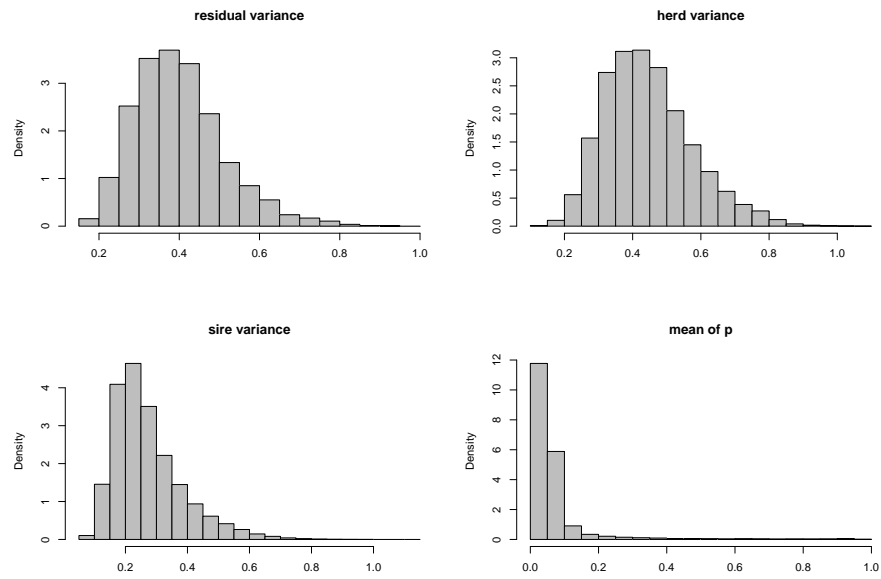
Figure 2. Posterior distribution of the dispersion components affecting $\mathbf{p}^*$ and distribution of the posterior means of the probability of the perfect state ($\mathbf{p}$).
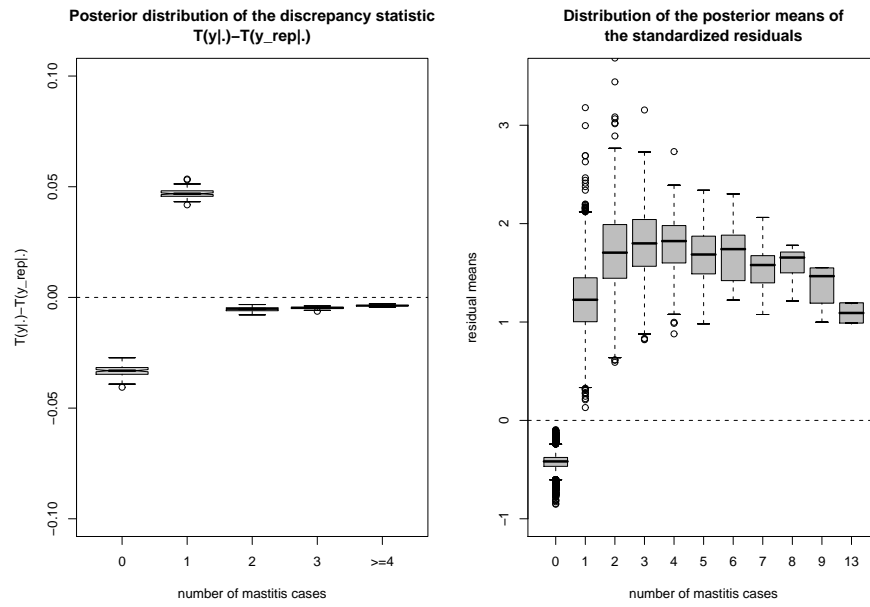
Figure 3.   Distribution of the discrepancy statistic, and of posterior means of the residuals under the correlated ZIP model.

Table 1
Inference on residual $(\sigma^2_{\varepsilon_\lambda}, \sigma^2_{\varepsilon_p})$, herd $(\sigma^2_{h_\lambda}, \sigma^2_{h_p})$ and sire $(\sigma^2_{a_\lambda}, \sigma^2_{a_p})$ variances affecting $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$ in ZIP models with or without genetic correlation between $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$.

	Posterior mean (standard error)	
	Uncorrelated ZIP	Correlated ZIP
$\sigma^2_{\varepsilon_\lambda}$	0.72 (0.04)	0.76 (0.04)
$\sigma^2_{h_\lambda}$	0.36 (0.02)	0.36 (0.02)
$\sigma^2_{a_\lambda}$	0.09 (0.01)	0.07 (0.01)
$\sigma^2_{\varepsilon_p}$	0.36 (0.10)	0.40 (0.11)
$\sigma^2_{h_p}$	0.41 (0.13)	0.45 (0.13)
$\sigma^2_{a_p}$	0.37 (0.10)	0.27 (0.11)