Journal of Data Science **8**(2010), 379-396

A Mixed Effects Model for Overdispersed Zero Inflated Poisson Data with an Application in Animal Breeding

Mariana Rodrigues-Motta[1], Daniel Gianola[2,3] and Bjørg Heringstad[3]
[1] *University of Campinas-UNICAMP,* [2] *DUniversity of Wisconsin and* [3] *Norwegian University of Life Sciences*

Abstract: Response variables that are scored as counts, for example, number of mastitis cases in dairy cattle, often arise in quantitative genetic analysis. When the number of zeros exceeds the amount expected such as under the Poisson density, the zero-inflated Poisson (ZIP) model is more appropriate. In using the ZIP model in animal breeding studies, it is necessary to accommodate genetic and environmental covariances. For that, this study proposes to model the mixture and Poisson parameters hierarchically, each as a function of two random effects, representing the genetic and environmental sources of variability, respectively. The genetic random effects are allowed to be correlated, leading to a correlation within and between clusters. The environmental effects are introduced by independent residual terms, accounting for overdispersion above that caused by extra-zeros. In addition, an inter correlation structure between random genetic effects affecting mixture and Poisson parameters is used to infer pleiotropy, an expression of the extent to which these parameters are influenced by common genes. The methods described here are 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.

Key words: Count data, overdispersion, random effects, ZIP models.

1. Introduction

Some traits in animal breeding are scored as counts, for example, litter size in pigs, embryo yield produced after superovulation, and number of mastitis cases in dairy cattle. When the number of zeros exceeds the amount expected under a certain density, as for example, the Poisson density, a possibility for modeling the extra-zeros has been proposed by Lambert (1992). In using the ZIP model in animal breeding studies, it is necessary to accommodate genetic and environmental covariances. For that, this study proposes to model the mixture

and Poisson parameters hierarchically, each as a function of two random effects, representing the genetic and environmental sources of variability, respectively.

Models for zero-inflated count data with random effects accounting for intra-group correlation and dependence of clustered observations either in the logistic regression model of the mixture parameter and/or the in log-linear model of the Poisson parameter have been discussed by Welsh *et al.* (1996), Hall *et al.* (2000), Wang *et al.* (2002), Kuhnert *et al.* (2005), and Min and Agresti (2005). In this article, modeling genetic effects of zero-inflated count data presents special challenges; in addition of the problem of extra zeros, the correlation within and between clusters, e.g., half-sibs families, needs to be taken into account. We explore a model where the correlated genetic random effects not only accommodate within but between cluster correlation. The ZIP model accomodates overdispersion (variance > mean) caused by extra zeros, however independent environmental effects (residual effects) accommodate overdispersion above that. Both genetic and environmental random effects are independent, and considered at the level of the mixture and Poisson parameters. As pleiotropy is the main genetic cause of correlation (Falconer, 1989), a correlation structure is also introduced between the genetic effects on these parameters, similar to the correlation between direct and maternal effects (Willham, 1963). The degree of correlation from pleiotropy express the extent to which the mixture and Poisson parameters are influenced by common genes. In a ZIP model, the mixture parameter p is interpreted as the "perfect state" probability (Rodrigues-Motta *et al.*, 2007), so a negative correlation between p and the Poisson parameter would be expected, meaning that genes in favor of the perfect state act against to the imperfect Poisson state. In the same way, a correlation close to zero indicates that there are no common genes affecting those parameters simultaneously.

A hierarchical ZIP model with correlated parameters is developed to an application to number of mastitis cases in Norwegian Red cows in first lactation. First, a hierarchical Bayes structure is presented. Second, estimation of the parameters are suggested via a Gibbs-sampling approach. Third, a model checking was conducted via an analysis of residuals and predictive ability.

2. The Mixed Effects ZIP Model with Independent Residual Effects

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 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

$$p(y_i = k|\lambda_i, p_i) = [p_i + (1 - p_i)e^{-\lambda_i}]^{I(k=0)}[(1 - p_i)e^{-\lambda_i}\lambda_i^k/k!]^{I(k>0)}, \qquad (2.1)$$

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 $\mathbf{p}^*$ satisfy the linear mixed model

$$\begin{bmatrix} \boldsymbol{\lambda}^* \\ \mathbf{p}^* \end{bmatrix} = \begin{bmatrix} \mathbf{X}_\lambda & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_p \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_\lambda \\ \boldsymbol{\beta}_p \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_\lambda & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_p \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_\lambda \\ \boldsymbol{\varepsilon}_p \end{bmatrix}, \quad (2.2)$$

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 (matrices of 0 and $1'$s). Factors included in $\boldsymbol{\beta}_\lambda$ may or may not be the same as those in $\boldsymbol{\beta}_p$.

In (2.2), $\boldsymbol{\varepsilon}_p$ and $\boldsymbol{\varepsilon}_\lambda$ are vectors of residuals which are assumed to follow a multivariate normal distribution on R^{2n} with mean zero and covariance matrix $\boldsymbol{\Sigma} \otimes \mathrm{I}_n$, where $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2_{\varepsilon_\lambda} & \mathbf{0} \\ \mathbf{0} & \sigma^2_{\varepsilon_p} \end{pmatrix}$, $\mathbf{I}_n$ 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

$$\begin{pmatrix} \boldsymbol{\lambda}^* | \boldsymbol{\beta}_\lambda, \mathbf{u}_1, \sigma^2_{\varepsilon_\lambda} \\ \mathbf{p}^* | \boldsymbol{\beta}_p, \mathbf{u}_2, \sigma^2_{\varepsilon_p} \end{pmatrix} \sim \text{Normal} \left[\begin{pmatrix} \mathbf{X}_\lambda \boldsymbol{\beta}_\lambda + \mathbf{Z}_\lambda \mathbf{u}_1 \\ \mathbf{X}_p \boldsymbol{\beta}_p + \mathbf{Z}_p \mathbf{u}_2 \end{pmatrix}, \boldsymbol{\Sigma}_u \otimes \mathbf{I}_n \right], \quad (2.3)$$

where $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$ are as in (2.2).

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

$$\mathbf{H} = \begin{pmatrix} \sigma^2_{h_\lambda} & 0 \\ 0 & \sigma^2_{h_p} \end{pmatrix} \text{ and } \mathbf{G} = \begin{pmatrix} \sigma^2_{a_\lambda} & \sigma_{a_{\lambda,p}} \\ \sigma_{a_{\lambda,p}} & \sigma^2_{a_p} \end{pmatrix}.$$

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 estimation

The likelihood, priors and the joint posterior distribution

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

$$l(\boldsymbol{\tau}; \mathbf{y}) = \prod_{y_i=0} [p_i + (1-p_i)e^{-\lambda_i}] \prod_{y_i>0} [(1-p_i)e^{-\lambda_i}\lambda_i^{y_i}/y_i!]. \tag{2.4}$$

To achieve a reasonably vague prior, an uniform distribution is assigned to each element of $\boldsymbol{\beta}$'s, with large absolute values of 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 $\mathbf{G}$. Replacing (2.4) by the models in (2.2), the joint posterior density can be written as

$$\begin{aligned}
p(\boldsymbol{\tau}|\mathbf{y}) \;\propto\;& l(\boldsymbol{\tau}; \mathbf{y}) \times (\sigma^2_{\varepsilon_\lambda})^{-\left(\frac{n+\nu_\varepsilon+2}{2}\right)} (\sigma^2_{\varepsilon_p})^{-\left(\frac{n+\nu_\varepsilon+2}{2}\right)} \\
\times\;& \exp\left\{ -\frac{1}{2\sigma^2_{\varepsilon_\lambda}} \left[\sum_{i=1}^{n} (\lambda_i^* - \mathbf{x}'_{i;\lambda}\boldsymbol{\beta}_\lambda - \mathbf{z}'_{i;1,\lambda}\mathbf{h}_\lambda - \mathbf{z}'_{i;2,\lambda}\mathbf{a}_\lambda)^2 \right.\right. \\
+\;& \left.\left. \nu_\varepsilon \delta_\varepsilon^2 \right] \right\} \\
\times\;& \exp\left\{ -\frac{1}{2\sigma^2_{\varepsilon_p}} \left[\sum_{i=1}^{n} (p_i^* - \mathbf{x}'_{i;p}\boldsymbol{\beta}_p - \mathbf{z}'_{i;1,p}\mathbf{h}_p - \mathbf{z}'_{i;2,p}\mathbf{a}_p)^2 \right.\right. \\
+\;& \left.\left. \nu_\varepsilon \delta_\varepsilon^2 \right] \right\} \\
\times\;& (\sigma^2_{h_\lambda})^{-\left(\frac{n_h+\nu_h+2}{2}\right)} \exp\left[-\frac{1}{2\sigma^2_{h_\lambda}} (\mathbf{h}_\lambda{}' \mathbf{h}_\lambda + \nu_h \delta_h^2) \right] \\
\times\;& (\sigma^2_{h_p})^{-\left(\frac{n_h+\nu_h+2}{2}\right)} \exp\left[-\frac{1}{2\sigma^2_{h_p}} (\mathbf{h_p}{}' \mathbf{h_p} + \nu_h \delta_h^2) \right] \\
\times\;& |\mathbf{G}|^{-\left(\frac{n_a+\nu_G+3}{2}\right)} \exp[-\frac{1}{2}\mathrm{tr}(\mathbf{G}^{-1}\mathbf{V}_G^*)],
\end{aligned} \tag{2.5}$$

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 i^{th} 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, and

$$\mathbf{V}_G^* = \begin{bmatrix} \mathbf{a}'_\lambda \mathbf{A}^{-1} \mathbf{a}_\lambda & \mathbf{a}'_\lambda \mathbf{A}^{-1} \mathbf{a}_p \\ \mathbf{a}'_\lambda \mathbf{A}^{-1} \mathbf{a}_p & \mathbf{a}'_p \mathbf{A}^{-1} \mathbf{a}_p \end{bmatrix} + \mathbf{V}_G^{-1}.$$

The joint posterior distribution with density as in (2.5) 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.

Sampling the parameters λ_i^* and p_i^*

From (2.5), the conditional posterior density of the vector of log-Poisson parameters $\boldsymbol{\lambda}^*$ is $p(\boldsymbol{\lambda}^*|\text{ELSE}, \mathbf{y}) \propto l(\boldsymbol{\tau}; \mathbf{y}) p(\boldsymbol{\lambda}^*|\boldsymbol{\beta}_\lambda, \mathbf{h}_\lambda, \mathbf{a}_\lambda, \sigma^2_{\varepsilon_\lambda})$, given all other parameters ("ELSE"). The λ_i^*'s are assumed independent, a priori, and their prior densities are normal with parameters according to (2.3). Hence,

$$\begin{aligned} p(\lambda_i^*|\text{ELSE}, \mathbf{y_i}) &\propto [\mathrm{e}^{p^*} + \mathrm{e}^{-\exp(\lambda_i^*)}]^{\mathbf{1}(y_i=0)} [\mathrm{e}^{-\exp(\lambda_i^*)+\lambda_i^* y_i}]^{\mathbf{1}(y_i>0)} \\ &\times \exp\left[-\frac{1}{2\sigma^2_{\varepsilon_\lambda}}(\lambda_i^* - \mathbf{x}'_{i,\lambda}\boldsymbol{\beta}_\lambda - \mathbf{z}'_{i;1,\lambda}\mathbf{h}_\lambda - \mathbf{z}'_{i;2,\lambda}\mathbf{a}_\lambda)^2\right], \quad (2.6) \end{aligned}$$

$i = 1, 2, ..., n$. It can be seen that these fully conditional distributions are independent. From (2.5), the conditional posterior density of the vector of logits $\mathbf{p}^*$ is $p(\boldsymbol{p}^*|\text{ELSE}, \mathbf{y}) \propto l(\boldsymbol{\tau}; \mathbf{y}) p(\boldsymbol{p}^*|\boldsymbol{\beta}_p, \mathbf{h}_p, \mathbf{a}_p, \sigma^2_{\varepsilon_p})$. The p_i^*'s are assumed independent, a priori, and their prior densities are normal according to (2.3). The fully conditional distribution of these parameters are independent, i.e.,

$$\begin{aligned} p(p_i^*|\text{ELSE}, \mathbf{y}) &\propto [\mathrm{e}^{p_i^*} + \mathrm{e}^{-\exp(\lambda_i^*)}]^{\mathbf{1}(y_1=0)} [1 + \mathrm{e}^{p_i^*}]^{-1} \\ &\times \exp\left[-\frac{1}{2\sigma^2_{\varepsilon_p}}(p_i^* - \mathbf{x}'_{i,p}\boldsymbol{\beta}_p - \mathbf{z}'_{i;1,p}\mathbf{h}_p - \mathbf{z}'_{i;2,p}\mathbf{a}_p)^2\right]. \quad (2.7) \end{aligned}$$

The densities in (2.6) and (2.7) do not have any obviously recognizable form. A Metropolis-Hastings algorithm was therefore tailored for drawing the $\lambda_i^* = \log(\lambda_i)$ and $p_i = \mathrm{e}^{p_i^*}/(1 + \mathrm{e}^{p_i^*})$, one at a time.

Sampling location effects affecting $\boldsymbol{\lambda}^*$ and $\boldsymbol{p}^*$

From (2.5), the fully conditional posterior distribution of the $\boldsymbol{\beta}$, $\mathbf{h}$ and $\mathbf{a}$ location parameters is $p(\boldsymbol{\beta}, \mathbf{h}, \mathbf{a}|\text{ELSE}, \mathbf{y}) \propto p(\boldsymbol{\lambda}^*, \mathbf{p}^*|\boldsymbol{\beta}, \mathbf{h}, \mathbf{a}) p(\boldsymbol{\beta}) p(\boldsymbol{h}|\mathbf{H}) p(\boldsymbol{a}|\mathbf{G})$. 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

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

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 $\mathbf{C}\hat{\boldsymbol{\theta}} = \mathbf{t}$, where $\mathbf{C} = \mathbf{M}^{'}(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n)\mathbf{M} + \boldsymbol{\Omega}$, with

$$\boldsymbol{\Omega} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}^{-1} \otimes \mathbf{I}_{n_h} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{G}^{-1} \otimes \mathbf{A}^{-1} \end{pmatrix} \quad \text{and} \quad \mathbf{t} = \mathbf{M}^{'}(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n) \begin{bmatrix} \boldsymbol{\lambda}^* \\ \boldsymbol{p}^* \end{bmatrix}.$$

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.

Conditional posterior distribution of the residual variances

From (2.5), and the fact that $\boldsymbol{\Sigma}$ is a diagonal matrix, it follows that

$$\begin{aligned} p(\sigma^2_{\varepsilon_\lambda}|\text{ELSE}, \mathbf{y}) &\propto p(\boldsymbol{\lambda}^*|\boldsymbol{\beta}_\lambda, \mathbf{h}_\lambda, \mathbf{a}_\lambda, \sigma^2_{\varepsilon_\lambda}) p(\sigma^2_{\varepsilon_\lambda}|\nu_\varepsilon, \delta_\varepsilon) \\ &\propto (\sigma^2_{\varepsilon_\lambda})^{-\left(\frac{n+\nu_\varepsilon+2}{2}\right)} \\ &\times \exp\left\{-\frac{1}{2\sigma^2_{\varepsilon_\lambda}}\left[\sum_{i=1}^{n}(\lambda^*_i - \mathbf{x}^{'}_{i;\lambda}\boldsymbol{\beta}_\lambda - \mathbf{z}^{'}_{i;1,\lambda}\mathbf{h}_\lambda - \mathbf{z}^{'}_{i;2,\lambda}\mathbf{a}_\lambda)^2\right.\right. \\ &+ \left.\left. \nu_\varepsilon\delta^2_\varepsilon\right]\right\} \end{aligned} \tag{2.8}$$

and

$$\begin{aligned} p(\sigma^2_{\varepsilon_p}|\text{ELSE}, \mathbf{y}) &\propto p(\boldsymbol{p}^*|\boldsymbol{\beta}_p, \mathbf{h}_p, \mathbf{a}_p, \sigma^2_{\varepsilon_p}) p(\sigma^2_{\varepsilon_p}|\nu_\varepsilon, \delta_\varepsilon) \\ &\propto (\sigma^2_{\varepsilon_p})^{-\left(\frac{n+\nu_\varepsilon+2}{2}\right)} \\ &\times \exp\left\{\frac{1}{2\sigma^2_{\varepsilon_p}}\left[\sum_{i=1}^{n}(p^*_i - \mathbf{x}^{'}_{i;p}\boldsymbol{\beta}_p - \mathbf{z}^{'}_{i;1,p}\mathbf{h}_p - \mathbf{z}^{'}_{i;2,p}\mathbf{a}_p)^2\right.\right. \\ &+ \left.\left. \nu_\varepsilon\delta^2_\varepsilon\right]\right\}. \end{aligned} \tag{2.9}$$

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

Conditional posterior distribution of the non-genetic variances

From (2.5), it follows directly that the two non-genetic variances are conditionally independent. In particular,

$$p(\sigma^2_{h_\lambda}|\text{ELSE},\mathbf{y}) \quad \propto \quad (\sigma^2_{h_\lambda})^{-\left(\frac{n_h+\nu_h+2}{2}\right)}\exp\left\{-\frac{1}{2\sigma^2_{h_\lambda}}\mathbf{h}_\lambda{}'\mathbf{h}_\lambda+\nu_h\delta^2_h\right\} \quad (2.10)$$

and

$$p(\sigma^2_{h_p}|\text{ELSE},\mathbf{y}) \quad \propto \quad (\sigma^2_{h_p})^{-\left(\frac{n_h+\nu_h+2}{2}\right)}\exp\left\{-\frac{1}{2\sigma^2_{h_p}}\mathbf{h_p}{}'\mathbf{h_p}+\nu_h\delta^2_h\right\}. \quad (2.11)$$

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

Conditional posterior distribution of genetic variance matrix G

From (2.5) it follows directly that

$$p(\mathbf{G}|\text{ELSE},\mathbf{y}) \quad \propto \quad |\mathbf{G}|^{-\left(\frac{n_a+\nu_G+3}{2}\right)}\exp\left\{-\frac{1}{2}\text{tr}(\mathbf{G}^{-1}\mathbf{V}^*_G)\right\}, \quad (2.12)$$

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

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 (2.6). 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 (2.7). The variances of the proposal distributions were chosen to attain acceptance rates between 30% and 50% (Gelman *et al.*, 2004). For the proposed model, a possible implementation of the algorithm at each step t is as follows, where $\boldsymbol{\tau}^0$ would be some starting value:

- Sample $[\lambda^*_{i,t}|\text{ELSE}_{t-1}, \mathbf{y}]$ via a Metropolis step applied to (2.6), $i = 1, ..., n$.
- Sample $[p^*_{i,t}|\text{ELSE}_{t-1}, \lambda^*_{i,t}, \mathbf{y}]$ via a Metropolis step applied to (2.7), $i = 1, ..., n$.
- Sample $[\boldsymbol{\theta}_t|\text{ELSE}_{t-1}, \boldsymbol{\lambda}^*_t, \mathbf{p}^*_t, \mathbf{y}]$ from the multivariate normal distribution with mean $\hat{\boldsymbol{\theta}}_t$ and covariance matrix $\mathbf{C}^{-1}$, or from an univariate normal distribution with mean $\tilde{\boldsymbol{\theta}}_{i,t}$ and variance $\mathbf{C}^{-1}_{i,i}$ if sampling is element-wise, as described in Section 2.1.
- Sample $[\sigma^2_{h_\lambda,t}, \sigma^2_{h_p,t}|\text{ELSE}_{t-1}, \boldsymbol{\lambda}^*_t, \mathbf{p}^*_t, \boldsymbol{\theta}_t, \mathbf{y}]$ from the scaled inverse chi-square distributions (2.10) and (2.11), respectively.
- Sample $[\sigma^2_{\varepsilon_\lambda,t}, \sigma^2_{\varepsilon_p,t}|\text{ELSE}_{t-1}, \boldsymbol{\lambda}^*_t, \mathbf{p}^*_t, \boldsymbol{\theta}_t, \sigma^2_{h_\lambda,t}, \sigma^2_{h_p,t}, \mathbf{y}]$ from the scaled inverse chi-square distributions (2.8) and (2.9), respectively.
- Sample $[\mathbf{G}_t|\text{ELSE}_{t-1}, \boldsymbol{\lambda}^*_t, \mathbf{p}^*_t, \boldsymbol{\theta}_t, \sigma^2_{h_\lambda,t}, \sigma^2_{h_p,t}, \sigma^2_{\varepsilon_\lambda,t}, \sigma^2_{\varepsilon_p,t}, \mathbf{y}]$ from the inverse Wishart distribution (2.12).

Above, "ELSE_{t-1}" stands for all parameters, other than those that have been updated in the preceding conditional distribution. 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 Model adequacy

Discrepancy statistic

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, $T_k(\mathbf{y},\boldsymbol{\tau}) = n^{-1}\sum_{i=1}^{n} I(y_i = k)$, where y_i is the i^{th} component of the observed vector $\mathbf{y}$, $T_{k,l}(\mathbf{y}_{rep,l},\boldsymbol{\tau}) = n^{-1}\sum_{i=1}^{n} I(y_{i;rep,l} = k)$, where $y_{i;rep,l}$ is the i^{th} component of the replicated vector $\mathbf{y}_{rep}$ of size n in sample $l = 1, ..., L$, $k = 0, 1, ...$, and $I(.)$ 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, $n^{-1}\sum_{i=1}^{n} I(y_{i;rep,l} = k)$ 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.

Residual analysis

Another tool used for model adequacy was a residual analysis. A residual for the i^{th} record was defined as $r_i = y_i - E(y_i|\tau)$. The posterior mean of the standardized residual was estimated as $\hat{r}_i = J^{-1}\sum_{j=1}^{J} \frac{(y_i - E(y_i|\tau^{(j)}))}{\sqrt{Var(y_i|\tau^{(j)})}}$, $i = 1, ..., n$, 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.

2.4 Model comparison

ZIP models were contrasted against a Poisson model; all specifications included genetic and residual effects in the structure $\boldsymbol{\lambda}^*$. Models were contrasted using the pseudo-Bayes factor (Geiser and Eddy, 1979; Gelfand, 1996). The predictive log-likelihood of each observed datum was estimated as the harmonic mean of the likelihood evaluated at samples from the posterior distribution (Gelfand and Dey, 1994; Newton and Raftery, 1994). At each iteration, the likelihood was stored for each observation, and the harmonic mean of the likelihood across samples was calculated as $\bar{L}_i = \left(J^{-1}\sum_{j=1}^{J}(L_i^{(j)})^{-1}\right)$, where $L_i^{(j)}$ is the likelihood of subject i, evaluated at draw j from the posterior distribution. The predictive likelihood of the model M was them estimated as $\hat{p}(\mathbf{y}|M) = \prod_{i=1}^{n}\bar{L}_i$, and the pseudo-Bayes factor is the ratio between the predictive likelihoods of competing models.

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". When cows have more than one case of mastitis during lactation, 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 the expected under Poisson sampling, so a ZIP model might be suitable. If so, an extension for quantitative genetics analysis of counts with an excess of zeros is needed. The objective of this application was to investigate alternative specifications for modeling number of incidences of mastitis via a ZIP model, and to make inferences about genetic (co)variation between the Poisson and mixture parameters involved.

The hierarchical ZIP model was fitted to a data set consisting of number of mastitis cases in 36,175 first-lactation Norwegian Red cows. The data is described in detail in Rodrigues-Motta *et al.* (2007). The ZIP model in terms of (2.2) 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 ($< 20, ..., 32, > 32$ months), the vector $\boldsymbol{\beta}_{2,j}$ included effects of 12 months at first calving, $\boldsymbol{\beta}_{3,j}$ consisted of effects of 3 years at first calving (1990, 1991 and 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 occurred 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 $\mathbf{H} \otimes I_{n_h}$, where $\mathbf{H} = \begin{pmatrix} \sigma^2_{h_\lambda} & 0 \\ 0 & \sigma^2_{h_p} \end{pmatrix}$ and $n_h = 5,286$. 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 $\mathbf{G} = \begin{pmatrix} \sigma^2_{a_\lambda} & \sigma_{a_{\lambda,p}} \\ \sigma_{a_{\lambda,p}} & \sigma^2_{a_p} \end{pmatrix}$, and $\mathbf{A}$ is a known matrix of additive genetic relationship

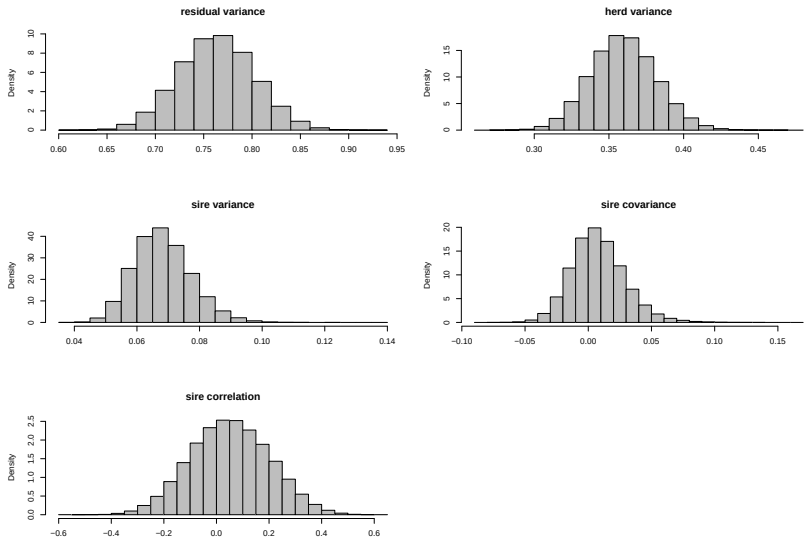

Figure 1: Posterior distributions of the residual, herd, and sire variance affecting $\boldsymbol{\lambda}^*$ and of sire covariances and correlations between $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$.

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-square and inverse Wishart distributions assigned as priors for variance components and matrix $\mathbf{G}$ were $\nu_\varepsilon = \nu_h = \nu_G = 5$. Further, $V_G = \begin{pmatrix} 0.45 & -0.025 \\ & 0.45 \end{pmatrix}$, $\delta_\varepsilon = \delta_h = 1$. Added to high dimension of the data set and the pedigree, the complexity of the hierarchical model demanded that 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

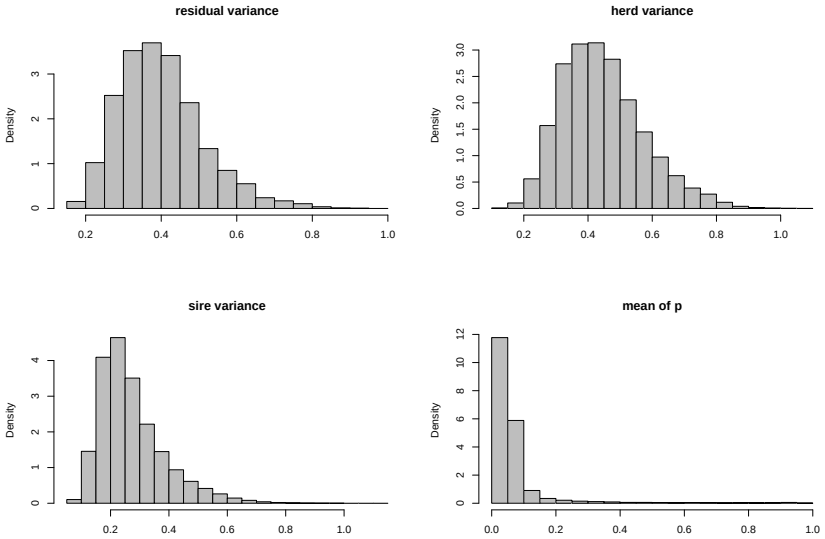

Figure 2: Posterior distribution of the residual, herd, and sire variance affecting $\mathbf{p}^*$ and distribution of the posterior means of the probability of the perfect state ($\mathbf{p}$) under the ZIP model with correlated $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$ parameters.

plots (results not shown) indicated that additional iterations would produce more accurate posteriori 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 were 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, and 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 distribution 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 $\boldsymbol{\lambda}^*$. Additionally, 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 average posterior mean 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 $(0.02; 0.22)$. The average of the posterior means of elements of $\mathbf{p}^*$ lead 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.

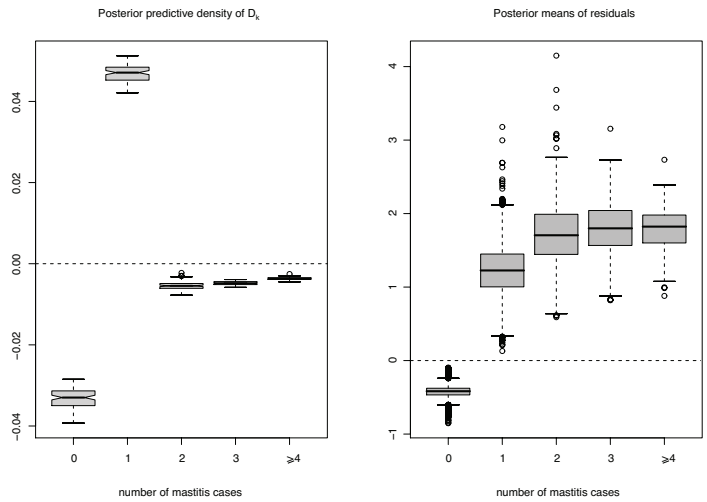

Figure 3: Posterior predictive density of $D_k = T_k(y|.) - T_k(y_{rep}|.)$ and posterior means of the residuals under the ZIP model with correlated $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$ parameters.

For model assessment, the posterior predictive density of $D_k = T_k(y|.) - T_k(y_{rep}|.)$ displayed in Figure 3 (left panel) indicates an agreement between the observed $(T_k(y|.))$ and replicate data $(T_k(y_{rep}|.))$, in special for number of mastitis

cases greater than 1. 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 is less successful for fitting number of mastitis cases larger than 0 because the mean residual values were far from zero. In the analysis of residuals, standardized residuals were obtained by replacing $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$ by point estimates; however, in the complex models considered with a huge number of random effects such residuals are far from being standard normal. In particular, the lack of model adequacy may have been aggravated by a poor estimation of the parameters in the case of small number of observations for number of mastitis cases greater than 0: 15.8, 5.1, and 1.6% cows had 1, 2, and 3 cases, respectively; only 315 cows had more than 3 episodes of clinical mastitis during first lactation.

For comparative purposes, two other models were fitted to the data: (1) a Poisson model having the same exploratory structure for $\boldsymbol{\lambda}^*$ and (2) a ZIP model having the same exploratory structure but uncorrelated $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$. The predictive log-likelihood values for the Poisson, ZIP with correlated $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$ and ZIP with uncorrelated $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$ models were -27059.8, -27019.9 and -27017.9, respectively, favoring the ZIP with uncorrelated $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$ model in a log-scale basis. The pseudo-Bayes factor was 40 between the ZIP model with correlated $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$ and the Poisson model, and 41.9 between the ZIP model with uncorrelated $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$ and the Poisson model, both favoring the ZIP model. The pseudo-Bayes factor between ZIP models with uncorrelated and correlated $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$, respectively, was 2, favoring weakly the ZIP with uncorrelated parameters. Estimates of the variance components affecting $\mathbf{p}^*$ and $\boldsymbol{\lambda}^*$ are summarized in Table 1. The posterior mean (SD) of the residual variance affecting $\boldsymbol{\lambda}^*$ was larger in the Poisson (0.9 (0.03)) than in the ZIP models (posterior means (SD) were 0.72 (0.04) and 0.76 (0.04) in the ZIP model with uncorrelated and correlated $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$, respectively), suggesting that the overdispersion due to zeros was absorbed by the residual term in the Poisson model. The posterior mean (SD) of the herd variance affecting $\boldsymbol{\lambda}^*$ was similar in all models: 0.35(0.02) in the Poisson model and 0.36(0.02) in the two ZIP models. The ZIP models captured more genetic variation affecting $\boldsymbol{\lambda}^*$, since the variation between sires was larger (posterior means (SD) were 0.09 (0.01) and 0.07 (0.01) in the ZIP model with uncorrelated and correlated $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$, respectively) than in the Poisson model (posterior mean (SD) was 0.05 (0.01)). The mean (SD) of the posterior distribution of $\mathbf{p}^*$ in the ZIP model with uncorrelated and correlated parameters were 0.1 (0.11) and 0.07 (0.11), respectively. Estimates were similar in the two ZIP models, except for σ_{a_p} where the mean was 37% larger in the ZIP model with uncorrelated $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$. Since the credible interval for the covariance between sire effects included zero, the principle of parsimony favors the ZIP model with uncorrelated $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$.

Table 1: Posterior mean (standard error) of 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 from Poisson and ZIP models with correlated and uncorrelated $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$.

	Posterior mean (standard error)		
	Poisson	ZIP with uncorrelated $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$	ZIP with correlated $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$
$\sigma^2_{\varepsilon_\lambda}$	0.90 (0.03)	0.72 (0.04)	0.76 (0.04)
$\sigma^2_{h_\lambda}$	0.35 (0.02)	0.36 (0.02)	0.36 (0.02)
$\sigma^2_{a_\lambda}$	0.05 (0.01)	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)

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 a 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 ($\mathbf{p}^*$); the correlated random effects accounted for correlation within and between half-sibs families. The model 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). Additionally, to account for overdispersion over and above that caused by extra zeros, residual terms were included in the structure of $\boldsymbol{\lambda}^*$ and $\mathbf{p}^*$. 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 absorbs overdispersion due to zeros in the residual term reasonable well. If this is so, the Poisson mixed model would produce poor estimates of the variance components. In the ZIP model, the components of variance affecting $\mathbf{p}^*$ were inferred less precisely

than those affecting $\boldsymbol{\lambda}^*$. The small number of observations in each combination of levels of random effects may have been the cause that large residuals were produced in the residual analysis (in special, for counts greater than 0). However, in a predictive analysis random effects were integrate out and the model seems to fit the data.

Although the scale reduction factor value proposed by Gelman and Rubin (1992) as a convergence criterion was satisfied, trace plots (not shown here) suggested that, in the case of variance components affecting $\mathbf{p}^*$, additional iterations are needed for convergence. However, this can be computationally very intensive, and 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. We found a high posterior correlation between the genetic variance affecting $\boldsymbol{\lambda}^*$ and 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 imposing mixing would be via a blockwise Gibbs-Metropolis sampler, as proposed by Gelman *et al.* (2004).

The fully conditional distribution 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.

Acknowledgements

Access to the data was given by the Norwegian Dairy Herd Recording System and the Norwegian Cattle Health Service in agreement number 004.2005. Research was funded by grants CAPES (BRAZIL) BEX 1758004, USDA $2003-35205-12833$, NSF DEB-0089742 and NSF DMS 0443771. Support by the Wisconsin Agriculture Experiment Station and by the Babcock Institute for Dairy Research and Development is acknowledged.

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 Quantitative Genetics*, 3rd. edition. 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*. Chapman and Hall.

Gelfand, A. E. (1996). Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice* (Edited by W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, 145-161). Chapman and Hall.

Gelfand, A. E. and Dey, D. (1994). Bayesian model choice: symptotics and exact calculations. *Journal of the Royal Statistical Society Series B* **56**, 501-514.

Geiser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association* **74**, 153-160.

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

Hall, D. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics* **56**, 1030-1039.

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.

Kuhnert, Petra M., Martin G. T., Mengersen, K. and Possignham, H. P. (2005). Assessing the impacts of grazing levels on bird density in woodland habit: a Bayesian approach using expert information. *Environmetrics* **16**, 717-747.

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

Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical modelling* **5**, 1-19.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society Series B* **56**, 1-48.

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.

Received November 25, 2008; accepted May 6, 2009.

Mariana Rodrigues-Motta
Department of Statistics-IMECC-P.O.Box 6065
University of Campinas-*UNICAMP*, 13083-970, Campinas
SP, Brazil
marianar@ime.unicamp.br

Daniel Gianola
Department of Animal Sciences
University of Wisconsin, 53706
Madison, U.S.A.
gianola@ansci.wisc.edu

Bjørg Heringstad
Department of Animal and Aquacultural Sciences
P.O. Box 5003
Norwegian University of Life Sciences
N-1432, Ås, Norway
bjorg.heringstad@umb.no