Locating multiple interacting quantitative trait loci with the zero-inflated generalized Poisson regression

[1]Vinzenz Erhardt, [2,3]Małgorzata Bogdan, [1]Claudia Czado,

[1] *Technische Universität München, Zentrum Mathematik, Garching, Germany*
[2] *Institute of Mathematics and Computer Science, Wrocław University of Technology, Poland*
[3] *Statistical Bioinformatics Center, Department of Statistics, Purdue University, USA*

Abstract

We consider the problem of locating multiple interacting quantitative trait loci (QTL) influencing traits measured in counts. In many applications the distribution of the count variable has a spike at zero. Zero-inflated generalized Poisson regression (ZIGPR) allows for an additional probability mass at zero and hence an improvement in the detection of significant loci. Classical model selection criteria often overestimate the QTL number. Therefore, modified versions of the Bayesian Information Criterion (mBIC and EBIC) were successfully used for QTL mapping. We apply these criteria based on ZIGPR as well as simpler models. An extensive simulation study shows their good power detecting QTL while controlling the false discovery rate. We illustrate how the inability of the Poisson distribution to account for over-dispersion leads to an overestimation of the QTL number and hence strongly discourages its application for identifying factors influencing count data. The proposed method is used to analyze the mice gallstone data of Lyons, Wittenburg, Li, Walsh, Leonard, Churchill, Carey, and Paigen (2003). Our results suggest the existence of a novel QTL on chromosome 4 interacting with another QTL previously identified on chromosome 5. We provide the corresponding *R* code.

1 Introduction

Despite a long history of QTL mapping (see e.g. Sax (1923)) this research field is still a very active area in which perpetually new statistical methodologies are developed. The majority of methods proposed in the literature, like classical interval mapping (Lander and Botstein (1989) and Haley and Knott (1992)), composite interval mapping (Zeng (1993), Zeng (1994)), multiple QTL mapping (Jansen (1993)

and Jansen and Stam (1994)) or multiple interval mapping (Kao, Zeng, and Teasdale (1999)) are designed for the situation when the trait has a normal distribution. Since in many practical cases this assumption is violated, we observe lately a considerable effort to develop new methods, which could handle other trait distribution types. In this context we mention recent articles on the analysis of ordinal traits (see e.g., Yi, Xu, George, and Allison (2004), Yi, Banerjee, Pomp, and Yandell (2007), Coffman, Doerge, Simonsen, Nichols, and Duarte (2005) or Li, Wang, and Zeng (2006)), nonparametric methods based on ranks (see e.g., Kruglyak and Lander (1995), Zou, Yandell, and Fine (2003) or Żak, Baierl, Bogdan, and Futschik (2007)), extension of multiple interval mapping to generalized linear models (Chen and Liu (2009)) or specific methods which can handle a "spike" in the trait distribution (see e.g., Broman (2003) and Li and Chen (2009)). In case the trait is a count variable it often occurs that it has a "spike" at zero. A clear example of such a phenomenon is provided by the gallstone data of Lyons et al. (2003), where the number of gallstones is considered and a large proportion of mice did not develop any disease symptoms. As illustrated by Cui and Yang (2009), such data can be efficiently modeled using the zero-inflated generalized Poisson regression (ZIGPR, Famoye and Singh (2003)). In contrast to the generalized Poisson regression ZIGPR allows for excess zeros, which may be due to other than genetic reasons. The simulations and the real data analysis reported in Cui and Yang (2009) show that interval mapping based on ZIGPR can efficiently locate QTL influencing the count traits. Cui and Yang (2009) also suggest to apply ZIGPR in order to locate several interacting QTL, based on the multiple interval mapping approach.

From the statistical point of view the most difficult part in fitting the multiple regression model lies in the estimation of the number of significant predictors. As discussed in Broman and Speed (2002) and Bogdan, Ghosh, and Doerge (2004), the classical model selection criteria have a strong tendency to overestimate the number of QTL when the number of markers is comparable to the sample size n. These experimental observations were confirmed by theoretical results in Bogdan, Ghosh, and Żak-Szatkowska (2008c) and Chen and Chen (2008), which show that the classical Bayesian Information Criterion (BIC, Schwarz 1978) is not consistent when the number of potential regressors increases to infinity quicker then $\sqrt{n}$. To correct for this behavior of BIC, several modifications of this criterion were proposed in the literature (e.g. see Ball (2001), Bogdan et al. (2004), Manichaikul, Moon, Sen, Yandell, and Broman (2009)). Specifically, Bogdan et al. (2004) propose to modify BIC by supplementing it with the Binomial prior distribution on the QTL number. If the expected value of this prior distribution does not depend on the number of markers, this leads to an additional "penalty" for the model dimension, which prevents overestimation. As illustrated by theoretical results in Bogdan et al. (2008c), mBIC controls the number of falsely detected QTL and has some asymp-

totic optimality properties in the context of selecting the best multiple regression model under sparsity. Recently, another interesting extension of BIC, EBIC, was proposed by Chen and Chen (2008). In its standard form (e.g., see Li and Chen (2009)) EBIC uses a non informative uniform prior on the number of QTL. Chen and Chen (2008) support EBIC by showing its consistency.

In a sequence of papers Baierl, Bogdan, Frommlet, and Futschik (2006), Baierl, Futschik, Bogdan, and Biecek (2007), Żak et al. (2007) and Bogdan, Frommlet, Biecek, Cheng, Ghosh, and Doerge (2008b) mBIC was successfully used to locate multiple interacting QTL. Specifically, Żak et al. (2007) proposed a nonparametric version of mBIC based on ranks, which can be used to analyze traits which do not have a normal distribution. However, the rank methods are only well justified if the trait has a continuous distribution. Therefore they have to be used with care when the trait has a "spiked" distribution, i.e. when some proportion of the trait data are concentrated at one point. Recently, a very interesting application of EBIC to the traits with "spiked" distributions was proposed in Li and Chen (2009). Li and Chen (2009) use the approach of Broman (2003) and model such traits with a mixture of a distribution concentrated at one point and a distribution from the general exponential family. They show that an appropriately modified BIC can be used successfully to locate QTL influencing such traits. Here we extend this approach and apply mBIC and EBIC for locating multiple interacting QTL based on the zero-inflated generalized Poisson regression. Note that this application goes beyond the framework of Li and Chen (2009), since the generalized Poisson distribution does not belong to the exponential family.

We illustrate the performance of mBIC and EBIC to a ZIGPR with an extensive simulation study. The results of this study show that the proposed methods allow for a good power of QTL detection, while keeping the false discovery rate at a reasonable level. They also clearly illustrate the superior performance of ZIGPR over other simplified methods analyzing count traits. Here, among other findings, we present the interesting phenomenon of overestimating the number of QTL by the standard Poisson regression. This behavior can be attributed to the inability of the Poisson regression to account for data over-dispersion and therefore it should not be applied for identifying QTL's based on count data. We also report results of the analysis of the mice gallstone data of Lyons et al. (2003), which confirms the good performance of mBIC applied to ZIGPR. Specifically, our method confirms the existence of a QTL on a chromosome 5, influencing the number of gallstones, and additionally suggests a novel QTL on chromosome 4. The program in R, which can be used for future real data analyses, is available at *http://www-m4.ma.tum.de/Papers/Erhardt/qtl-zigp-code.rar*.

The outline of the paper is as follows. In Section 2 we introduce and discuss our ZIGPR model for QTL mapping. In Section 3 we introduce the correspond-

ing versions of mBIC and EBIC. In Section 4 we present results of the extensive simulation study comparing ZIGPR to simpler versions of Poisson regression as well as with a standard least squares regression with regard to the performance of mBIC and EBIC. Section 5 contains the results of the analysis of mice gallstone data of Lyons et al. (2003) and Section 6 contains a summary as well as directions for further research.

2 Zero-inflated generalized Poisson regression

One of the simplest distributions which can be used to model count traits is the Poisson distribution. However, the range of applications of this distribution is very limited due to the lack of its flexibility. Specifically, the standard Poisson model assumes that the trait variance is equal to its mean. As discussed later in this paper, this weakness becomes particularly disturbing when the Poisson distribution is used together with model selection tools for locating multiple interacting QTL.

There are two natural extensions of the Poisson distribution, which allow for modeling a difference between the mean and the variance: the Negative Binomial (or Poisson-Gamma) distribution and the generalized Poisson distribution. In this paper we will use the generalized Poisson distribution $GP(\mu,\varphi)$, which was first introduced by Consul and Jain (1970) and subsequently studied in detail by Consul (1989). In the context of QTL mapping GP was applied e.g. by Thomson (2003). In this article we refer to the mean parametrization of GP(see e.g. Consul and Famoye (1992)):

$$\text{for } \; y \in \{0,1,\ldots\} \quad P(Y=y|\,\mu,\varphi) = \frac{\mu(\mu+(\varphi-1)y)^{y-1}}{y!}\varphi^{-y}e^{-\frac{1}{\varphi}(\mu+(\varphi-1)y)}\,, \tag{2.1}$$

where μ and φ are larger than 0. For $Y \sim GP(\mu,\varphi)$ we have $E(Y)=\mu$ and $Var(Y)=\varphi^2\mu$. This allows for modeling over- or underdispersion. However, in the case of underdispersion ($\varphi \in (0,1)$), the support of the distribution depends on μ and φ, which is difficult to enforce when μ and φ need to be estimated. Therefore, in this article we restrict to equi- and overdispersion; $\varphi \geq 1$.

When comparing to the Negative Binomial (NB) distribution, the GP distribution has several advantages. While the NB distribution with pmf

$$P(Y=y|\mu,\Psi) = \frac{\Gamma(y+\Psi)}{\Gamma(\Psi)y!}\left(\frac{\Psi}{\mu+\Psi}\right)^{\Psi}\left(\frac{\mu}{\mu+\Psi}\right)^{y},$$

and $E(Y)=\mu$, $Var(Y)=\mu(1+\frac{\mu}{\Psi})$ contains the basic Poisson distribution only as a limiting case for $\Psi\to\infty$, the GP distribution contains the Poisson class for $\varphi=1$.

Second, unlike the NB distribution the dispersion factor in GP is independent of the mean. Hence, in the NB distribution the statistical modeling of overdispersion is less transparent than in case of the GP. For a detailed comparison between GP and NB we refer the readers to Joe and Zhu (2005).

A zero-inflated generalized Poisson (ZIGP) distribution is a further extension of the GP distribution, which allows to model a "spike" at zero. Such a "spike" occurs quite often when the response variable counts disease symptoms (like e.g. the gallstones). In the context of QTL mapping, the ZIGP distribution was first applied by Cui, Kim, and Zhu (2006). As explained by Cui and Yang (2009), the over-excess of zeros may result from the fact that a certain fraction of a population was not exposed to the disease virus.

The ZIGP distribution is defined as a mixture of a distribution concentrated at 0, denoted as δ_0, and the generalized Poisson distribution:

$$ZIGP(\mu,\varphi,\omega) = \omega\delta_0 + (1-\omega)GP(\mu,\varphi) \ , \tag{2.2}$$

where $\omega \in [0,1]$ is the zero-inflation parameter. Mean and variance of $Y \sim ZIGP$ are given by

$$E(Y) = (1-\omega)\mu \quad \text{and} \quad \sigma^2 := Var(Y) = E(Y)\left(\varphi^2 + \mu\omega\right). \tag{2.3}$$

To model the dependence of the count response variable on explanatory variables Famoye and Singh (2006) introduced a zero-inflated generalized Poisson regression model for independent $Y_i \sim ZIGP(\mu_i,\varphi,\omega_i)$, where μ_i and ω_i are defined through the log-linear and logit link functions, respectively. In this article we will restrict to the case when the zero-inflation parameter ω does not depend on genetic factors, while the dependency of μ_i on explanatory variables is given through the log-linear link function

$$\log\mu_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ji} \ .$$

The constant ω can be interpreted as the fraction of the population which was not exposed to the disease virus.

The class of ZIGPR models, considered in this paper contains the subclasses of zero-inflated Poisson regression (ZIPR, $\varphi = 1$), generalized Poisson regression (GPR, $\omega = 0$) and standard Poisson regression (PoiR, $\varphi = 1$, $\omega = 0$).

Remark 1. In our preliminary research we also considered the situation when both ω and μ were influenced by genetic covariates. However, we observed that due to the fact that both μ and ω directly influence the expected trait value and the probability that the trait is equal to zero, a precise separation of regressors influencing

these two parameters was hardly possible with the sample sizes typically used for QTL mapping. Therefore, the extension of our model to include the dependency of ω on the genetic factors did not bring the expected benefits over the restricted version. We believe that our choice of a constant ω is justified in many situations, like e.g. in the case where it is interpreted as the probability of not having a contact with the disease virus. An alternative ZIGPR model for QTL mapping was proposed in Cui and Yang (2009). This model, based on the parametrization of Lambert (1992), assumes that both logit(ω) and $\log \mu$ are proportionally influenced by the same genetic covariates and explicitly "confounds" μ and ω. The model selection methods proposed in this article can be used also for the Cui and Yang (2009) parameterization.

3 mBIC and EBIC for ZIGPR

Consider the problem of locating multiple interacting QTL in experimental populations. In this case precise estimators of QTL positions and their effects can be obtained with the multiple interval mapping, MIM (see e.g. Kao et al. (1999)), which for a variety of different trait distributions has been implemented in the popularly used packages *QTL Cartographer* and *R/qtlbim*. The application of MIM for ZIGPR is quite straightforward and has been recently discussed in Li and Chen (2009). However, according to our knowledge, the implementation of MIM for ZIGPR is not available yet.

The general idea of MIM is to fit the corresponding regression model at a large number of possible QTL positions and estimate QTL locations by maximizing the corresponding likelihood function. If the QTL is located between the markers, the trait distribution is modeled as a mixture of distributions corresponding to the possible QTL genotypes. The mixture coefficients are defined by the conditional probabilities of QTL genotypes, given the genotypes of flanking markers. The parameters of the linear model are usually estimated by the EM algorithm or by replacing the unknown QTL genotypes with the expected values of the corresponding dummy variables, conditional on the genotypes of flanking markers (for a comparison of these two approaches in the context of the least-squares regression see e.g. Kao (2000)). While MIM should be recommended for the precise QTL analysis, it creates a huge computational burden when it needs to be repeated many times in large scale simulation studies. On the other hand, simulation results of Dupuis and Siegmund (1999) and Bogdan et al. (2008b) show that multiple interval mapping does not substantially increase the power of QTL detection in comparison to the search over marker positions. Therefore, to reduce the computational complexity, interesting genome regions can be initially chosen by selecting the best regression

model (possibly with interactions), relating the trait values to the marker genotypes. Since the main purpose of the present article is the comparison of different Poisson regression models with respect to the power of QTL detection, we restrict the attention to such a search over markers.

In case of a backcross design or recombinant inbred lines there are only two genotypes possible at every locus and each of the markers may be represented by just one dummy variable: $X_{ij} = \frac{1}{2}$ or $X_{ij} = -\frac{1}{2}$, depending on the number of alleles from the reference parental line present at marker j for the i^{th} individual. In case of an intercross design there are three possible genotypes and, according to the Cockerham's model (see Kao and Zeng (2002)), each of the markers can be represented by two dummy variables:

$$\text{Additive Effect for individual } i\text{:} \quad X_{aij} = \begin{cases} 1 & \text{if the } j^{th} \text{ marker has a genotype } g_{ij} = AA, \\ 0 & \text{if the } j^{th} \text{ marker has a genotype } g_{ij} = aA, \\ -1 & \text{if the } j^{th} \text{ marker has a genotype } g_{ij} = aa. \end{cases}$$

$$\text{Dominance Effect for individual } i\text{:} \quad X_{dij} = \begin{cases} 1/2 & \text{if } j^{th} \text{ marker has a genotype } g_{ij} = Aa, \\ -1/2 & \text{otherwise .} \end{cases}$$

Let $Y = (Y_1, Y_2, \ldots, Y_n)^T$ denote the vector of values of some quantitative trait for n individuals and let $X_{n \times N_m}$ denote the corresponding design matrix, whose columns contain dummy variables corresponding to all available markers. Note that for the backcross and recombinant inbred lines $N_m = m$, where m is the number of available markers, while for the intercross $N_m = 2m$.

We assume that the relationship between QTL genotypes (coded as above) and the count trait can be described by a zero-inflated generalized Poisson regression model. As already discussed, we will focus on identification of markers which are closest to the QTL. In our search, apart from main effects (additive and dominance), we may include two-way interactions (epistatic effects). Thus our task consists in choosing the best model of the form $Y_i \sim ZIGP(\mu_i, \varphi, \omega)$, with

$$\log(\mu_i) = \beta_0 + \sum_{j \in I} \beta_j X_{ij} + \sum_{(u,v) \in U} \gamma_{uv} X_{iu} X_{iv}, \tag{3.1}$$

where I is a subset of the set of indices $N = \{1, \ldots, N_m\}$ of all dummy variables coding QTL genotypes and U is a subset of $N \times N$. Note that the total number of potential two-way interactions is equal to $N_e = N_m(N_m - 1)/2$.

Remark 2. Our model allows to include interaction effects without the corresponding main effects. This modeling strategy is motivated by the well documented find-

ings of genes which do not have main effects and influence the trait only by interactions with other genes (see e.g., Fijneman, De Vries, Jansen and Demant (1996) and the real data analysis in the present paper). In principle, the model (3.1) could be extended to include also interactions of higher order. However, due to the increased multiple testing problem, the power for identification of such interactions is very limited for sample sizes typically used in QTL mapping. Therefore, genome-wide searches for high-order interactions are rarely carried out.

Remark 3. In case of an intercross design there are four terms in the linear model (3.1) which describe the interaction between the j-th and k-th marker: additive-additive term $X_{aij}X_{aik}$, additive-dominance term $X_{aij}X_{dik}$, dominance-additive term $X_{dij}X_{aik}$ and dominance-dominance term $X_{dij}X_{dik}$. In our approach we separately add these terms to the model. Compared to the approach where all these terms are included together, our method allows to reduce the penalty (or the number of degrees of freedom) for the interaction and allows for a larger power of detecting epistasis, when only one or two of the interaction components are substantially different from zero.

Since we do not know the QTL number nor their locations, we use a model selection procedure for choosing the best regressors in model (3.1). One popular method for this purpose is the Schwarz Bayesian Information Criterion (BIC). However, when locating QTL with the standard least-squares regression, BIC was found to have a strong tendency to overestimate the QTL number (see e.g. Broman and Speed (2002)). As discussed in Bogdan et al. (2008c), this phenomenon is closely related to the well known multiple testing problem. Specifically, in Bogdan et al. (2008c) it is proved that under the orthogonal design the expected number of "false discoveries" produced by BIC converges to infinity if $\frac{N_m}{\sqrt{n}} \to \infty$. In Bogdan et al. (2004) an alternative Bayesian explanation is provided. The Bayesian model selection suggests choosing the model M_j that has the highest posterior probability

$$P(M_j|Y) \propto L(Y|M_j)\pi(M_j) \ ,$$

where $L(Y|M_j)$ is the likelihood of the data given the model M_j and $\pi(M_j)$ is a prior probability of M_j. The standard BIC neglects $\pi(M_j)$ and uses the Laplace approximation for $\log L(Y|M_j)$ (e.g. see Ghosh, Delampady and Samanta (2006)), which results in

$$BIC = \log(L(Y|M_j, \hat{\delta}_j)) - \frac{1}{2}k_j \log(n) \ ,$$

where $\hat{\delta}_j$ is the maximum likelihood estimate of the parameter vector in model M_j and k_j denotes the dimension of δ_j.

In Bogdan et al. (2004) it is observed that neglecting $\pi(M_j)$ corresponds to assigning the same prior probability to each model. It is easy to check that this leads to the implicit Binomial $B\left(N_m, \frac{1}{2}\right)$ prior on the number of main effects. This prior is concentrated mainly on the interval $\left(\frac{N_m - 3\sqrt{N_m}}{2}, \frac{N_m + 3\sqrt{N_m}}{2}\right)$ and assigns an unsuitably large prior probability to the event that the true number of QTL is close to $\frac{N_m}{2}$. This in turn causes the BIC to choose relatively large models. To solve this problem, in Bogdan et al. (2004) a modified version of the BIC, called mBIC, has been proposed. The mBIC criterion allows to take prior information on the number of QTL into account. Let $E(k)$ and $E(r)$ denote the expected values of the prior distributions for the number of main and epistatic effects, respectively. In mBIC the parameter $p = \frac{1}{2}$ in the Binomial prior distribution for the number of true regressors is replaced with $p_a = \frac{E(k)}{N_m}$ for the main effects and $p_e = \frac{E(r)}{N_e}$ for the interactions.

After some simple algebra (for details see e.g., Bogdan et al. (2004) or Żak-Szatkowska and Bogdan (2010)), we obtain that mBIC selects the model which maximizes the expression

$$mBIC := 2\log(L(Y|M_j, \hat{\delta}_j))) - (k_j + r_j)\log(n) - 2k_j\log(l-1) - 2r_j\log(u-1) \ , \tag{3.2}$$

where k_j and r_j are the numbers of main and interaction effects in the model M_j, $l = \frac{1}{p_a}$ and $u = \frac{1}{p_e}$. In the case of no prior information, Bogdan et al. (2008c) suggest using

$$l = \frac{N_m}{4} \ , \tag{3.3}$$

when the scan is restricted to main effects only and

$$l = \frac{N_m}{2.2} \quad \text{and} \quad u = \frac{N_e}{2.2} \ , \tag{3.4}$$

when epistatic effects are considered as well.

In comparison to BIC, the standard version of mBIC for detecting main effects and two-ways interactions contains the additional penalty term

$$2k_j\log\left(\frac{N_m}{2.2} - 1\right) + 2r_j\log\left(\frac{N_e}{2.2} - 1\right) \ ,$$

which depends on the number of markers used in the genome scan. As shown in Bogdan et al. (2008c), in case of the standard least squares regression this additional term allows to deal with the multiple testing problem and guarantees that the overall type I error does not exceed 0.08 for a sample size of 200 and more than 30 markers. Due to the consistency of mBIC, the probability of the type I error decreases when the sample size increases.

Choosing the same penalty constant (namely 2.2) for main and interaction effects results in dividing the probability of the overall type I error in two approximately equal parts: the probability of detecting a "false" additive effect and the probability of detecting a "false" interaction. Thus, the expected number of falsely detected interactions is approximately equal to the number of falsely detected main effects. Note that since $N_e >> N_m$, this choice implies a larger penalty for interaction terms than for main effects. As a result the power of detecting interaction effects by the standard version of mBIC is substantially smaller than the power of detecting main effects of the same "size". This choice is a deliberate decision. Since $N_e >> N_m$, equating the penalty coefficients for main and interaction effects in such a way that the probability of the overall type I error is still controlled would lead to a decrease of the power for main effects, without having much effect on the power for interactions. Our proposed approach can easily be extended to higher order interactions, even without a significant sacrifice of power of detecting lower order terms. However, other choices are also possible and the penalty coefficients can easily be adjusted according to prior knowledge and preferences of the researcher.

Remark 4. As discussed in Bogdan et al. (2008b), the specific choice of the penalty coefficients for mBIC is related to the Bonferroni correction for multiple testing. This correction works well when the corresponding test statistics are independent and is usually quite conservative when they are strongly correlated. Therefore, the calibration of mBIC is particularly suitable for sparse marker maps. Despite these concerns, the simulation study reported in Bogdan et al. (2004) shows that in the case of a backcross design, where the correlation between marker genotypes is particularly strong, mBIC works well if the average distance between markers is larger or equal than $5cM$. The performance of mBIC for dense marker maps and multiple interval mapping is investigated in Bogdan et al. (2008b), where a method for scaling the corresponding penalty coefficients is proposed. The results of Bogdan et al. (2008b) suggest that if the average distance d between markers is smaller than $5cM$, the penalty "weight" of each additive and interaction term should roughly be proportional to d. Thus, if markers are very densely spaced, the corresponding penalty for the additive effects depends on the length of the chromosome rather than the number of markers. In case of interactions, the penalty coefficient still depends on the number of markers, but this dependence is substantially weaker than for a sparse map.

The calculations presented in Bogdan et al. (2004), Bogdan et al. (2008b) and Bogdan et al. (2008c), which lead to the specific choices of l and u, are based on the assumption that the likelihood ratio statistics for testing the significance of specific explanatory variables have asymptotically the chi-square distribution. Since,

under some mild regularity conditions, this assumption is satisfied for the Generalized Linear Models (see e.g. Shao (1999)), the proposed choices for l and u are appropriate also in this case. An extensive simulation study, confirming good properties of the mBIC in the context of logistic and Poisson regression, can be found in Żak-Szatkowska and Bogdan (2010). Note however that ZIGPR does not fit the general framework of GLM, since the ZIGP distribution does not belong to the exponential family. In this case a standard choice of l and u can be justified by theoretical results on the asymptotic normality of the maximum likelihood estimate of $\delta = ((\beta_j)_{j\in I}, (\gamma_{uv})_{(u,v)\in U}, \varphi, \omega)$, presented in Czado, Erhardt, Min, and Wagner (2007) based on Min and Czado (2010). This implies that under the null hypothesis the corresponding likelihood ratio test statistics have also asymptotically a chi-square distribution. The appropriateness of the standard choice of l and u for ZIGPR is confirmed by the simulation study, presented in the next section.

A similar modification of BIC was recently proposed by Chen and Chen (2008), who introduce an extended BIC (EBIC), based on the different prior choices for the model dimension. In comparison to mBIC, the priors used by EBIC substantially prefer models of larger dimensions. Specifically, the standard, most restrictive version of the EBIC, assumes that the prior distribution on the number of main effects is uniform on the set $\{0, 1, \ldots, N_m\}$ (see Li and Chen (2009)). After assigning the same prior probability to all models of the same dimension this results in $\pi(M_j) = \frac{1}{N_m+1}\binom{N_m}{k}^{-1}$. Interestingly, the same prior is proposed in Scott and Berger (2008), where it results from the application of a hierarchical model with a non informative, uniform prior on the proportion of true regressors p. The choice between mBIC and EBIC should depend on the prior expectations concerning the QTL number. As illustrated by theoretical results discussed in Bogdan et al. (2008b) and proved in Bogdan, Chakrabarti, and J.K.Ghosh (2008a), mBIC has some asymptotic optimality properties in the context of selecting the best multiple regression model under sparsity. Therefore mBIC seems to be especially appropriate in case when one expects that the number of true predictors is much smaller than the number of columns in the "total" design matrix. To compare these two criteria, in the next section we present results of an extensive simulation study, in which we identify important main effects with the standard version of mBIC

$$mBIC := 2\log(L(Y|M_j, \hat{\delta}_j))) - k\log(n) - 2k\log\left(\frac{N_m}{4} - 1\right) \quad , \tag{3.5}$$

and the standard version of EBIC

$$EBIC := 2\log(L(Y|M_j, \hat{\delta}_j))) - k\log(n) - 2\log\binom{N_m}{k} \quad . \tag{3.6}$$

4 Simulation study

Simulations are carried out to investigate the performance of our proposed methods of QTL detection for a backcross design. We simulate genotypes of $N_m = 100$ markers located on 20 mice chromosomes. These marker positions are identical to the ones in the data set investigated by Lyons et al. (2003). The marker positions are supplemented by $k = 10$ fictional QTL's (not matching any of the markers) located on chromosomes 1 to 6. Figure 1 plots the marker and QTL positions on these 6 chromosomes.

Trait values are generated from the ZIGPR model, $Y_i \sim ZIGP(\mu_i(\beta), \varphi, \omega))$, with

$$\mu_i(\beta) := \exp\left\{2.05 + X'_{Q,i}\beta\right\}, \quad i = 1, \ldots, n \ , \tag{4.1}$$

where $X_{Q,i} = (X_{Q1,i}, \ldots, X_{Q10,i})'$ denotes the vector of 10 QTL genotypes coded as $-1/2$ and $1/2$ for homozygotes and heterozygotes, respectively, and parameter values are chosen as

$$\beta = (-0.20, 1.00, 0.25, -0.60, 0.80, 1.20, 0.70, -0.15, -0.40, 1.50)' \ .$$

Additionally, we choose $\varphi = 2$ and investigate small as well as medium sized zero-inflation of $\omega \in \{20\%, 40\%\}$.

Our simulation results are based on $N = 1000$ replicates for the sample sizes $n = 200$ and $n = 500$. In each run new random markers and QTL genotypes are generated from the map, the coefficients β , however, are kept identical. In order to handle the computational complexity of a large scale simulation study, in each of these replicates model selection is carried out using a forward selection procedure. We start with the Null model, i.e. we fit $ZIGPR(\mu_i(\beta_0), \varphi, \omega))$, where β_0 is the coefficient of an intercept. We add sequentially the marker which increases the standard version of the mBIC (3.5) the most, as long as the mBIC grows. Since our simulated QTL are widely spaced, we expect the model selected by forward selection to be identical or close to the optimal model with respect to mBIC in most of the replicates (see e.g., Broman (1997)). Additionally, we carry out forward selection based on mBIC with a Gaussian linear model (LM), a Poisson regression (PoiR), generalized Poisson regression (GPR) and zero-inflated Poisson regression (ZIPR). We include the standard least squares regression LM,

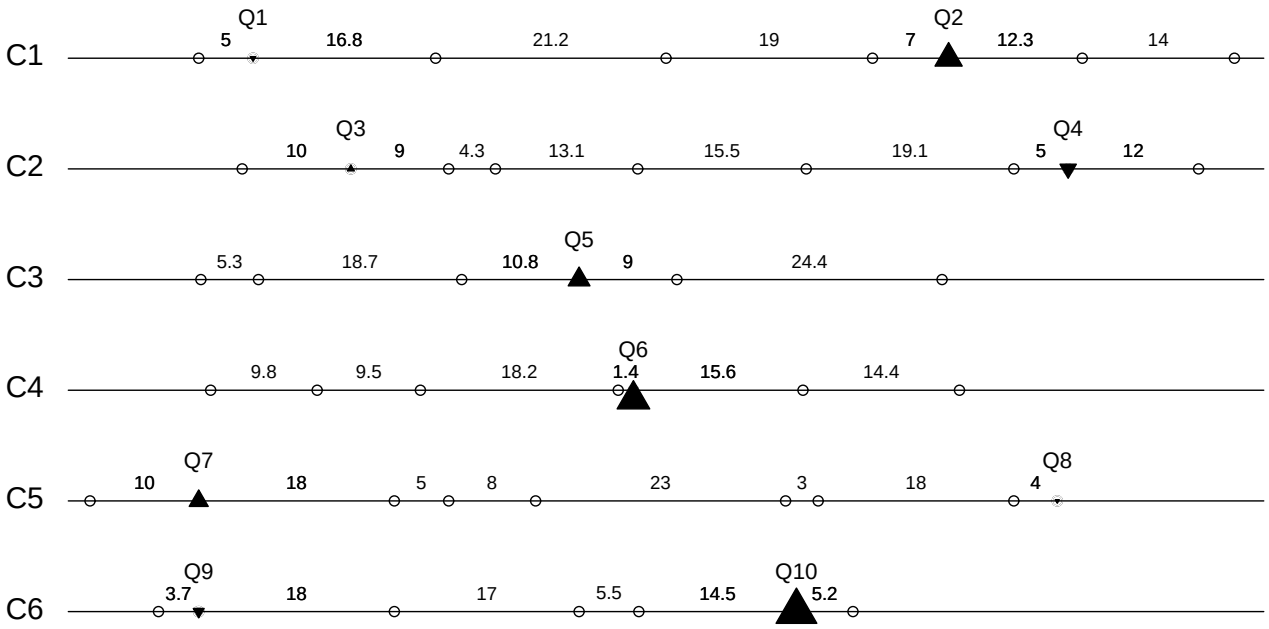

Figure 1: Marker positions and positions of the true QTL, where positive effects are denoted by point-up triangles, negative ones by point-down triangles. The sizes of the triangles are proportional to the magnitude of the coefficient.

since it is capable of identifying correlations between explanatory and response variables, and may perform reasonably choosing important predictors for the ZIGP data. Also, due to the central limit theorem, we expect that mBIC with LM will control the number of false positives, even when the true data are generated according to ZIGPR. Additionally, for each model class we perform model selection based on the standard version of EBIC given in (3.6).

Results of the simulation study are compared for the five model classes. We consider the following statistics:

- true positives (TP): number of selected effects whose distance to the simulated QTL's was less or equal 20 *cM*; if more than one effect was caught in the interval around a certain QTL only one of them was counted
- false positives (FP): number of selected effects whose distance to the simulated QTL's was higher than 20 *cM*
- misclassification error, ME = false positives (FP) + false negatives (FN), where $FN = 10 - TP$
- power: $TP/10$
- observed false discovery rate : $FDR = FP/(FP+TP)$

In Table 1 we will tabulate the averages of FP, ME, power and FDR. Figure 2 plots the estimated power against the magnitude of the true regression coefficients β.

From Table 1 and Figure 2 we see that a higher number of observations substantially eases the detection of significant effects. On the other hand, higher zero-inflation makes the detection of correct effects more difficult even in the correctly specified ZIGPR model. Also, according to Figure 2, the power of detection clearly increases with the magnitude of the true regression coefficients.

Our simulations show that mBIC and EBIC based on the ZIGP model provide a low false discovery rate while maintaining relatively high power rates. These criteria are also definitely the best with respect to the misclassification error ME. Note that if the cost of the false positive is the same as the cost of the false negative, ME is proportional to the cost of the statistical inference. Interestingly, the second best group of procedures with respect to ME is formed by the criteria based on the standard least squares regression model, LM. While LM clearly performs worse than the correct ZIGPR model, it outperforms other misspecified models based on Poisson regression. Specifically, the LM class offers a much larger power than the corresponding Generalized Poisson Regression (GPR) class without zero-inflation parameter. In case of the models without the overdispersion parameter, PoiR and ZIPR, we observe the opposite, i.e., the corresponding criteria offer a much higher

	n = 200									
	mBIC									
	$\omega = 20\%$					$\omega = 40\%$				
	LM	PoiR	ZIPR	GPR	ZIGPR	LM	PoiR	ZIPR	GPR	ZIGPR
FP	0.125	21.075	11.309	0.658	0.357	0.106	27.033	9.614	0.296	0.373
ME	6.944	22.903	13.486	7.952	5.371	8.260	29.025	12.460	9.631	6.623
Power	0.318	0.817	0.782	0.271	0.499	0.185	0.801	0.715	0.066	0.375
FDR	0.036	0.711	0.575	0.188	0.062	0.050	0.764	0.558	0.282	0.088
	EBIC									
	$\omega = 20\%$					$\omega = 40\%$				
FP	0.149	40.869	19.746	0.880	0.604	0.090	53.568	15.354	0.281	0.614
ME	6.879	41.808	21.377	7.943	5.196	8.368	54.276	17.671	9.633	6.397
Power	0.327	0.906	0.837	0.294	0.541	0.172	0.929	0.768	0.065	0.422
FDR	0.040	0.809	0.682	0.213	0.090	0.044	0.846	0.645	0.270	0.116
	n = 500									
	mBIC									
	$\omega = 20\%$					$\omega = 40\%$				
	LM	PoiR	ZIPR	GPR	ZIGPR	LM	PoiR	ZIPR	GPR	ZIGPR
FP	0.112	24.662	14.818	0.642	0.215	0.117	30.817	13.813	0.405	0.234
ME	4.830	25.519	15.723	6.102	3.541	5.949	31.847	15.088	8.381	4.047
Power	0.528	0.914	0.909	0.454	0.667	0.417	0.897	0.873	0.202	0.619
FDR	0.019	0.723	0.607	0.112	0.028	0.025	0.770	0.599	0.142	0.033
	EBIC									
	$\omega = 20\%$					$\omega = 40\%$				
FP	0.144	40.428	26.274	0.984	0.466	0.120	48.520	23.765	0.465	0.435
ME	4.662	40.936	26.878	6.145	3.397	5.815	49.110	24.665	8.490	3.940
Power	0.548	0.949	0.940	0.484	0.707	0.430	0.941	0.910	0.198	0.649
FDR	0.023	0.805	0.725	0.149	0.055	0.024	0.835	0.708	0.154	0.057

Table 1: Average number of false positives (FP), misclassification error (ME), power and false discovery rate (FDR) based on mBIC and EBIC for different model classes and $n = 200, 500$ and $\omega = 20\%, 40\%$

power than the criteria based on LM, or even ZIGPR, but instead lead to the detection of a large number of false positives. The FDR of PoiR and ZIPR systematically exceeds 50%, which implies that the number of false positives usually exceeds the number of true discoveries. We found this phenomenon very interesting, since according to our theoretical results and simulation studies reported in Żak-Szatkowska and Bogdan (2010), the mBIC with PoiR performs very well with respect to FDR and ME if the data are generated exactly according to the Poisson regression. Also, under the total null hypothesis, mBIC with PoiR controls the overall type I error at the assumed level. We believe that when the data are generated by ZIGPR the criteria based on PoiR and ZIPR pick too many regressors in order to account for the data heterogeneity caused by overdispersion.

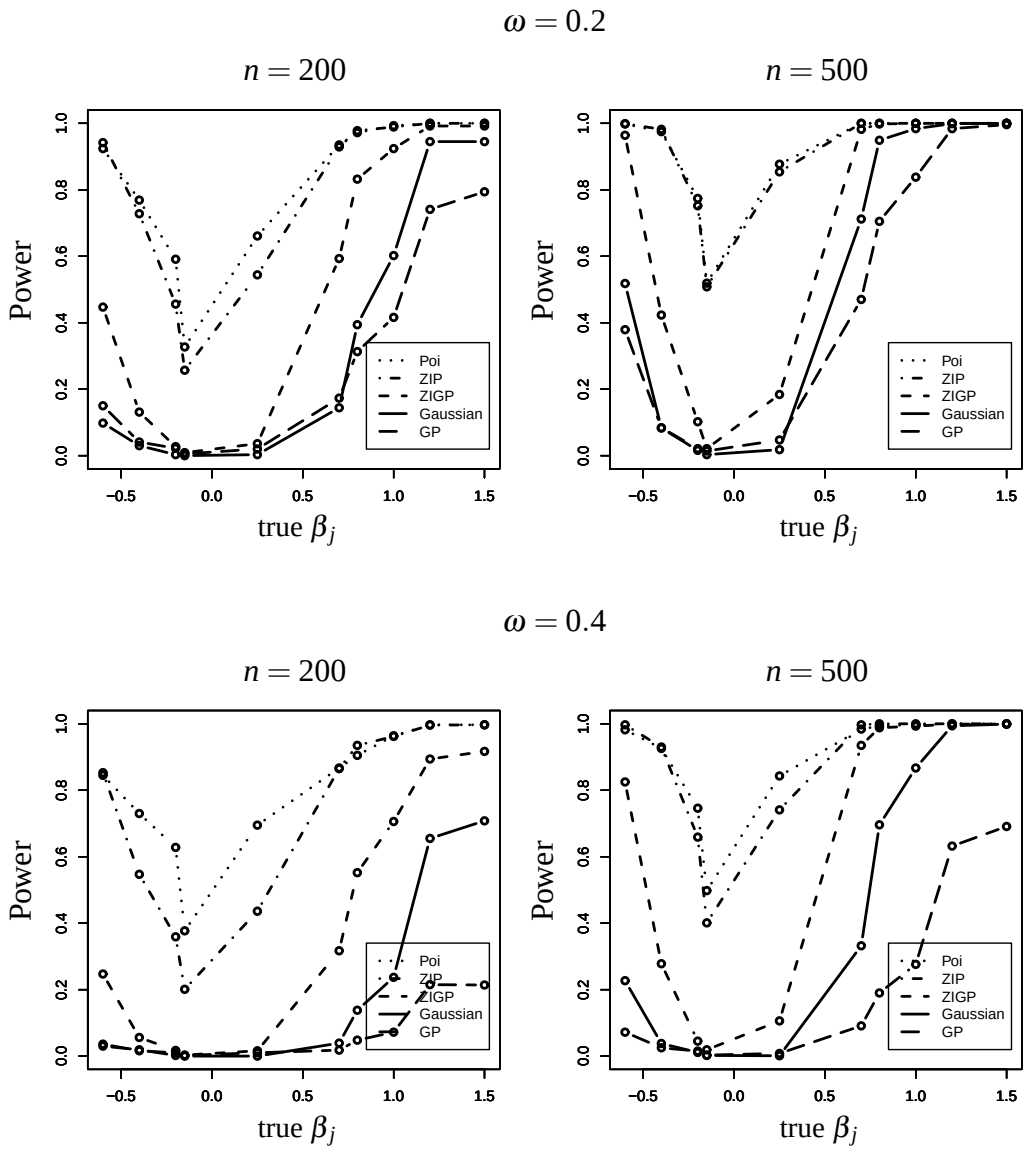

Figure 2: Power for different sizes of true regression coefficients based on several model classes. Note that the lines are linearly interpolated to increase visual comparability.

	n = **200**, mBIC				
	LM	PoiR	ZIPR	GPR	ZIGPR
FP	0.095	8.200	8.150	0.405	0.410
ME	5.285	9.830	9.810	3.920	3.930
Power	0.481	0.837	0.834	0.648	0.648
FDR	0.018	0.476	0.475	0.053	0.053

Table 2: Average number of false positives (FP), misclassification error (ME), power and false discovery rate (FDR) based on mBIC for different model classes when the traits come from a Poisson distribution

In Table 2 we report the results of a further simulation study, in which the data were generated according to the standard Poisson regression model PoiR, with μ_i defined by (4.1). Since this is only meant to be an illustrative example, we restrict to the case of a scan based on mBIC for $n = 200$ mice. In this case PoiR and ZIPR perform similarly bad, while GPR and ZIGPR are similarly good. The reason is simply that in the model classes allowing for excess zeros, the zero-inflation parameter for Poisson traits is estimated to be close to zero, hence the performance only depends on the underlying distribution, which is not inflated (i.e. Poisson and GP, respectively). Interestingly, also in this case mBIC with PoiR and ZIPR substantially overestimates the number of QTL. The number of false positives produced by these criteria is approximately equal to the number of true discoveries, with FDR close to 50%. At the same time mBIC based on GPR and ZIGPR work very well, maintaining a reasonable power and FDR at the level close to 5%. It turns out that the poor behavior of mBIC based on PoiR or ZIPR results from the model misspecification, caused by the discrepancy between the marker and QTL location. Here we give a simple illustrative example: we generate Poisson traits with 10 true effects $X_i := (X_{i1}, \ldots, X_{i10})'$ and $\mu_i := \exp(2.05 + X_i'\beta)$, where β is chosen as before. Then we fit two GPR models, one using X_i as regressors and one using misspecified $X_i^{mis} := (X_{i1}^{mis}, \ldots, X_{i10}^{mis})'$, which are random and reflect genotypes referring to a recombination fraction with distance of $10cM$ to X_i in each component. In the left panel of Table 3 we see that in the first case φ is estimated to be 1.01. This illustrates that the GPR class contains the PoiR class and that the dispersion can be estimated with a very good precision. In the second case, however, the regressors are misspecified by not knowing the exact trait loci and the marker genotypes X^{mis} are used instead. Now φ is estimated to be 3.25 (see right panel of Table 3), i.e. the estimated variance exceeds the estimated mean by a factor of more than 10. As one can see in Table 2 this leads to a dramatic overfit when using mBIC with PoiR since this model cannot reflect the additional overdispersion and picks too many regressors in order to account for the data heterogeneity. Zero-inflation also leads to

overdispersion, however one can see in Table 1 for the ZIPR case that zero-inflation alone is insufficient to compensate the lack of the overdispersion parameter.

	Estimate	Std. Error	$Pr(>\|z\|)$		Estimate	Std. Error	$Pr(>\|z\|)$
Interc.	2.064	0.031	$< 2 \cdot 10^{-16}$	Interc.	2.197	0.081	$< 2 \cdot 10^{-16}$
X_1	-0.211	0.031	10^{-11}	X_1^{mis}	-0.097	0.097	0.316
X_2	0.920	0.038	$< 2 \cdot 10^{-16}$	X_2^{mis}	0.836	0.104	$7 \cdot 10^{-16}$
X_3	0.266	0.037	$5 \cdot 10^{-13}$	X_3^{mis}	0.211	0.103	0.041
X_4	-0.566	0.029	$< 2 \cdot 10^{-16}$	X_4^{mis}	-0.673	0.097	$5 \cdot 10^{-12}$
X_5	0.812	0.035	$< 2 \cdot 10^{-16}$	X_5^{mis}	0.626	0.105	$3 \cdot 10^{-9}$
X_6	1.228	0.049	$< 2 \cdot 10^{-16}$	X_6^{mis}	1.137	0.118	$< 2 \cdot 10^{-16}$
X_7	0.696	0.038	$< 2 \cdot 10^{-16}$	X_7^{mis}	0.379	0.102	$2 \cdot 10^{-4}$
X_8	-0.174	0.033	10^{-7}	X_8^{mis}	-0.191	0.097	0.049
X_9	-0.417	0.033	$< 2 \cdot 10^{-16}$	X_9^{mis}	-0.310	0.099	0.002
X_{10}	1.518	0.046	$< 2 \cdot 10^{-16}$	X_{10}^{mis}	1.199	0.114	$< 2 \cdot 10^{-16}$

Table 3: GP fit of Poisson data ($n = 200$) based on the 10 true effects $(X_1, \ldots, X_{10})'$ (left panel) with $\hat{\varphi} = 1.01\ (0.051)$. GP fit of the same data based on 10 misspecified effects correlated with $(X_1, \ldots, X_{10})'$ which are $10cM$ away from the true effects (right panel), $\hat{\varphi} = 3.25\ (0.477)$.

Comparing the performance of mBIC and EBIC under the most appropriate ZIGPR model we observe that both these criteria perform very well and their results do not differ much. As expected, EBIC offers slightly larger power at the price of a larger, but still reasonable, FDR. Our simulations show that the power of these criteria increases and the expected number of false positives decreases as the samples size goes up, which strongly suggest that these criteria are consistent also under the ZIGPR model.

5 Real data analysis

The data by Lyons et al. (2003) considers different phenotypes related to gallstones. While Lyons et al. (2003) focus on the gallstone weight, a score for solid gallstones and the gallbladder volume, we will focus on the number of gallstones the 277 male mice developed. The data is publicly available at

http://phenome.jax.org/phenome/protodocs/QTL/QTL-Lyons3.xls

and refers to an intercross of CAST/Ei and 129S1/SvImJ inbred mice. Since the phenotypes considered in Lyons et al. (2003, Figure 5) are related to the number

of gallstones the mice developed, we perform a preselection of interesting chromosomes based on this figure. Hence we restrict our search to eight chromosomes accounting for 41 markers, i.e. we consider the chromosomes 2, 3, 4, 5, 7, 17, 18 and 19. We replace missing genotypes by their expected values, given the flanking markers (see for instance Haley and Knott (1992)). Additive and dominance effects are added separately, according to the specification provided in Section 3, with a corresponding to the CAST/Ei allele. As a search method we used forward selection with mBIC based on ZIGPR. The reason for which we chose mBIC rather than EBIC, is that mBIC has been adapted for the search of interaction effects. In this case mBIC adjusts to the increased "multiple testing" problem by changing the penalty constant from 4 to 2.2 (see (3.2)). The adaptation of EBIC for the search of interactions is not obvious and we are not aware of existing solutions to this problem.

We performed two different analyses. At first we searched only for main effects with the standard version of mBIC (3.2) and a penalty constant provided in (3.3). In this case mBIC identifies one additive effect at D5Mit183 ("D5Mit183(a)"). This is in line with the result of Lyons et al. (2003), which found this marker to be significant for all three Gallstone related traits considered in their study. A model summary is given in the upper panel of Table 4. Note that the asymptotic normality of the maximum likelihood estimates of the dispersion parameter φ and zero-inflation parameter ω has been shown in Czado et al. (2007, Theorem 1). Therefore we report the p-values of the Wald test also for these estimates. Additionally, we performed the search for both additive and interaction effects using mBIC (3.2) with constants provided in (3.4). In this search we detected an additive-additive interaction term between two markers: D5Mit183 and a novel suggestive QTL, D4Mit42. A model summary is given in the middle panel of Table 4. Additionally, in the lower panel of Table 4 we provide the results of the analysis based on the model including additive effects of both D5Mit183 and D4Mit42 and their interaction. Interestingly, the p-value corresponding to the interaction term between D5Mit183 and D4Mit42 is substantially smaller than the p-values corresponding to the additive effects, which suggests that the interaction between D5Mit183 and D4Mit42 plays a very important role in determining the expected number of gallstones. This observation is confirmed by the graphical representation in Figure 3. In accordance with the results of the search for main effects this figure suggests that the expected number of gallstones decreases when the number of 129S1/SvImJ alleles at D5Mit183 increases. However, according to the bottom graph, the effect of D5Mit183 strongly depends on the genotype at D4Mit42, and is most pronounced for mice who are homozygous for 129S1/SvImJ allele at D4Mit42. Specifically, the average number of gallstones is decisively the largest in the group of mice with the combination of dummy variables equal to (-1,1), which corresponds to the mice homozygous for

	Estimate	Std. Error	z value	$Pr(>\|z\|)$
Intercept	0.067	0.983	0.068	0.946
D5Mit183(a)	−1.292	0.432	−2.991	0.003
φ	6.799	3.560	1.909	0.056
ω	0.631	0.362	1.743	0.081
Intercept	−0.156	0.572	−0.272	0.786
D5Mit183(a):D4Mit42(a)	−2.298	0.495	−4.647	$3.4 \cdot 10^{-6}$
φ	5.776	2.520	2.293	0.022
ω	0.575	0.167	3.437	0.001
Intercept	−0.864	0.573	−1.510	0.131
D5Mit183(a)	−1.244	0.442	−2.817	0.005
D4Mit42(a)	−0.215	0.476	−0.451	0.652
D5Mit183(a):D4Mit42(a)	−2.177	0.548	−3.973	$7.1 \cdot 10^{-5}$
φ	5.387	2.185	2.466	0.014
ω	0.458	0.163	2.809	0.005

Table 4: ZIGP model summaries of forward selection based on mBIC for different regression designs.

CAST/Ei allele at D5Mit183 and for 129S1/SvImJ allele at D4Mit42. Finally we add that we also carried out a scan based on the LM. Neither in the search over main effects nor in the search including epistatic effects a significant effect could be caught.

6 Discussion

We investigated the applicability of different versions of Poisson regression and the modified Bayesian Information Criterion for locating multiple interacting quantitative trait loci influencing count traits. Our research demonstrates very good properties of the zero-inflated generalized Poisson regression in this context. ZIGPR takes into account both the overdispersion and an over-excess of zeros and performs much better than simplified versions of Poisson regression in case when both these parameters play an important role. Moreover, we found out that the overdispersion parameter allows to compensate for a model misspecification due to the discrepancy between marker and QTL locations. Therefore, the search for markers associated with the count trait based on ZIGPR gives much better results than the one based on the standard Poisson regression, even when the data are generated according the latter. Also, our simulations illustrate very good properties of the modified versions of the Bayesian Information Criterion, mBIC and EBIC, as applied to select

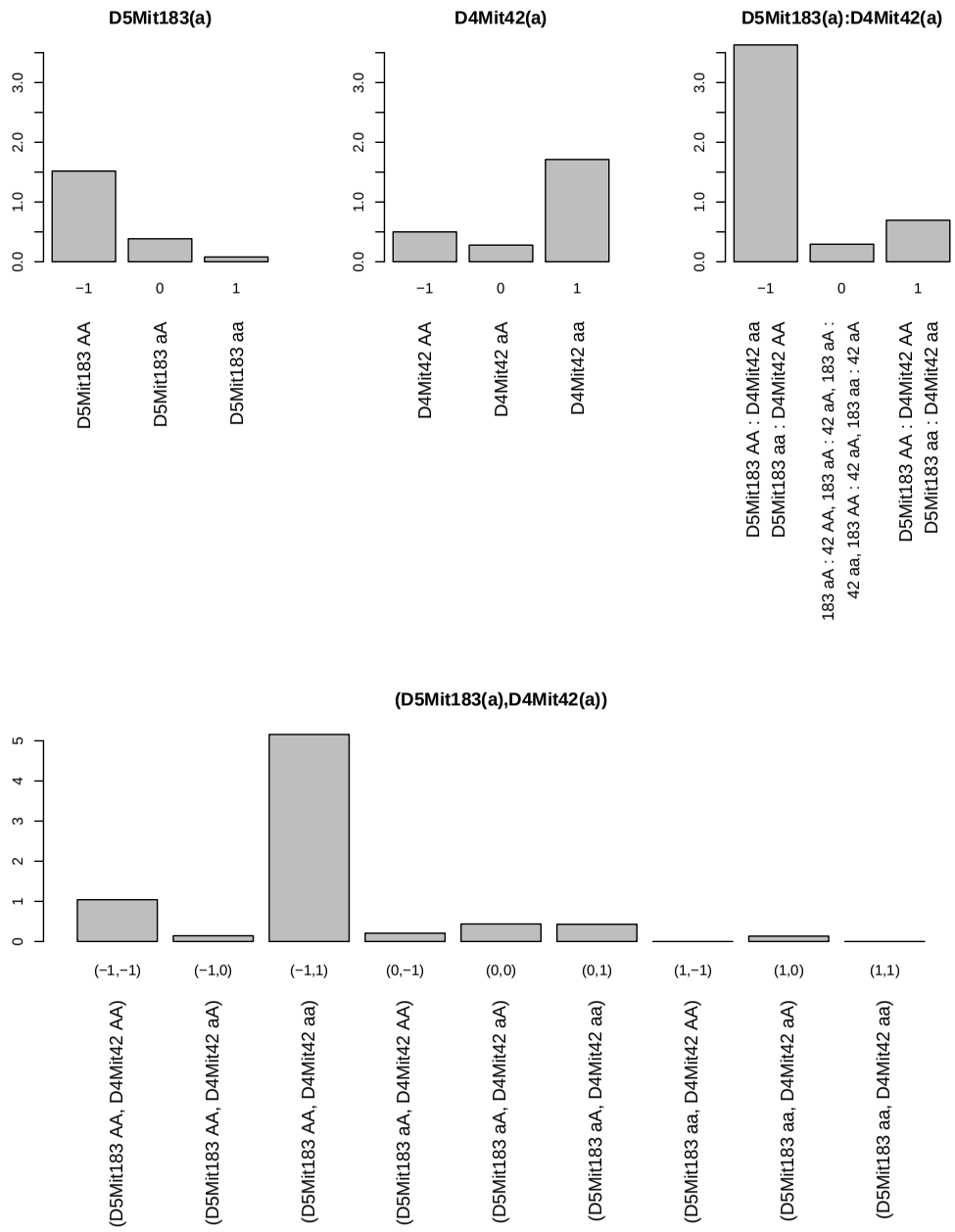

Figure 3: Average number of gallstones in different groups of mice, specified by dummy variables corresponding to the additive effects of D5Mit183 and D4Mit42.

important predictors for ZIGPR. Both these criteria perform in a similar way and guarantee a good power of QTL detection, while keeping the false discovery rate at a low level. The reported real data analysis shows the possible gains, which can be obtained when ZIGPR with mBIC is used for detection of interacting QTL.

Good properties of mBIC in the context of sparse orthogonal multiple regression were confirmed by the results on its asymptotic optimality, proved in Bogdan et al. (2008a). Our preliminary results suggest that similar asymptotic optimality results can be proved for EBIC. However, the extension of these results to the nonorthogonal designs and ZIGPR models presents a major challenge and remains a topic for future research.

Due to the complexity of a large scale simulation study, whose main purpose was the comparison of different Poisson regression models, we reduced the attention to the search over markers. Note that the computational effort for the simulation study carried out in Table 1 was very high. We made quite some effort to optimize the *R* code, nevertheless the repeated search for significant effects over the 100 main effects was running for more than 20 days on a parallelized 32-core cluster with 2.6 GHz processors. However, an extension of the proposed methodology to the multiple interval mapping is in general quite straightforward and, concerning the estimates of QTL effects and positions, goes along the line of an interval mapping for ZIGPR, as proposed in Cui and Yang (2009). Concerning the estimate of a QTL number, a successful application of EBIC for the multiple interval mapping with mixture General Linear Models was presented in Li and Chen (2009). Also, the results reported in Bogdan et al. (2008b) show that if markers are on the average distant by more than 5 cM then mBIC may be successfully used with the multiple interval mapping. However, the results reported in Bogdan et al. (2008b) show also that if markers are densely spaced (less than 5 cM apart) then the neighboring marker genotypes are strongly correlated and the penalty in mBIC and EBIC could be substantially relaxed. We believe that the corresponding scaling coefficients provided in Bogdan et al. (2008b) would work well also for ZIGPR but an exact verification requires a very intensive simulation study and is out of the scope of the present paper.

To reduce the complexity of our simulation study we identified the best regression model with a forward selection. Our simulations, as well as results reported in Broman (1997), Broman and Speed (2002), and Bogdan et al. (2004), show that the forward selection usually performs well in the context of QTL mapping. However, the real data analysis reported in Bogdan et al. (2008b) illustrates that in the case when there are many linked QTL this procedure may fail to identify the optimal model. The uncertainty related to the model choice can be well expressed within the Bayesian framework by the posterior model probabilities. The Bayesian approach for the analysis and comparison of ZIGPR models was investi-

gated e.g. in Gschlößl and Czado (2006). However, the computational complexity of the full Bayes analysis by Markov Chain Monte Carlo (MCMC) substantially limits its range of applications in the context of localizing multiple interacting QTL. Note however that both mBIC and EBIC allow an approximation to the posterior probabilities of different models according to

$$P(M_i|Y) \approx \frac{\exp(xBIC(i)/2)}{\sum_j \exp(xBIC(j)/2)} \ , \tag{6.1}$$

where *xBIC* denotes mBIC (3.2) or EBIC (3.6) and the sum in the denominator is over all possible ZIGPR models. Thus, to estimate the posterior probability of a given model by the modified BIC it is enough to visit each of the plausible models just once. This allows to substantially reduce the computational burden in comparison to the MCMC methods, which typically require multiple visits of each model, and then estimate the posterior probability by the frequency of such visits. However, the estimate of $P(M_i|Y)$ provided in (6.1) may be accurate only if the majority of plausible models is represented in the denominator. Therefore, to use mBIC or EBIC in a Bayesian context, a suitable, computationally efficient search strategy still needs to be developed.

Acknowledgment. MB gratefully acknowledges the support from the Department of Mathematics of Munich University of Technology through the *Women for Math Science Award.* We also would like to thank two anonymous referees for helpful comments and suggestions.

References

Baierl, A., M. Bogdan, F. Frommlet, and A. Futschik (2006). On Locating Multiple Interacting Quantitative Trait Loci in Intercross Designs. *Genetics 173*(3), 1693–1703.

Baierl, A., A. Futschik, M. Bogdan, and P. Biecek (2007). Locating multiple interacting quantitative trait loci using robust model selection. *Computational Statistics and Data Analysis 51*, 6423–6434.

Ball, R. (2001). Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian information criterion. *Genetics 159*(3), 1351–1364.

Bogdan, M., A. Chakrabarti, and J.K.Ghosh (2008a). Optimal rules for multiple testing and sparse multiple regression. *Technical Report I-18/08/P-003*. Institute of Mathematics and Computer Science, Wroclaw University of Technology, *www.im.pwr.wroc.pl/∼mbogdan/Preprints*.

Bogdan, M., F. Frommlet, P. Biecek, R. Cheng, J. Ghosh, and R. Doerge (2008b). Extending the Modified Bayesian Information Criterion (mBIC) to dense markers and multiple interval mapping. *Biometrics 64*(8), 1162–1169.

Bogdan, M., J. Ghosh, and R. Doerge (2004). Modifying the Schwarz Bayesian Information Criterion to locate multiple interacting quantitative trait loci. *Genetics 167*(2), 989–999.

Bogdan, M., J. Ghosh, and M. Żak-Szatkowska (2008c). Selecting explanatory variables with the modified version of Bayesian Information Criterion. *Quality and Reliability Engineering International 24*, 627–641.

Broman, K. (1997). Identifying quantitative trait loci in experimental crosses. PhD dissertation. Department of Statistics, University of California, Berkeley, CA.

Broman, K. (2003). Mapping quantitative trait loci in the case of a spike in the phenotype distribution. *Genetics 163*(3), 1169–1175.

Broman, K. and T. Speed (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *J. Roy. Stat. Soc. B 64*, 641–656.

Chen, J. and Z. Chen (2008). Extended Bayesian Information criteria for model selection with large model spaces. *Biometrika 95*(3), 759–771.

Chen, Z. and J. Liu (2009). Mixture Generalized Linear Models for Multiple Interval Mapping of Quantitative Trait Loci in Experimental Crosses. *Biometrics 65*(2), 470–477.

Coffman, C., R. Doerge, K. Simonsen, K. Nichols, and C. Duarte (2005). Model selection in binary trait locus mapping. *Genetics 170*(3), 1281–1297.

Consul, P. C. (1989). *Generalized Poisson distributions*, Volume 99 of *Statistics: Textbooks and Monographs*. New York: Marcel Dekker Inc. Properties and applications.

Consul, P. C. and F. Famoye (1992). Generalized Poisson regression model. *Comm. Statist. Theory Methods 21*(1), 89–109.

Consul, P. C. and G. C. Jain (1970). On the generalization of Poisson distribution. *Ann. Math. Statist. 41*, 1387.

Cui, Y., Kim, D.-Y., and Zhu, J. (2006). On the generalized Poisson regression mixture model for mapping quantitative trait loci with count data. *Genetics 3*: 2159-2172.

Cui, Y. and W. Yang (2009). Zero inflated generalized Poisson regression mixture model for mapping quantitative trait loci underlying count trait with many zeros. *Journal of Theoretical Biology 256*(2), 276–285.

Czado, C., V. Erhardt, A. Min, and S. Wagner (2007). Zero-inflated generalized Poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates. *Statistical Modelling 7*(2), 125–153.

Dupuis, J. and D. Siegmund (1999). Statistical methods for mapping quantitative trait loci from a dense set of markers. *Genetics 151*, 373–386.

Famoye, F. and K. P. Singh (2003). On inflated generalized Poisson regression models. *Adv. Appl. Stat. 3*(2), 145–158.

Famoye, F. and K. P. Singh (2006). Zero-inflated generalized Poisson model with an application to domestic violence data. *Journal of Data Science 4*(1), 117–130.

Fijneman, R. J. A., S.S. De Vries, R. C. Jansen and P. Demant (1996) Complex interactions of new quantitative trait loci, sluc1, sluc2, sluc3, and sluc4, that influence the susceptibility to lung cancer in the mouse. *Nat. Gen. 14*, 465–467.

Ghosh J.K., M. Delampady and, T. Samanta (2006) *An introduction to Bayesian analysistheory and methods*. Springer, Berlin / Heidelberg.

Gschlößl, S. and C. Czado (2006) Modelling count data with overdispersion and spatial effects. *Statistical Papers.* DOI 10.1007/s00362-006-0031-6

Haley, C. and S. Knott (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity 69*, 315–324.

Jansen, R. (1993). Interval mapping of multiple quantitative trait loci. *Genetics 135*(1), 205–211.

Jansen, R. and P. Stam (1994). High resolution of quantitative traits into multiple loci via interval mapping. *Genetics 136*(4), 1447–1455.

Joe, H. and R. Zhu (2005). Generalized Poisson distribution: the property of mixture of Poisson and comparison with negative binomial distribution. *Biom. J. 47*(2), 219–229.

Kao, C. (2000). On the Differences Between Maximum Likelihood and Regression Interval Mapping in the Analysis of Quantitative Trait Loci. *Genetics 156*, 855–865.

Kao, C. and Z. Zeng (2002). Modeling Epistasis of Quantitative Trait Loci Using Cockerham's Model. *Genetics 160*, 1243–1261.

Kao, C., Z. Zeng, and R. Teasdale (1999). Multiple Interval Mapping for Quantitative Trait Loci. *Genetics 152*(3), 1203–1216.

Kruglyak, L. and E. Lander (1995). A nonparametric approach for mapping quantitative trait loci. *Genetics 139*(3), 1421–1428.

Lambert, D. (1992). Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing. *Technometrics 34*, 1–14.

Lander, E. and D. Botstein (1989). Mapping mendelian factors underlying quantitative traits using rflp linkage maps. *Genetics 121*(1), 185–199.

Li, J., S. Wang, and Z.-B. Zeng (2006). Multiple interval mapping for ordinal traits. *Genetics 173*(3), 1649–1663.

Li, W. and Z. Chen (2009). Multiple interval mapping for quantitative trait loci with a spike in the trait distribution. *Genetics 182*(2), 337–342.

Lyons, M. A., H. Wittenburg, R. Li, K. A. Walsh, M. R. Leonard, G. A. Churchill, M. C. Carey, and B. Paigen (2003). New quantitative trait loci that contribute to cholesterol gallstone formation detected in an intercross of CAST/Ei and

129S1/SvImJ inbred mice. *Physiol. Genomics 14*(3), 225–239.

Manichaikul, A., J. Moon, S. Sen, B. Yandell, and K. Broman (2009). A model selection approach for the identification of quantitative trait loci in experimental crosses, allowing epistasis. *Genetics 181*(3), 1077–1086.

Min, A. and C. Czado (2010). Testing for zero-modification in count regression models. *Statistica Sinica 20*, 323–341.

Mullahy, J. (1986). Specification and testing of some modified count data models. *J. Econometrics 33*(3), 341–365.

Sax, K. (1923). The association of size difference with seed-coat pattern and pigmentation in phaseolus vulgaris. *Genetics 8*(6), 552–560.

Scott, J.G. and J.O. Berger (2008). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. Duke University Department of Statistical Science, Discussion Paper 2008-10, to appear in *Ann. Statist.*.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics 6*, 461–464.

Shao, J. (1999). *Mathematical Statistics*. Springer-Verlag, New York.

Thomson, P. (2003). A generalized estimating equations approach to quantitative trait locus detection of non-normal traits. *Genet. Sel. Evol. 3*, 257–280.

Yi, N., S. Banerjee, D. Pomp, and B. Yandell (2007). Bayesian mapping of genomewide interacting quantitative trait loci for ordinal traits. *Genetics 176*(3), 1855–1864.

Yi, N., S. Xu, V. George, and D. Allison (2004). Mapping multiple quantitative trait loci for complex ordinal traits. *Behav. Genet. 34*, 3–15.

Zeng, Z. B. (1993). Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci. *Proc. Natl. Acad. Sci. USA 90*, 10972–10976.

Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. *Genetics 136*(4), 1457–1468.

Zou, F., B. Yandell and J. Fine (2003). Rank based statistical methodologies for QTL mapping. *Genetics 165*(3), 1599–1605.

Żak, M., A. Baierl, M. Bogdan, and A. Futschik (2007). Locating multiple interacting quantitative trait loci using rank-based model selection. *Genetics 176*(3), 1845–1854.

Żak-Szatkowska M. and M. Bogdan (2010). Applying generalized linear models for identifying important factors in large data bases. *Technical Report I-18/2010/P-001*. Institute of Mathematics and Computer Science, Wroclaw University of Technology, *www.im.pwr.wroc.pl/∼mbogdan/Preprints*.