Bayesian analysis of high-dimensional count data

by

Ignacio Alvarez-Castro

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:

Jarad Niemi, Major Professor

Alicia Carriquiry

Peng Liu

Dan Nettleton

Dan Nordman

Iowa State University

Ames, Iowa

2017

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank, Jarad for his advice, commitment, patience, and understanding greatly helped me in my first steps into the statistical research.

Also, I feel grateful to Dan Nettleton and Phil Dixon, together with Jarad are the professors with I have shared most of my academic work at ISU, learning something new every time.

Alicia helped me arriving in Ames and to have an excellent beginning in the program.

There are many friends and family to thank for the love and support they have sent me locally, nationally and overseas.

I must thanks to Naty, as in any other aspect of my life, she is the energy that kept me moving forward in this project.

CHAPTER 1. General introduction

This thesis describes my research work in past years in the Statistic Department of Iowa State University. There are several key statistical features common to the whole thesis.

In the first place, all the statistical methods are developed taking a Bayesian perspective to conduct the statistical inference. A second common feature of the two main parts is that both correspond to high-dimensional problems. In the first case because large amount of information for a few individuals is available, and in the second part due to model space is really large which brings computational intractability issues. Finally, the response variable in all data used here is a positive count, in the first part it is associated with the gene expression while in the second part it represents a number of automobile crashes.

Nevertheless, this thesis can be organized in two main parts dealing with different applications.

1.1 Hierarchical Bayesian analysis of allele-specific gene expression data

In recent years, next generation sequencing (NGS) technology has evolved enough to offer a more accurate and cost effective means of studying a variety of genomic signals with a wide range of applications (Datta and Nettleton, 2014). We define the gene's expression level as the amount of messenger RNA (transcript abundance) produced. For each gene, RNA-sequencing (RNA-seq) is a count positively correlated with the transcript abundance.

Diploid organisms have two copies of each genes (alleles) that can be separately transcribed. The RNA abundance of any particular allele is known as allele-specific expression (ASE). When an mRNA read can be identified to particular allele, ASE can be studied with RNA-seq read count data. Reads counts that can be unambiguously attributed to a specific allele are

correlated with allele's expression. (Sun and Hu, 2014).

In plant breeding, it is common that hybrid lines show improvements in several phenotype traits compared with its parent lines. The effect that a heterozygous hybrid is better compared with the average of its homozygous parents is called hybrid vigor or *heterosis* (Schnable and Springer, 2013). A relationship between heterosis and some gene ASE patterns has been suggested, for instance, Paschold et al. (2012) found the differential allele-specific expression relates to non-additive patterns in total RNAseq expression.

One of the main goals of this research is to build statistical methodologies to identify genes with preferential mRNA expression from one of the two alleles. In order to do this, statistical models for ASE counts from hybrid cross lines are developed. The most important characteristics specific to ASE compared to total RNA-seq counts are addressed by the statistical models proposed in this work.

Additionally, this thesis proposes measures for assessing the statistical association of ASE patterns and gene heterotic patterns. These types of measures might help in improve the connections among gene expression patterns at the allele-specific level with patterns indicating hybrid vigor. To achieve this goal, ASE counts and total RNA-seq for inbred lines and it hybryd cross are used.

Chapter 2 presents a modeling strategy for ASE information, in the case of having data from one single hybrid genotype. A hierarchical Poisson-lognormal mixture model is proposed addressing the main characteristics of ASE data. In Chapter 3 jointly heterotic patterns and allelic imbalance are modeled jointly, expanding the model from Chapter 2 to more genotypes and total RNA-seq as well as ASE gene expression data. Some association measures between them heterotic patterns and allelic imbalance are proposed.

A key aspect of the proposed models is its hierarchical nature. A hierarchical model make use of latent information between the groups that only can be seen when the estimation problem consider all groups together (Efron, 1992). Within a context of high dimensional problems, where there are a large number of groups but only a few observations in each group, sharing information among groups. In particular, in the analysis of gene expression data the information sharing approach has been extremely successful.

Most traditional hierarchical models promote the information sharing among one set of parameters. However the models proposed to deal ASE counts have more than one set of parameters, and the borrowing of information process occurs in each set of parameters simultaneously. Another piece contained in this work consist in an exploration of the interactions between the learning of several sets of group specific parameters distributions.

Chapter 4 deal with hierarchical models for both mean and variances simultaneously in sparse high dimensional context, which is an important aspect of the model used in Chapter 2. Using log-transformed ASE counts, we focus on the effects of variance hierarchical modeling on the mean vector inference.

1.2 Approximate Bayesian computation for crash frequency models

Approximate Bayesian computation (ABC) is a field of Bayesian research that has gained much popularity in recent years (Marin et al., 2012). This constitutes a powerful estimation technique based on simulations, these methods are designed for complex problems where the likelihood is computationally or analytically intractable.

The methods presented in Chapter 5 can potentially work with areal-referenced count data. The response variable is a discrete variable which is available in a set of locations, while the covariates are continuous variables also available at location level. Spatial dependence is introduced through the neighborhood structure, locations might be connected or not, according the neighbor structure.

The modeling of crash frequency data is a field of extensive research. At the national level, motor vehicle fatalities account for approximately 30% of all injury deaths in the United States every year. In Iowa, about 400 lives are lost annually in traffic accidents and crashes represent a total cost of 1 billion dollars per year (McDonald, 2012). Therefore, preventing crashes or at least minimizing the loss of life and major injuries due to crashes is critically important.

Chapter 5 proposes a model crash frequency at the intersection level while introducing spatial correlation among intersections. A Winsorized Poisson Markov random fields (PMRF) is used (Kaiser and Cressie, 1997) for this purpose. However, the model is not computationally tractable, an ABC approach is used in order to obtain Bayesian inference.

CHAPTER 2. Fully Bayesian analysis of allele-specific RNA-seq data using an hierarchical, overdispersed, count regression model

Abstract

Diploid organisms have two copies of each gene (alleles) that can be separately transcribed. The RNA abundance of any particular allele is known as allele-specific expression (ASE). When two alleles have sequences of polymorphisms in transcribed regions, ASE can be studied with RNA-seq read count data. Reads counts that can be unambiguously attributed to a specific allele are correlated with allele's expression.

In plant breeding, hybrids are developed to take advantage of the genetic phenomenon known as heterosis or hybrid vigor. Heterosis occurs when hybrid offspring possess superior levels of one or more traits relative to their inbred parents. ASE is relevant for the study of this phenomenon at the molecular level. One possible reason for the occurrence of heterosis are genes where two distinct alleles at a heterozygous locus are differentially expressed.

ASE has some characteristics different from the regular RNA-seq expression: ASE cannot be assessed for every gene, measures of ASE can be biased towards one of the alleles (reference allele), and presents subsampling.

We present statistical methods for modeling ASE and detecting genes with differential allele expression. We propose a hierarchical overdispersed Poisson model to deal with ASE counts. The model accommodates gene-specific overdispersion, it has an internal measure of the reference allele bias, and use random effects to model the gene-specific regression parameters. Fully Bayesian inference is obtained using `fbseq` package that implements a parallel strategy to make the computational times reasonable. Simulation and real data analysis suggest the proposed model is a practical and powerful tool for the study of differential allele expression.

2.1 Introduction

Over the past decade, RNA-sequencing (RNA-seq) has been replacing the microarray technology as the method used to measure gene expression (Datta and Nettleton, 2014). In a biological sample, the amount of messenger RNA (transcript abundance) produced by a gene is known as the gene's expression level. For each gene, RNA-seq is a count positively correlated with the transcript abundance. A diploid genome has two sets of chromosomes, one from each parent, so every gene has two copies. One of the advantages of next generation sequencing is that makes possible to measure the expression of each gene copy, we call allele-specific expression (ASE) to refer this measure. ASE can be obtained using single nucleotide polymorphism (SNP) that makes it possible to distinguish the expression of the two alleles (Sun and Hu, 2014).

The study of ASE may provide some explanation for the so-called heterosis effects. In plant breeding, phenotypic heterosis occurs when hybrid lines show improvements in several phenotype traits compared with its inbred parent lines(Schnable and Springer, 2013). Heterozygous hybrid varieties might take advantage of having two alleles with different genotypes in order to adapt to environmental conditions by promoting the selection of the superior allele. The uneven expression of alleles might be related to the superior adaptation of hybrids, so it might be related to the occurrence of gene heterosis (Paschold et al., 2012; Bell et al., 2013). Other biological questions where ASE is relevant may include identifying imprinting or parent-of-origin effects, which occurs in genes where only one parental allele is expressed, the distinction between cis-acting and trans-acting regulation DNA relies on ASE since cis-acting is associated with differentially expressed alleles while trans-acting has effects both alleles. (Sun and Hu, 2014)

Given the total ASE, i.e., the sum of counts in both alleles, the reference allele count can be modeled as binomially distributed. Differentially expressed genes can be obtained applying a binomial test for each gene and adjusting p-values to control false discovery rate (FDR) (Bell et al., 2013). Binomial distribution can be combined with a beta distribution for the probability parameter to create a hierarchical model. Pirinen et al. (2014) use a mixture of betas as the

probability parameter prior, with 3 components corresponding to a degree of allelic difference (none, moderate, and strong).

Beta-binomial distribution has also been proposed for modeling the reference allele count (Sun and Hu, 2013), this model includes gene-specific overdispersion. Both, total RNA-seq expression and ASE can be combined to distinguish factors that affect the gene expression in an allele-specific manner (*cis*-QTL) from factors that affect the gene expression of the two alleles at the same time (*trans*-QTL). A likelihood ratio test distinguishes *cis* and *trans* regulation by combining ASE beta-binomial model with a model for the total RNA-seq counts (Sun and Hu, 2014). The model is extended in Hu et al. (2015) to incorporate isoform-specific information and haplotype modeling. However, the beta-binomial proposed model uses haplotype information (or models it), so it is hard to apply in plant experiments where hybrids varieties are created from inbred lines.

Instead of modeling ASE counts based on a binomial distribution, it is possible to adapt models originally designed for dealing with total RNA-seq transcript abundance counts using Poisson or negative binomial distributions. Lorenz et al. (2014) provide an extensive review of the methods to detect differential expression for total RNA-seq data. A Poisson generalized linear model per gene can be fit, including random terms for experimental unit and overdispersion effect, and controlling FDR to determine genes with differential allele expression (Paschold et al., 2012). Generalized Poisson distribution (Srivastava and Chen, 2010) has been adapted to analyze ASE (Wei and Wang, 2013)

In this paper, a hierarchical overdispersed count regression model is proposed to study allele-specific expression, in the case where data from a single hybrid genotype are available. A Poisson data model for ASE appropriately treat both allele counts as random variables, while models based on binomial implicitly treat the non-reference allele count as given. At the same time, the Poisson data model is flexible enough to include more alleles, genotypes or even the total RNA-seq count. We describe how the proposed model is able to capture the key features of ASE data and show this is done with a simulation study. In addition, a fully Bayesian analysis of this model is possible using `fbseq` package (Landau and Niemi, 2016).

The next Section describe the main characteristic ASE data different from the total RNA-

seq expression, using a maize experiment ASE data set. Section 2.3 describes the hierarchical overdispersed count regression model we propose to analyze ASE data. Sections 2.4 and 2.5 presents results from a simulation study and a real ASE data set analysis, respectively. Finally, 2.6 presents a summary of the main findings and comments on the next steps in this line of research.

2.2 Allele-specific expression

RNA-seq counts are obtained mapping short reads to an annotated genome. In the case of ASE counts, a distinction between two genomes is needed, a small difference in the genome sequence called single nucleotide polymorphism (SNP) provides a way to make that distinction. An important issue of ASE data is that when no SNP is discovered for a gene therefore there is no ASE information for that gene. The proportion of genes having ASE information available depends on how similar the two genomes are and with the read length (Sun and Hu, 2014).

Let assume the ASE counts for a single hybrid variety are available, we refer to the hybrid variety with ASE data as BM, and its parents as the varieties B and M. The dataset is then formed by two transcript abundance counts per gene in the hybrid BM, each count positively correlated with the gene expression of allele B and allele M respectively. In plant breeding experiments, is common that the parental varieties, B and M, are inbred lines. Because of this, haplotypes are known and identical to the genotype, in other words there is no information in the haplotype. Then we cannot use models like proposed in Sun and Hu (2013) to analyze this data set.

Let m_{ga} be the average gene-expression level of gene $g \in \{1, \ldots, G\}$ and allele $a \in \{B, M\}$, averaged over all available replicates. Consider two summary measures per gene, a measure of the average ASE abundance (A_g) and a measure of the ASE ratio among the two alleles (M_g). These summary measures per gene are computed in log-scale and can be written as follows:

$$A_g = \log(m_{gB} + m_{gM}), \quad M_g = \log\left(\frac{m_{gB}}{m_{gM}}\right) .$$

Figure 2.1 illustrates some characteristics usually present in ASE gene transcript abundance data using the summary measures A_g and M_g. Some basic details of the specific dataset used

to produce Figure 2.1 are described in Section 2.5.

Left panel in Figure 2.1 presents a histogram of the gene-specific allele expression ratio, M_g, the blue line is a Gaussian density with the sample mean and sample variance. The plot suggests most of the genes present very small differences in the ASE counts within the same gene. The empirical distribution of M_g is more concentrated around 0 and heavier tails than a normal distribution with the observed mean and variance. This characteristic is not exclusive of ASE counts, differential expression measures in total RNA-seq counts and microarrays are usually sparse effects.

Right panel in Figure 2.1 shows a two-dimensional histogram of the pairs (A_g, M_g), where the color intensity of the cells indicate how frequent is that region in the data. The most frequent cells are close to zero allele difference ($M_g = 0$) for any level of average ASE (A_g), suggesting that most of the genes present very small differences in the ASE counts. In addition, right panel in Figure 2.1 shows all frequent cells seems to be in $M_g > 0$ region, and the average of the difference among all genes (red line) is also positive. This suggests that allele B present larger ASE counts than allele M, on average across all genes.

While it could be some biological reason to observe one of the alleles consistently more expressed than the other one in every gene, it is known the ASE measurement process results in towards the reference allele. In practice, there is only one genome to compare with the short reads, this is called the reference allele. Reads matching the reference genome are assigned to the reference allele while reads having a mismatch with the reference genome are assigned to the non-reference allele. It is known this procedure implies a bias that favors the reference allele (Panousis et al., 2014). Reference genome is (almost) totally known, and many times is not possible to distinguish mismatches due to errors from genuine mismatches due to the read corresponding to a non-reference genome. Then a read that truly matches the reference genome is more likely to be counted than a read matching the non-reference genome, creating a bias towards reference allele counts. The bias toward reference allele might be alleviated when the ASE data is obtained (Degner et al., 2009; Vijaya Satya et al., 2012; Stevenson et al., 2013). But it is not clear how to deal with the bias effect at the modeling stage. A conservative strategy consists in only consider genes with significant allele imbalance against the reference

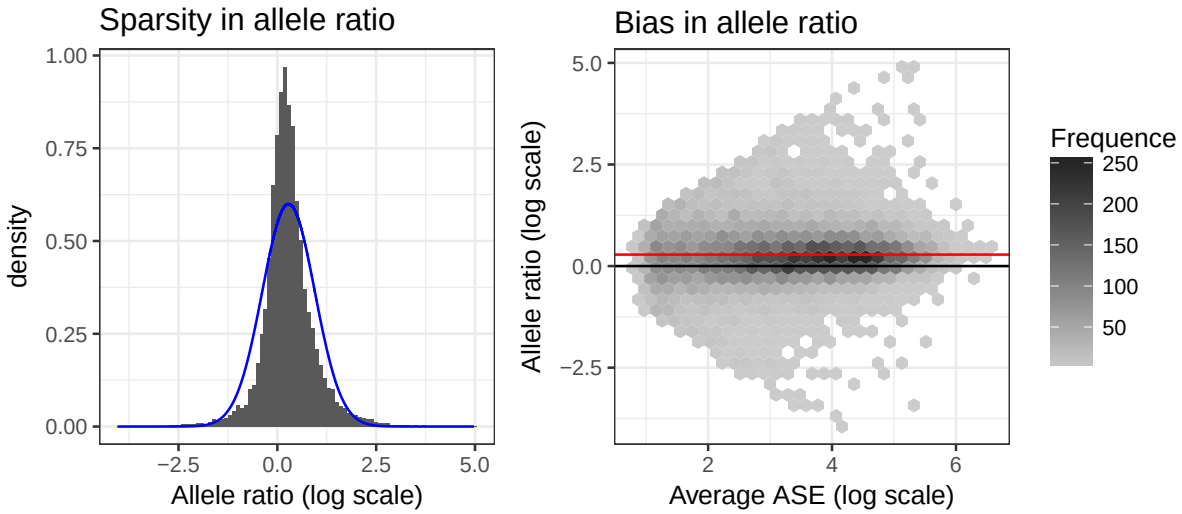

Figure 2.1: (Left panel) Histogram of the ratio of allele means per gene in logs. Blue line is a normal density curve with mean and variance equal to mean and variance of the ratio across all genes. (Right Panel) Hexagon binning plot of means per gene and allele. Vertical axis is the ratio of allele means per gene in logs while horizontal is the allele-specific mean expression in logs. The red line indicates the overall mean difference expression among the two alleles means.

allele (Paschold et al., 2012).

2.3 Hierarchical overdispersed count regression model

We model the RNA-seq count of each allele with an overdispersed hierarchical overdispersed count regression model. The ASE observed count for each allele should be connected with a random variable since there is uncertainty in the expression measurement of both alleles counts. Furthermore, modeling allele counts will it make easier to enlarge the model to include more genotype types, tissue types, to deal with cases with more than two alleles, or to include the total RNA-seq count. Larger models for ASE will allow researchers to approach more complex biological questions. The count regression model is composed of two main blocks: a data model setting up a count regression model based on a Poisson-lognormal mixture distribution, and a hierarchical distribution block for gene-specific regression coefficients and overdispersion variance. The hierarchical distributions allow sharing information across all genes, improving the posterior distribution inference of the gene-specific relevant quantities (defined later). As the hierarchical distribution is learned, the amount of shrinkage or shared information is determined by the data. The rest of this Section is dedicated to describing in more detail each block of the model.

2.3.1 Data model

Let Y_{gn} be the allele-specific RNA-seq count of gene g, in observation n, there are two observational sub-samples within each biological sample in the data presented in previous Section. Equation (2.1) describes the Poisson lognormal mixture we use as data model.

$$\begin{aligned} Y_{gn} &\overset{\text{ind}}{\sim} P(e^{h_n + x_n^\top \beta_g + \epsilon_{gn}}) \\ \epsilon_{gn} &\overset{\text{ind}}{\sim} N(0, \gamma_g) \end{aligned} \tag{2.1}$$

The factor h_n represents the log of the library size for sample n, including both alleles counts in the library size since came from the same experimental unit.

Model matrix X is the same for all genes, formed by $x_n^\top$ on its rows. It is a $N \times p$ dimensional matrix, where N represent the numbers of allele-specific subsamples and p the

number of covariates or effects included in the model. The columns of the model matrix X determines the interpretation of the β_g parameters. There might be many columns in the model matrix X specific to particular applications, for instance to represent blocking factors or relevant covariate effects. However, there are 2 columns that should be present in models dealing with ASE counts. We assume the first column corresponds to an intercept term and denote its associated coefficient as β_{g1}. Moreover, we assume the second model matrix column take value 1 for observed counts from the reference allele and the value -1 for observed counts of the non-reference allele. Then, the regression coefficient associated with the second column, β_{g2}, represents the half difference of gene ASE, genes with $\beta_{g2} = 0$ shows evidence of equally expressed alleles. Lastly, if there are more than one biological replicate (which is usually the case), a third column should be included to represent the grouping effect of the allele-specific sub-sampling. We assume this effect it corresponds to the last column in X, its associated coefficient is β_{gp}.

The third piece, are the overdispersion effects, ϵ_{gn}, are normally distributed with a gene-specific variance, γ_g. This effect implies quadratic mean-variance relationship that could differ across genes, and admits the partition of the total gene variability into technical and biological components similar to the Poisson-gamma mixture (Chen et al., 2014).

2.3.2 Gene-specific layer

A second layer in the model hierarchy is composed of distributions for the gene-specific parameters β_g and γ_g.

One feature common to RNA-seq and ASE counts is that in many cases, the effects of interest are only present for a small group of genes, which usually translate in that the gene-specific regression coefficients are concentrated around its mean, we refer this as a sparsity pattern in regression coefficients. For instance, the difference among alleles within each gene presented in left panel of Figure 2.1, follows this pattern.

Sparsity pattern can be addressed using shrinkage distributions, i.e., a distribution placing a great mass around its mean together with heavy tails. Equation (2.2) presents alternative shrinkage prior we consider as possible hierarchical distribution of the gene-specific regression

parameters, i.e., the components of β_g vector, β_{gk}. All these hierarchical distribution are formed as normal scale mixture.

$$\begin{array}{lll} \text{Cauchy} & \beta_{gk} \overset{\text{ind}}{\sim} N(\theta_k,\ \sigma_k^2\xi_{gk}) & \xi_{gk} \sim IG(1/2, 1/2) \\ \text{Student-t} & \beta_{gk} \overset{\text{ind}}{\sim} N(\theta_k,\ \sigma_k^2\xi_{gk}) & \xi_{gk} \sim IG(3, 2) \\ \text{Laplace} & \beta_{gk} \overset{\text{ind}}{\sim} N(\theta_k,\ \sigma_k^2\xi_{gk}) & \xi_{gk} \sim Exp(1) \\ \text{horseshoe} & \beta_{gk} \overset{\text{ind}}{\sim} N(\theta_k,\ \sigma_k^2\xi_{gk}) & \xi_{gk} \sim Ca^{+}(0, 1) \end{array} \tag{2.2}$$

A scaled t distribution is defined as If $Z \sim \mathrm{t}_d(m, v)$ then for z and m on the real line and $v > 0$, $d > 0$, its density function is

$$f_Z(z) = \frac{\Gamma((d+1)/2)}{\Gamma(d/2)} \frac{1}{\sqrt{d\pi}v} \left(1 + \frac{1}{d}(\frac{z-m}{v})^2\right)^{-(d+1)/2},$$

the rest of hierarchical distributions from (2.2) correspond to Laplace distribution (Park and Casella, 2008), horseshoe distribution (Carvalho et al., 2009), and Cauchy distribution i.e. a special case of t with 1 degrees of freedom.

Shrinkage distributions have receive a lot of attention in recent years to use as prior distributions, in that context horseshoe distribution is proposed as a default prior to use in sparse scenarios (Hahn and Carvalho, 2015). Usually, these shrinkage distribution contains only a scale parameter, that regulate the global amount of shrinkage in a particular application. However, in the model we propose here, both parameters (scale and mean) of the hierarchical distribution of the gene-specific regression coefficients are learned from data. Then, is not clear which shrinkage distribution might work better in this context. We include the hierarchical distribution as a relevant factor in a simulation study to determine which one to use in the data analysis.

A second set of gene-specific parameters are the overdispersion variances, γ_g. We model these variances as independent inverse gamma distributions conditional on ν and τ, and independent from the regression coefficients β_g.

$$\gamma_g \overset{\text{ind}}{\sim} IG(\tfrac{\nu}{2}, \tfrac{\nu\tau}{2}) \tag{2.3}$$

Overdispersion is controlled by (ν, τ) the two hyperparameters of the distribution of γ_g. With the parametrization shown in equation (2.3), mean, variance and coefficient of variation

are

$$E(\gamma_g) = \frac{\nu}{\nu-2}\tau \quad Var(\gamma_g) = (\frac{\nu}{\nu-2}\tau)^2\frac{2}{\nu-4} \quad CV(\gamma_g) = \sqrt{\frac{2}{\nu-4}}$$

so the parameter ν controls the amount of shrinkage around the mean. Parameter τ is related to the location of the distribution, there is no close form for the quantiles of IG distribution but is possible to compute them numerically. Figure 2.2 show the relationship between τ and selected quantiles, for different values of ν. The plots shows that median value is mostly affected by τ, and the largest effect of ν value occurs in the right tail of the distribution. In summary, τ

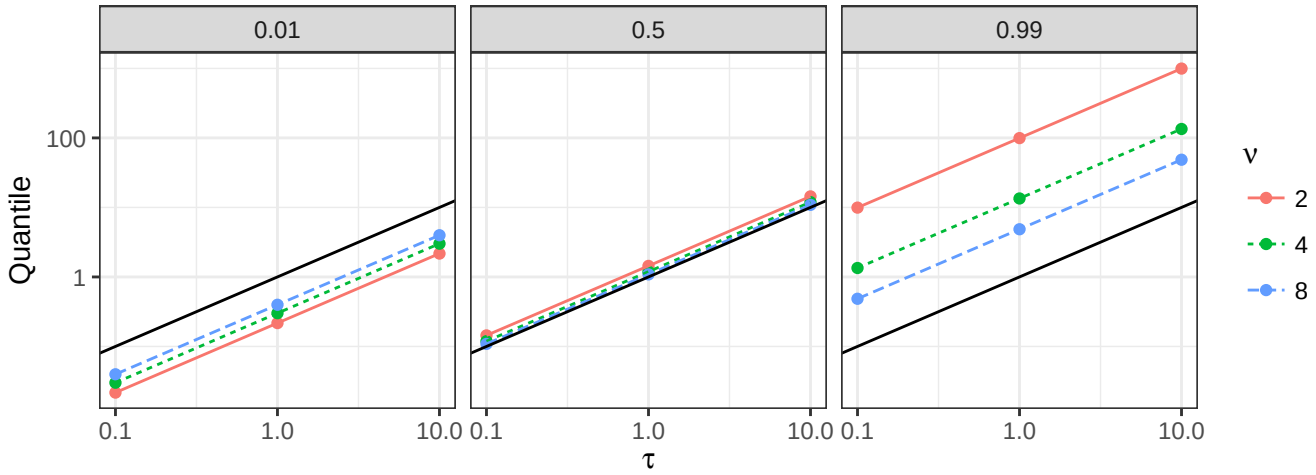

Figure 2.2: Quantiles of $IG(\frac{\nu}{2}, \frac{\nu\tau}{2})$ against τ. Facets correspond to quantiles, color and type of lines corresponds to ν value. Black line is the $y = x$ line.

controls the central location of the overdispersion across all genes, while ν controls the amount of shrinkage around τ and the right tail probability. For instance, if all genes would present the same overdispersion level, τ should be close to that value and $\nu \to \infty$ (or at least be large) , when most gene show no overdispersion but there are few genes with high overdispersion, both ν and τ should be small.

2.3.3 Allele effect (Δ_g)

Another important characteristic present in ASE experimental data is common to observe a higher transcription for one of the alleles on average across all genes, due to the positive

bias towards the reference allele mentioned in Section 2.2 (see Figure 2.1). These systematic difference among alleles are not of interest as the goal is to identify genes showing differences among allele larger than what is explained by systematic factors.

We define the *allele effect* to be the difference between alleles that is not due to bias, i.e.,

$$\Delta_g = \beta_{g2} - \theta_2. \tag{2.4}$$

We consider the overall mean across all genes, θ_2, as a measure of the systematic difference among alleles commonly due to bias towards the reference allele.

In order to obtain inference about the gene-specific regression parameters, the posterior mean and variance from the MCMC samples can be used to create a normal approximation of its posterior distribution (Landau et al., 2016). A similar strategy could be used to obtain credible intervals for the allele effect, Δ_g, in this case the posterior mean and variance of Δ_g are

$$\begin{aligned} E(\Delta_g|y) &= E(\beta_{g2}|y) - E(\theta_2|y) \\ Var(\Delta_g|y) &= Var(\beta_{g2}|y) + Var(\theta_2|y) - 2Cov(\beta_{g2}, \theta_2|y) \end{aligned}$$

A problem with this approach is that `fbseq` package provides posterior means and standard deviations for gene-specific parameters and full MCMC samples for all hyperparameters and only a few gene-specific parameters, so there is no information to compute $Cov(\beta_{g2}, \theta_2|y)$. However, the variability of the hyperparameters is negligible compared to gene-specific parameters, since the amount of information directly relevant for θ_2 is really large. This implies that $Var(\theta_2|y) \approx 0$ and $Cov(\beta_{g2}, \theta_2|y) \approx 0$.

Therefore, is possible to approximate $Var(\Delta_g|y) \approx Var(\beta_{g2}|y)$, to obtain a normal approximation of the posterior of the allele effect and then compute credible intervals for Δ_g. We show this approximation is reasonable later in Section 2.4

2.3.4 Differentially expressed alleles detection

The main goal of the proposed model is to identify genes with differentially expressed alleles (DEA). We approach this goal by using the posterior probability of DEA for each gene. A gene is DEA when one allele show an increase in the expression level, $|\Delta_g| \geq c$, where c represents a

threshold that must be adapted to specific applications or experiments. Here we follow Lithio and Nettleton (2015) in setting as the DEA threshold a 25% increase in the expression level, i.e., $c = log(1.25)/2$.

We want to detect genes with high posterior probability of presenting DEA. However, if the posterior distribution for Δ_g is too diffuse, for instance because is gene with very overdispersed counts, then $P(|\Delta_g| \geq c|y)$ is going be very large even when is no clear that gene is DEA. To avoid this problem, we first denote as H_{0g} the event that the gene g does not present DEA, which acts as null hypothesis. Then, we compute the null hypothesis posterior probability as follows

$$P^*(H_{0g}|y) = \min\{P(\Delta_g < c), P(\Delta_g > -c)\}$$

where the subscript distinguish for the regular probability and stands for *corrected*, this correction is proposed by Van De Wiel et al. (2013) and use it in Lithio and Nettleton (2015).

Next, a decision rule based on the posterior probability of DEA is needed to finally determine which genes present are flagged as DEA. One alternative is to use a Bayesian FDR, computed as the average of smallest $P^*(H_0|y)$ (Ventrucci et al., 2011; Van De Wiel et al., 2013). However, since we use a fully hierarchical Bayesian model and the null hypothesis has a positive probability multiplicity corrections are not needed (Muller et al., 2006; Bar et al., 2014).

Here we follow a more traditional Bayesian approach to derive a decision rule, i.e., choose an optimal rule in the sense that minimizes an expected loss function. Let d_g be an indicator that gene g has DEA feature, $h_g = P^*(H_{0g}|y)$, and consider the following expected loss

$$L(d, y) = q\sum_g d_g(1 - h_g) + \sum_g (1 - d_g)h_g = qFD + FN$$

where FD and FN are the posterior expected false discoveries total and false negative total respectively, and q is the relative cost associated to FD. Müller et al. (2004) shows the optimal rule that minimize $L(d, y)$ is

$$d_g = I\left(h_g \leq \frac{1}{q+1}\right).$$

Setting $q = 19$ we would declare as DEA every gene with posterior probability of H_{0g} lower than 5%.

2.3.5 Bayesian inference

Posterior distribution for gene-specific parameters is learned by borrowing information across all genes trough its hyperparameters, since hyperparameters posterior distribution uses complete data set. This fact helps to deal with the small sample size inherent in RNA-seq experiments.

Equation (2.5) shows the prior distributions for the hyperparameters in the model, we use normal distribution for regression means and uniform distribution for regression variances. Parameters controlling overdispersion effect have uniform prior, ν, and gamma prior, τ. The values of the parameters of these priors are set to obtain diffuse distributions. As number of genes is very large, there is lot of information about the hyperparameters in the data and the hyperparameter prior will not have a large impact in the gene-specific parameters (Ghosh et al., 2006)

$$\begin{aligned} \theta_k &\overset{\text{ind}}{\sim} \text{N}(0, c_k) \\ \sigma_k &\overset{\text{ind}}{\sim} \text{Unif}(0, s_k) \\ \nu &\sim \text{Unif}(0, d) \\ \tau &\sim Ga(a, b) \end{aligned} \tag{2.5}$$

Uniform priors for variance parameters, as σ_k^2, in hierarchical models have been propose as a good non-informative alternative (Gelman, 2006), a similar argument also applies for a variance related parameter like ν. Normal prior for location parameters θ_k is widely used choice (Gelman et al., 2013), it can be weakly informative maintaining conditional conjugacy. Similarly the gamma prior for a location-related parameter τ represents a good balance between computation convenience and being weakly informative.

Models where gene-specific parameters have fully specified distributions, i.e. non-hierarchical models, can be estimated using MCMC methods (León-Novelo et al., 2014). However, fully Bayesian inference of the hierarchical models is computationally demanding, since the number

of groups (or genes) is big. Usually, approximations like empirical Bayes (Niemi et al., 2015) or integrated nested Laplace approximation (Van De Wiel et al., 2013) are used to obtain inference results.

Parallel computing is a way to tackle down the computational intractability of this models, Landau and Niemi (2016) propose to use graphics processing units (GPU) to take advantage of the embarrassingly parallel nature of the MCMC algorithms in conditionally independent hierarchical models. We use `fbseq` package (Landau and Niemi, 2016) to obtain fully Bayesian inference of the proposed model.

2.4 Simulation Study

A simulation study is performed to explore how the model captures several characteristics of interest in the data. In this Section, we describe the data sets simulation scenarios, the analysis of each simulated data, and finally simulation study results are presented.

2.4.1 Model to simulate data

In order to obtain simulated data sets close to the specific data we have, we fit an initial model to use it as base for simulate new data, here we describe this initial model.

The specific dataset we use later in Section 4.6 as application example, has 8 allele-specific observations per gene, corresponding to 4 biological replicates of a single hybrid genotype distributed in two blocks. As an initial step we use a negative binomial model,

$$Y_{gn} \overset{\text{ind}}{\sim} NB(e^{h^*_{gn}+x_n^\top \beta_g}, \phi_g),$$

where h^*_{gn} are normalization factors and ϕ_g control the overdispersion. Model matrix X is formed by $x_n^\top$ on its rows, it has 8 rows and is the same for all genes. In order to ensure independence among regression coefficients, we use a zero-sum parametrization for X, the

response and design matrix for one gene g are as follows:

$$Y_g = \begin{bmatrix} Y_{g11} \\ Y_{g12} \\ Y_{g13} \\ Y_{g14} \\ Y_{g21} \\ Y_{g22} \\ Y_{g23} \\ Y_{g24} \end{bmatrix} \quad X = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & -1 & 1 & 1 & 0 \\ 1 & 1 & 1 & -1 & 0 \\ 1 & -1 & 1 & -1 & 0 \\ 1 & 1 & -1 & 0 & 1 \\ 1 & -1 & -1 & 0 & 1 \\ 1 & 1 & -1 & 0 & -1 \\ 1 & -1 & -1 & 0 & -1 \end{bmatrix} \tag{2.6}$$

This particular model matrix X implies β_{g1} corresponds to the intercept and β_{g2} to the half allele difference, as in the general model presented previously. Here we include a column to capture the difference between the two blocks, associated with coefficient β_{g3}. Also, two columns for block and replicate interaction, (β_{4g}, β_{5g}) are included, which represent the half difference between replicates within each block, they capture the effect of the common biological sample of each pair of measures. Note that usually the set of effects related with grouping factors as biological samples, share a common variance, while the model proposed here allows $\sigma_4 \neq \sigma_5$ which encompasses the common variance case.

We obtain point estimates of the gene-specific regression coefficients and gene-specific overdispersion parameters using `edgeR` (Robinson et al., 2010), the program also provide values of normalization factors h^*_{gn} based on the method proposed by Anders and Huber (2010). These point estimates and normalization values are used to obtain the simulated data sets.

2.4.2 Simulation scenarios

A simulation scenario is defined by four simulation design parameters: (w, s, p, T), latin letters are used for these design parameters to differentiate them from the unknown model parameters. We investigate the impact of sparsity (w) and strength (s) of the allele effect, bias toward reference allele (p), and overdispersion effects (T). Table 2.1 shows together all values for the design parameters, in total there are 24 scenarios as a full factorial combination of the

design parameters values. Each scenario represents ASE count for $G = 5000$ total genes with 8 observations per gene, and is replicated 2 times.

Table 2.1: Simulation study design parameter values

Description	Sparsity	Strength	Bias	Overdispersion
Parameter	w	s	p	T
Values	$(.5, .95)$	$(1, 1.8)$	$(1, .5)$	$(0.25, 1, 4)$

The input for each scenario is formed by the estimates $(h^*_{gn}, \hat{\beta}_{g1}, \hat{\beta}_{g2}, \hat{\phi}_g)$, kept from NB model described above.

The first step to create a simulated dataset constitute a random selection of genes. All genes are split in two groups

$$S1 : |\hat{\beta}_{g2}| \leq c \qquad S2 : |\hat{\beta}_{g2}| > c$$

and a stratified random sample with replacement of model coefficients, with w proportion from $S1$ and $(1 - w)$ proportion from $S2$. Then, design parameter w controls the amount of sparsity in the allele effects, two sparsity scenarios are considered $w = (0.5, 0.95)$.

The second step is to compute the gene-specific effects. Allele effects are computed as $sx_a\hat{\beta}_{g2}$, where x_a takes value 1 for the reference allele and -1 in the non-reference allele. The design parameter s controls the signal strength, we set $s = (1, 1.8)$ as weak and strong signal cases respectively. Meanwhile, overdispersion effects are computed as $T\hat{\phi}_g$, we use 3 scenarios determined by the value of $T = (.25, 1, 4)$. Then, for the selected genes and the computed gene specific effects we simulate 8 counts per gene as

$$Y^*_{gn} \overset{\text{ind}}{\sim} NB(e^{\bar{h}+\hat{\beta}_{g1}+sx_a\hat{\beta}_{g2}}, T\hat{\phi}_g)$$

where $\bar{h}$ is the mean of the computed normalization factors.

Reference allele bias. Finally we induce some bias toward reference allele by setting

$$\begin{cases} Y_{gn} = Y^*_{gn} & \text{if } x_a = 1 \\ Y_{gn} \sim Bin(Y^*_{gn}, p) & \text{if } x_a = -1 \end{cases}$$

this implies that for allele-specific counts from reference allele ($x_a = 1$) we maintain the first negative binomial simulated value, but in the non-reference allele-specific counts the initial

simulation serve as the size parameter in a binomial distribution. Design parameter p is the probability of actually assigning one short read to the non-reference allele, so on average $(1-p)$ non-reference reads are lost. We set two values for the design parameter that control reference allele bias, $p = (1, .5)$, corresponding to the case with no bias and a case where 50% of the non-reference reads are lost on average, respectively. This last step implies the counts from non-reference allele will be smaller on average than reference allele counts, to see this more clearly we can integrate the two final steps together as

$$\begin{cases} Y_{gn} \overset{\text{ind}}{\sim} NB(e^{\bar{h}+\hat{\beta}_{g1}+s\hat{\beta}_{g2}}, T\hat{\phi}_g) & \text{if } x_a = 1 \\ Y_{gn} \overset{\text{ind}}{\sim} NB(e^{\bar{h}+\hat{\beta}_{g1}-s\hat{\beta}_{g2}+\log(p)}, T\hat{\phi}_g) & \text{if } x_a = -1 \end{cases}$$

this implies that the mean of β_{g2} coefficients, θ_2, should be close to $-\log(p)/2$ since β_{g2} captures the gene-specific half difference among the two alleles and $\log(p)$ represent the allele between alleles averaging all genes.

We perform 10 analyses on each simulated dataset. First, every data is analyzed using data model 2.1 with each of the four shrinkage distributions from equation (2.2) for β_{g2}, and also a normal distribution for both regression parameters. The main reason for this is to assess the impact of the hierarchical distribution of the regression coefficients on the posterior inference. Additionally, for each hierarchical model we also fit its non-hierarchical version, i.e., fixing hyperparameters at $\theta_k = 0$, $\sigma_k^2 = 3^2$, $\tau = .1$, $\nu = 1$.

We run 3 MCMC chains with 40000 iterations for hierarchical models, and set thinning value of 5 in Cauchy and horseshoe cases. Still, horseshoe distribution shows lack of convergence in many scenarios for θ_k parameters, therefore we do not present horseshoe distribution simulation results nor consider it for the real data analysis. Hahn and He (2016) have recently pointed out that horseshoe distribution may have poor mixing in high-dimensional problems, and propose to use an elliptical slice sampler to improve it. Non-hierarchical models inference is obtained with 3 MCMC chains with 20000 and no thinning.

All computations are done in `R`. MCMC results are obtained using `fbseq` Landau () package, convergence is assessed using potential scale reduction factor statistic. Other data management and plots we use `dplyr` Wickham and Francois (2016), `ggplot` Wickham (2009), `plotROC` Sachs

(2016), `tidyr` Wickham (2017).

2.4.3 Simulation results

We construct receiver operating characteristic (ROC) curves for each simulation scenario, in order to describe model performance to detect genes with differential allele expression. The posterior probability of having an allele effect outside the null region, $P^*(\Delta_g \geq c|y)$, where $c = \log(1.25)/2$, is used as a continuous score to compute the ROC curves.

Figure 2.3 presents the ROC curves for simulation scenarios in which reference allele bias is present ($p = 0.5$). Each panel corresponds to a particular scenario combining the other 3 simulation design parameters, overdispersion (T) is represented in row faces while strength (s) and sparsity (w) of the allele effects determine the column facets. The dashed lines represent the non-hierarchical models and continuous lines represent the hierarchical models. Finally the ROC curve color indicates the hierarchical distribution for the regression coefficients.

There are two effects that applied to every model. As we might expect, increasing the signal strength and decreasing the overdispersion level produce better detection rates. Another overall effect is the fact that non-hierarchical models fail into correct the bias towards reference allele, and they present worse detection rates in every scenario compared with the hierarchical counterparts. An initial simulation study, included in appendix A, shows similar results, comparing Figures A.3 and A.5 is clear that non-hierarchical models are more affected by overdispersion, sparsity and bias. Then, we describe the remaining effects results only for hierarchical models.

Among hierarchical models, Figure 2.3 suggests hierarchical distribution for the regression parameters have some effects on model performance in many scenarios. Cauchy distributions shows larger true positive fraction (TPR) than the rest when overdispersion level is high and in the highly sparse scenarios when 95% of the genes does not have DEA. On the other hand, using a normal distribution have the worst results in terms of ROC curves, in this case the more clear negative effect for the normal model is the sparsity level. Performance of Laplace and t distributions appear to be somewhat in the middle, between Cauchy and normal.

ROC curves can be summarize computing the area under the ROC curve (AUC), a perfect

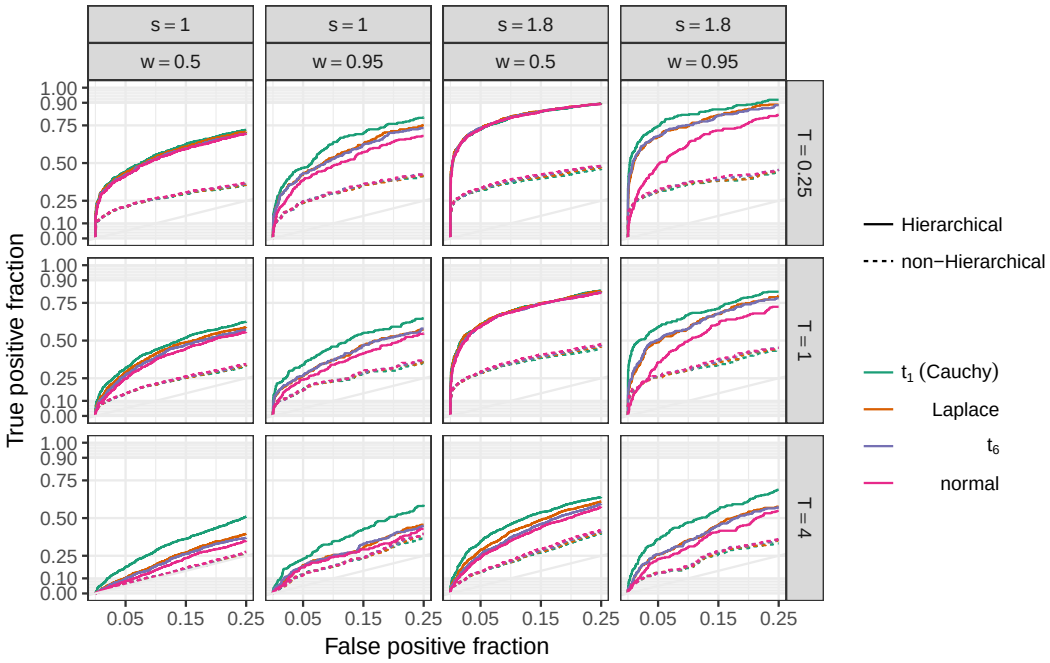

Figure 2.3: ROC curves for scenarios with reference allele bias present ($p = 0.5$). Row facets corresponds to overdispersion level (T), while column facets combine signal strength and sparsity (s,w). Line color indicates the hierarchical distribution. Hierarchical models are plotted with continuous lines and dashed lines correspond to non-hierarchical models.

detection rate would have AUC value of 1. Typically, only the region with small false positive fraction (FPR) is of interest, we compute a partial AUC as the area under the ROC curve but only over the region where FPR is less than 10%. Partial AUC results are presented in two separate plots, Figure 2.4 shows the effect of simulation design parameter within each hierarchical distribution, meanwhile Figure 2.5 compares partial AUC among hierarchical distribution for the regression model parameters.

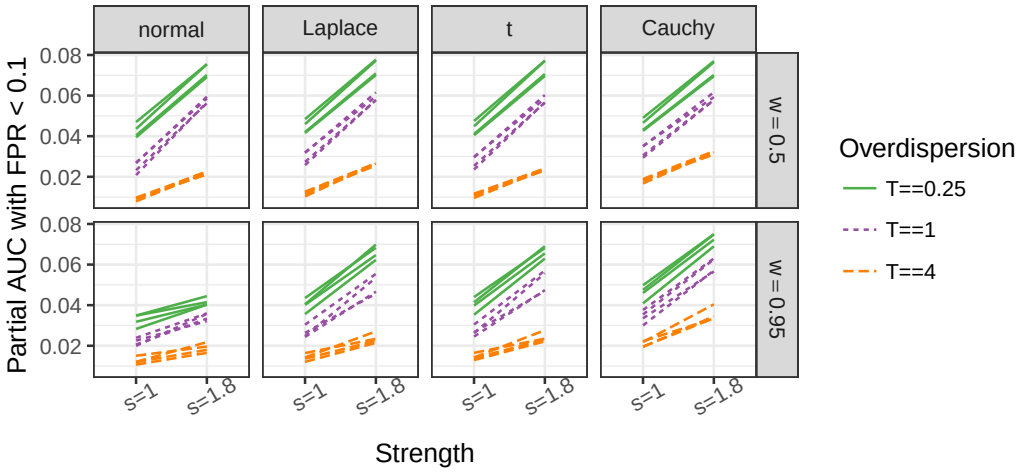

Figure 2.4: Partial area under ROC curve (AUC), over region with false positive rate lower than 10%. Facets represent the hierarchical distribution used in the model, and row facets represent sparsity level. Each line corresponds to a scenario, color and type of the line indicate the overdispersion level (T)

Figure 2.4 shows partial AUC measure results against signal strength. Column facets corresponds to hierarchical distributions, row facets corresponds to sparsity level, and the color and type of the lines corresponds to overdispersion level.

As it was comment in Figure 2.3 overdispersion level and signal strength have the largest impact on the signal detection performance. Reference allele bias does not seem to have an impact in signal detection within hierarchical models, even when half of the non-reference allele reads are lost we still be able to identify genes with allele effects correctly. There might be some interaction effects among the simulation design factors, for instance, signal strength gain

in performance is larger when overdispersion is low or moderate, also in the normal case, signal strength does not show a big change when sparsity is high in any overdispersion level.

Figure 2.5 shows the partial AUC measure results against hierarchical distributions. The facets combine the signal strength level (columns) and sparsity level (rows), the color and type of the lines indicate the overdispersion level. Each line corresponds to one simulation scenario, we want to compare the β_{g2} hierarchical distribution effect on the model performance.

In every case, Cauchy present better results than the rest while normal distribution has the poorest results. This is the same result than in Figure 2.3, it is not surprising Cauchy distribution having better performance in this case, Cauchy accommodates a lot of probability mass close to zero and its heavy tails can capture the genes with real effects. Normal distribution has light tails which produce a shrinkage effect on all parameters including the ones that are further apart from 0. Perhaps, the comparison among Cauchy and t distributions is not so common, after all, Cauchy distribution a special case of t, the results suggest the degrees of freedom parameter have an impact on how the model borrows information across genes.

Another relevant aspect of the model performance is the ability to discover the genes that are truly differentially expressed. We declare a gene as having allele differentially expressed when $P(|\Delta_g| \geq c|y) < 0.05$ which correspond to minimize the expected loss $19FD + FN$.

Figure 2.6 shows the proportion of false discoveries (FDR) against the proportion of discoveries Dp, i.e., the ratio of total discoveries over the total number of differentially expressed genes. Ideally, we would like to discover all differentially expressed genes with no false discoveries. We only plot scenarios with the same overdispersion level that observed data ($T = 1$).

Only looking at the right panel, with strong signals, the proportion of discoveries drops from somewhat higher than 40% when no sparsity is present to be close to zero when $w = 0.95$ for normal, t and Laplace distributions. Meanwhile, Cauchy maintains the proportion of discoveries around 30% in all cases, and showing small false discovery rates. In the left panel, weak signals, a similar effect of the sparsity is observed, but there are some cases with large FDR when using Cauchy distribution. It should be kept in mind that in this case the proportion could be variable since its denominator is very small.

We finish this Section showing how the proposed model captures the bias towards reference

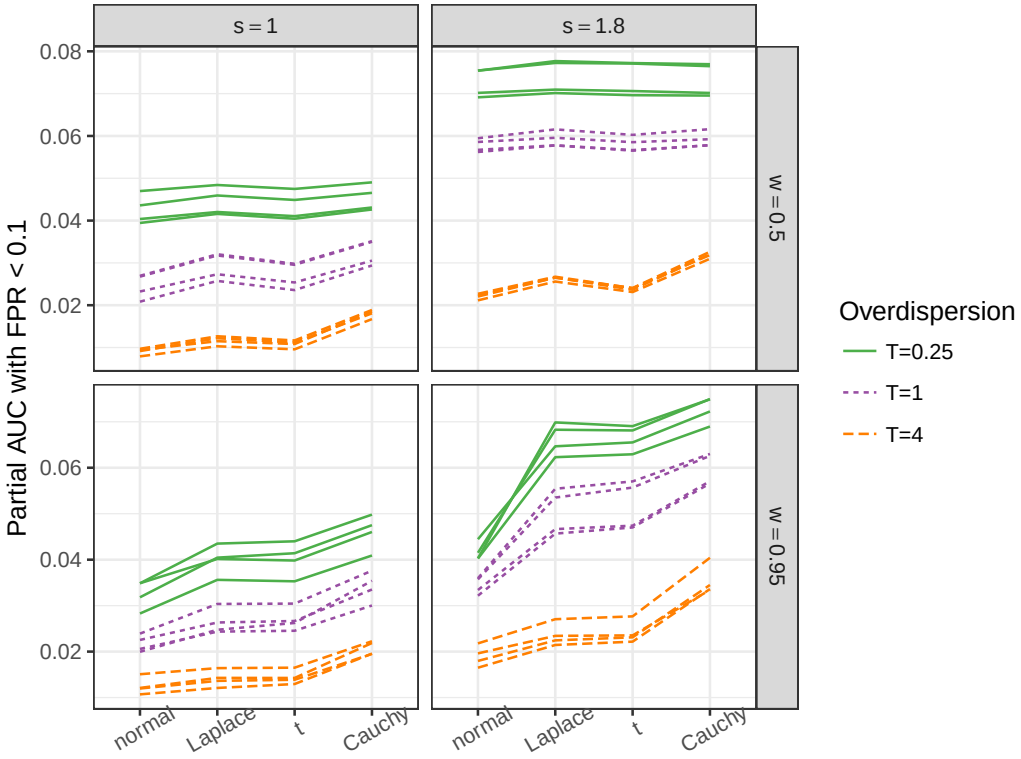

Figure 2.5: Partial area under ROC curve (AUC), over region with false positive rate lower than 10%. Facets represent the overdispersion level (T). Each line corresponds to a scenario, color indicates the signal strength and the line type represent the sparsity.

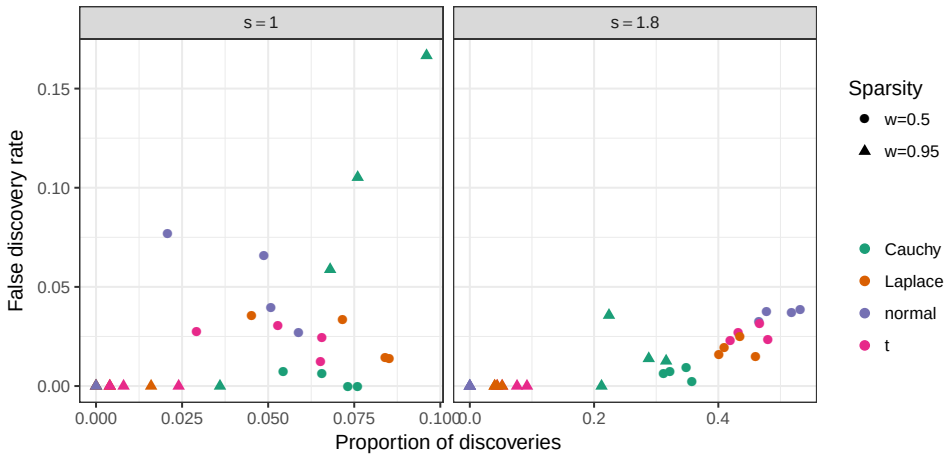

Figure 2.6: Scatter plot of false discovery rate against proportion of discoveries. Color of points indicates the hierarchical distribution for regression parameters and shape indicates sparsity level.

allele, and how well the posterior variance of the allele effect, Δ_g, is approximated by the variance of the regression coefficient β_{g2}.

Above we mentioned that parameter θ_2 should capture half of the bias in log scale, i.e., we expect i.e. $E(\theta_2|y) \approx -\log(p)/2$. Figure 2.7 shows a scatter plot of $E(\theta_2|Y)$ against $\log(p)$. The plot suggests posterior expectation of θ_2 captures the bias towards the reference allele, is possible to use it as an estimate of the bias and remove it when making inference about the allele effect.

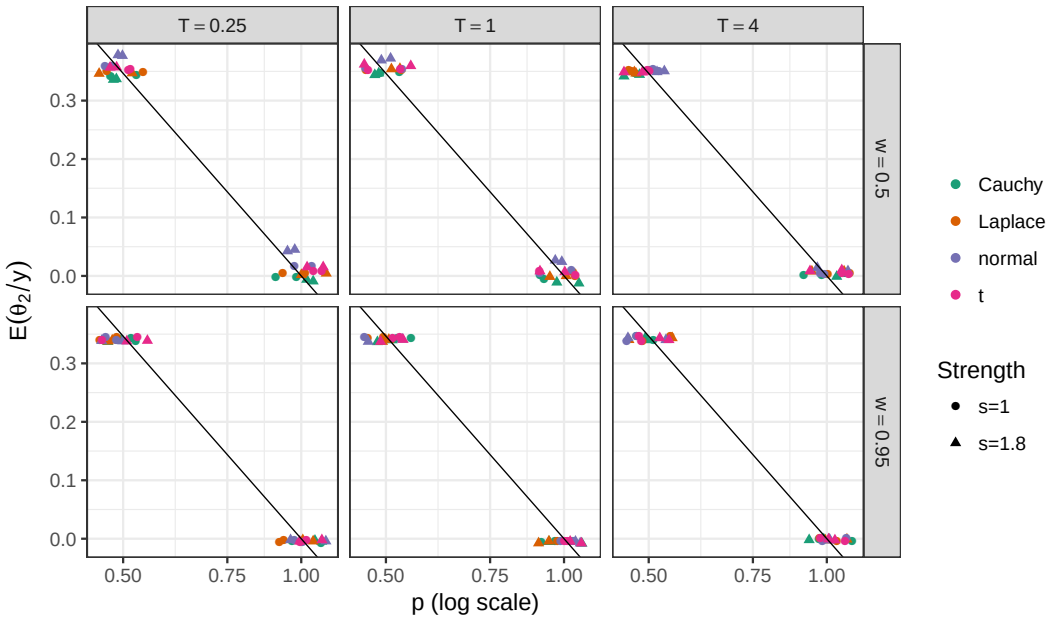

Figure 2.7: Scatter plot of θ_2 posterior mean against $\log(p)$ parameter. Row facets represents sparsity and column facets the overdispersion level. The line corresponds to $y = -\frac{x}{2}$ line.

In Section 2.3 we stated that $Var(\Delta_g|y) \approx Var(\beta_{g2}|y)$, we can test the approximation from a few genes having the complete MCMC samples. Figure 2.8 presents scatter plots of the variance of the allele effect against the variance of the regression coefficient β_{g2}, the facets represents the hierarchical distribution used in the model and color of points represent the overdispersion level. There is a close relationship among the two plotted variances, suggesting the approximation $Var(\Delta_g|y) \approx Var(\beta_{g2}|y)$ is reasonable.

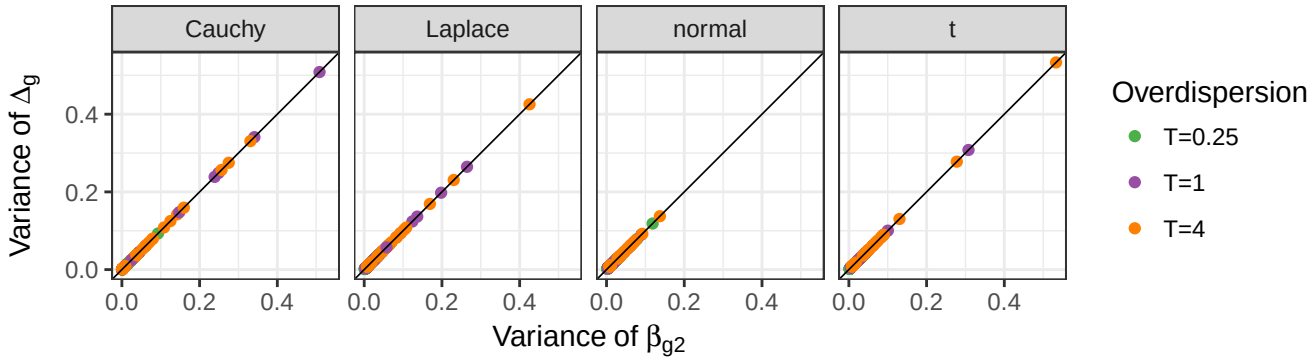

Figure 2.8: Scatter plot of variance of allele effect against variance of regression coefficient. Facets correspond to the hierarchical distribution and color indicates the overdipersion level.

2.5 ASE in maize experiment

In this Section we apply the methods described in Section 2.3 to a RNA-seq data set with allele-specific counts from maize that constitute a portion of the experimental data obtained by Paschold et al. (2012). Data set includes four replicate samples of a hybrid genotype (B73xMo17) distributed in 2 flow cell blocks and two allele count measures per sample. RNA-seq transcript abundance count information for 39656 genes is obtained using Iliumna® technology. However, many of them have little or no ASE information. To avoid genes with extremely low observed expression, only use genes were the average of allele-specific counts is bigger than 1. The resulting data set corresponds to the ASE counts of 16380 genes, which is 41% of the total. For those genes there is a count measuring the expression of the B73 allele and another for the Mo17 allele. Genome B73 is the reference allele since is the genotype that has been sequenced (Schnable et al., 2009).

An initial exploration of this data was presented in Section 2.2 to illustrate the main features of an ASE data set. The specific model matrix we use is the same presented in (2.6) used to create a model to simulate data. Gene-specific intercept is normally distributed while the rest of regression parameters are Cauchy distributed. The choice of $\beta_{gk} \sim Ca(\theta_k, \sigma_k)$ is based on the results from simulation study, the models using Cauchy hierarchical distribution results in better partial AUC measures in particular in sparse cases or cases with large overdispersion

levels. A more practical approach could be to use a normal distribution as initial step, and then based on histograms of β_{kg} posterior means determine if more shrinkage is needed.

In this Section we present the main results from the analysis, we start with some remarks about the posterior inference relative to the hyperparameters of the model, and after that we focus on the results relative to gene-specific allele effects. We present credible intervals for Δ_g and identify genes differentially expressed between alleles.

Table 2.2 present posterior summaries for all hyperparameter in the model. Posterior means and credible interval for (ν, τ) suggest most genes show very little or none overdispersion present, but there are a few genes with large overdispersion effects.

Table 2.2: Hyperparameter posterior summaries for ASE counts of B73xMo17 data

param	mean	Credible.Interval.95
ν	3.6	(3 , 4.3)
τ	0.0023	(0.0019 , 0.0028)
θ_1	2.4	(2.4 , 2.4)
θ_2	0.12	(0.12 , 0.13)
θ_3	-0.025	(-0.029 , -0.021)
θ_4	-0.026	(-0.029 , -0.024)
θ_5	0.002	(0.00015 , 0.0038)
σ_1^2	1.7	(1.7 , 1.8)
σ_2^2	0.012	(0.011 , 0.013)
σ_3^2	0.013	(0.013 , 0.014)
σ_4^2	0.0011	(0.00094 , 0.0012)
σ_5^2	0.000015	(0.0000095 , 0.000023)

Posterior mean of θ_2 is positive representing the bias towards reads from B73 allele. The results suggest that expression from Mo17 allele is only 78% of the expression count from allele B73 on average across all genes. In other words, 1 out of 5 reads from Mo17 is lost presumably because is compared with a different genome.

A final comment from Table 2.2 is related to the variances of common biological sample effects, i.e., σ_4 and σ_5. We mentioned in Section 2.4 that this type of effects, included due to grouping factors, are usually amused to have equal variances. The results suggest in this example those variances are different with $\sigma_4^2 > \sigma_5^2$ by a factor of 100.

Left panel Figure 2.9 is a volcano plots, shows the probability of a gene is differentially

expressed against the allele effect posterior mean (Δ_g). As expected, when $\Delta_g \in (-c, c)$ its probability of being differentially expressed is less than 50% and genes with allele effects large have a probability of begin differentially expressed close to 1. There also some genes with relatively small allele effect but with large probability.

The right panel of Figure 2.9 presents 95% credible intervals of allele effects against gene expression with color highlighting genes with differentially expressed alleles. The log-expression is computed as $\beta_{g1} + h_n$, i.e. the posterior mean of the gene-specific intercept plus the offset value. Genes with large expression show smaller allele effects and shorter credible intervals than genes with low expression. There are some genes flagged as differentially expressed among alleles with very low expression level.

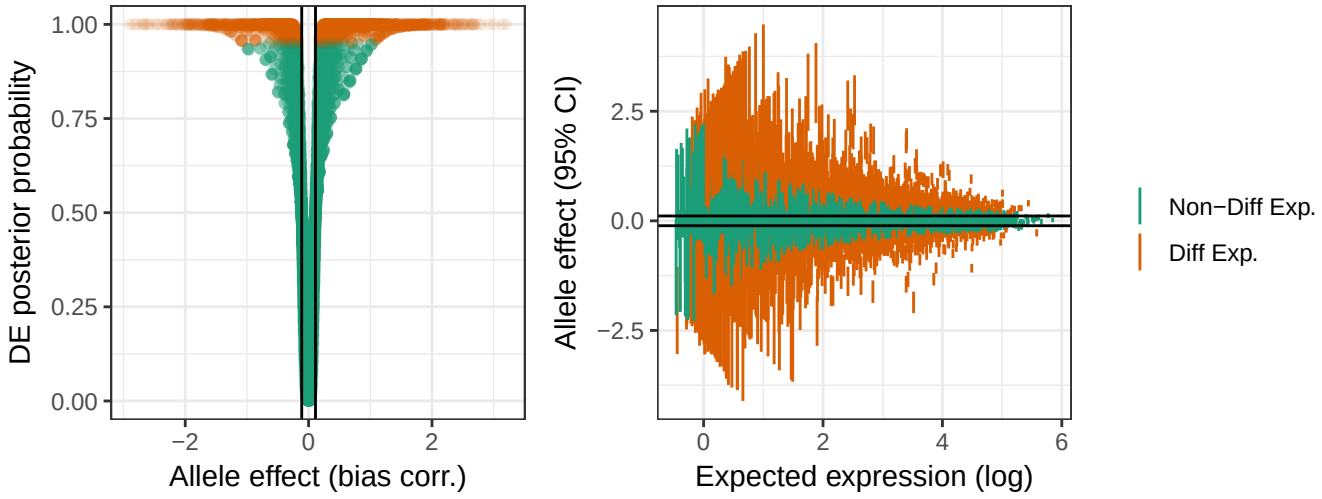

Figure 2.9: Allele effects for ASE counts of B73×Mo17 hybrid data. Left: Scatter plot of probability of differential expression against allele effect. Right: 95% credible intervals of allele effect against overall gene expression. In both panels, color indicates if the gene is declared as differentially expressed or not.

Results indicate that 17% of the genes shows allele differential expression. This might seems a small proportion relative to previously reported for this data set (Paschold et al., 2012), however we define a *region* of non-differential expression instead of a point value. The

expected number of false discoveries can be obtained simply as $\sum_g P(|\Delta_g| < c|y)$ among all genes declare with differential expression. We expect 23 genes to be false discoveries.

2.6 Discussion

Allele-specific expression refers to a transcript abundance count associated with each gene copy (allele). We propose a hierarchical overdispersed count regression model to deal with ASE data from a single hybrid variety.

The proposed method handles the main features of ASE data correctly. The model has an internal measure of reference allele bias which can be used to obtain gene allele effects free from that bias. Credible interval for the allele effects and a strategy to flag genes with large allele effects are provided. Sparsity pattern in the allele effects are managed by using shrinkage distributions in the gene-specific regression parameters hierarchical distribution.

Model inference is performed in a fully Bayesian fashion, this is not commonly used due to computational restrictions. MCMC algorithm to sample from the proposed model is embarrassingly parallel when updating the gene-specific parameters. A parallel strategy computing is then used for computational efficiency. We test the model's performance in simulated and experimental data.

Simulations suggested that overdispersion level and the signal strength level has a large impact on the model performance. Among four shrinkage distributions, the better performance results in terms of signals detection were showed by Cauchy distribution. Cauchy is informative enough to produce information sharing across genes and at the same time is flexible enough (due to the heavy tails) to accommodate large true signals. There are performance gains in learning the hierarchical distributions of gene-specific parameter from data. Non-hierarchical models performance is more heavily affected by to overdispersion and sparsity level, more importantly, they cannot accommodate the reference bias that ASE data usually show.

In addition to work for modeling ASE data from a single hybrid variety, the method proposed in the paper could serve as the base for a larger model including more varieties (other genomes) and total RNA-seq expression. This will allow us to obtain more relevant contrast other than differential expression among alleles. In particular, we could directly study the re-

lationship among hybrid vigor and allelic imbalance. Simulations suggested some interactions among overdispersion, signal strength and sparsity, another line of future work is to continue exploring these interactions.

CHAPTER 3. Bayesian hierarchical model to analyze heterosis and allelic imbalance relationship

Abstract

Heterosis, also known as hybrid vigor, has been exploited in agricultural production for over 100 years. Interestingly, the reasons for the occurrence of hybrid vigor are still considered an open problem. The uneven expression of alleles might be related with the increased ability of adaptation of hybrids, so it might be related to the occurrence of heterosis. The main focus of this paper is to jointly model heterotic patterns and allelic imbalance, and develop association measures between them.

A Poisson-lognormal mixture model is proposed to analyze RNA-seq transcript abundance counts, including total RNA-seq and ASE counts in the same model. There are five gene expression patterns of interest, three of those patterns corresponds to mid-parent and extreme-parent heterosis and the other two are related to the differential expression of the alleles in hybrids plants. Two gene-specific measures relating heterotic gene expression patterns and allelic differential expression patterns are proposed: linear correlation coefficient and the conditional probability of heterosis given allelic imbalance are the.

Results from a case of study using RNA-seq data set suggest that genes with allelic-specific expression available present low probability of showing extreme-parent heterosis, and genes with large differential expression between alleles display gene expression patterns consistent with mid-parent heterosis explained by the dominance hypothesis.

3.1 Introduction

In plant breeding, it is common that hybrid lines show improvements in several phenotype traits compared with its parent lines. The effect that a heterozygous hybrid is better compared with the average of its homozygous parents is called *heterosis*. (Schnable and Springer, 2013)

Heterosis, also known as hybrid vigor, has been exploited in agricultural production for over 100 years, there are studies documenting heterosis as early as Darwin (1876) and Shull (1908). Interestingly, the reasons for the occurrence of hybrid vigor are still considered an open problem (Hallauer et al., 2010). Several genetic models had been proposed as explanations for heterosis, the two more used are the dominance and overdominance hypothesis (Swanson-Wagner et al., 2006).

The dominance model is based on complementation of deleterious alleles, the hybrid display heterotic pattern when a deleterious allele in one inbred parental line is complemented with a superior allele from the other inbred parent. It has been hard to explain the cumulative empirical evidence of heterotic patterns only based on the dominance model. Overdominance model suggests the heterosis occurs due to allele interactions in the hybrid that create non-additive patterns when the hybrid progeny is compared with its parental lines. More recently a third family of reasons has been proposed, where heterosis can arise from epistatic interactions among alleles from different loci (Ryder et al., 2014).

One way to approach the analysis of hybrid vigor occurrence is to look for gene expression heterosis, i.e. identify individual genes presenting heterotic pattern expression. Ji et al. (2014) proposed an approach to identify genes showing mid, low or high parent heterosis using microarray data. Later Niemi et al. (2015) extend this approach to deal with RNA-seq data. In this paper, we continue this line of research but incorporating allele-specific expression into the models.

Allele-specific expression (ASE) refer to the gene expression of each copy of the gene. When applied to heterozygous individuals, Next generation sequencing technology makes possible, in some cases, to distinguish reads from different alleles. Allele-specific effects concern the additive and dominant nature of regulatory interactions between parental alleles (Bell et al., 2013).

Differential allele-specific expression has been related to non-additive patterns in total RNA-seq expression (Paschold et al., 2012). A possible reason is that the uneven expression of alleles might be related with the increased ability of adaptation of hybrids, so it might be related to the occurrence of heterosis. The main focus of this paper is to model jointly heterotic patterns and allelic imbalance, and develop association measures between them.

In the next section, a description of the main heterotic and allelic gene expression patterns is presented. Along with a maize experimental data set, which is used as motivation example in the rest of the paper. In Section 3.3, a hierarchical Poisson-lognormal mixture model for studying the relationship among gene expression heterosis and ASE is proposed. Results from the data analysis are presented in Section 3.4, and a discussion in Section 3.5.

3.2 Gene patterns and RNA-seq data

Before describe the model we describe the main characteristic of the data for which the model is relevant. As a motivating example we use gene expression data obtained in a maize experiment whose details are described later (Paschold et al., 2012). In addition, we illustrate the main gene expression patterns of interest.

3.2.1 Maize experimental data

We assume RNA-seq counts and ASE counts for three varieties are available, the three varieties are composed by an hybrid variety, BM, and the inbred parental lines, B and M, with a few biological samples per variety. Furthermore, we assume there are 3 types of RNA abundance obtained for each gene in every biological sample: the total abundance and the abundance for the two alleles. Total gene expression abundance is measure with the total RNA-seq count and the allele-specific abundance is measure with ASE counts for each allele.

Using upper case letters, B, BM, and M we refer to the genotypes or varieties for one parental line, the hybrid and the second parental line respectively. Meanwhile, lower case letters, b and m are use to indicate the expression count type as the ASE matching the varieties B and M respectively, while t indicates the total RNA-seq count.

Figure 3.1 presents boxplots of each count (in logs 2), row facets represent the sample variety, column facets represent the expression type, and the color represents the replicate. Several characteristics of the data can be described based on this Figure.

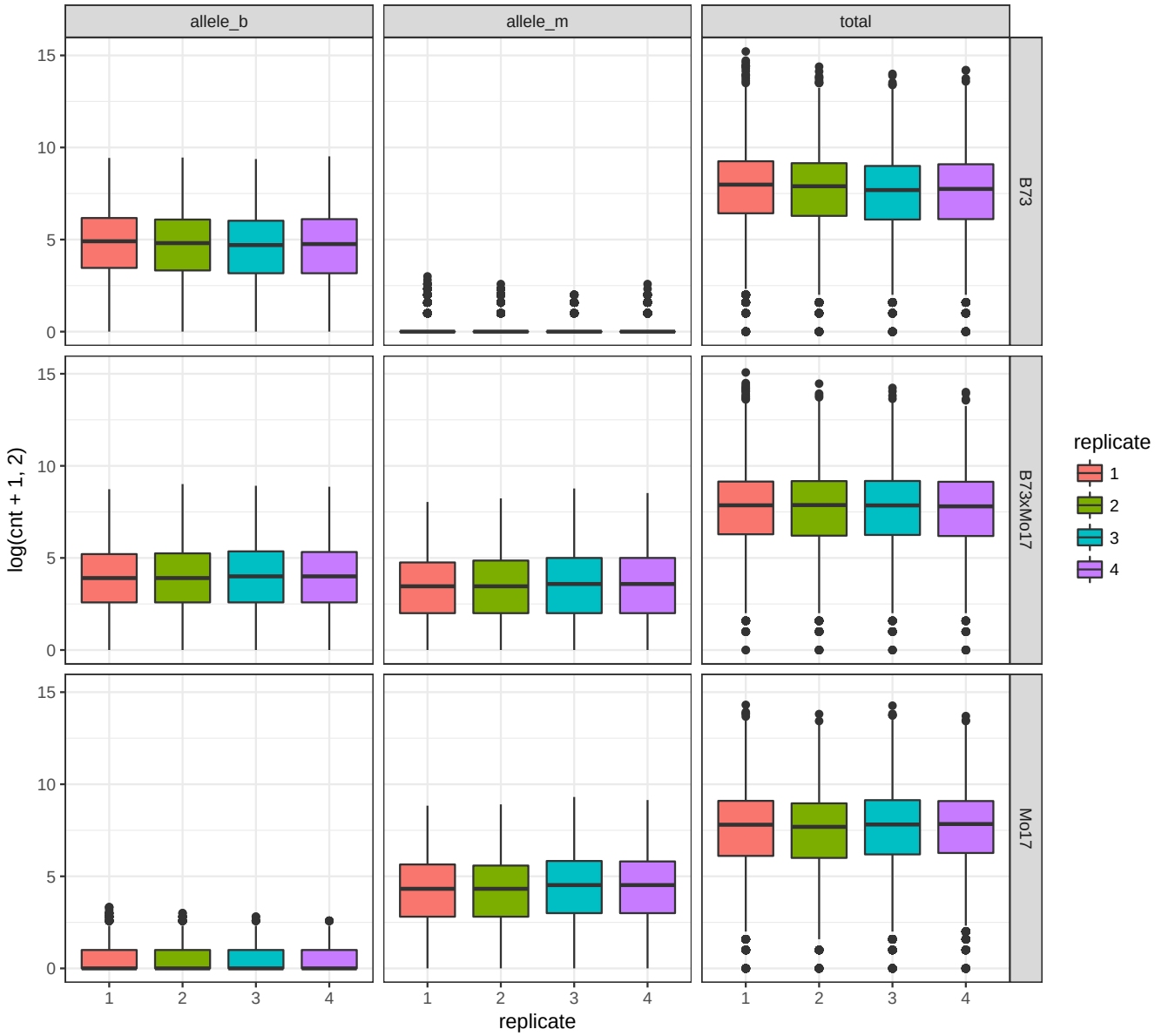

Figure 3.1: Boxplots of observed counts (in log2) in genes with ASE information available. Row facets represent the biological sample variety (B, BM, M), column facets represent the expression type (b,m,t), and the color represents the replicate.

Firstly, within each variety, each row facet in the plot, the total expression shows larger counts than the sum of both ASE. This happens since to be able to distinguish which allele

correspond a particular read, a single nucleotide polyformisms (SNP) need to be present in the in transcribed region. Therefore, for many of the reads is possible to assign it to a particular gene, but t is no possible to determine from which gene copy (allele) corresponds because there is no SNP. This can happen to all reads corresponding a gene, in that case, no ASE information is present for that gene. For instance, in the data from Paschold et al. (2012), genes with some ASE information represents only 40% of the genes with total RNA-seq abundance counts.

A second feature from Figure 3.1 can be observed by focusing in the hybrid variety (mid-row facets) and the allelic expression types (the first two columns). The overall abundance from allele b is larger than allele m, which is an expression of what is called reference allele bias. The hierarchical models for ASE in the hybrid proposed in chapter 2, use hyperparameter posterior expectations to remove this bias.

A third remark is that in some combinations of variety and allele the counts are really low. For instance, the second column panel from first row correspond to ASE counts of allele m from the B variety. Actually, all of those ASE counts should be equal to 0, since the B variety is an inbred, there should be no reads matching the genotype from variety M. It could be possible to use these counts as they provide information about false positives counts, and use it to correct in some way the other relevant counts. A similar pattern occur in the first column panel from the last row, this corresponds to ASE counts from variety M matching the genotype from variety B. However in this case, 50% of the genes show positive counts. This is presumably another consequence of the reference bias which constitute a first note of caution on how to use the false positive correction

Finally, there are 3 boxplots with the same color within each variety (facet row) corresponding to the 3 abundance counts produced by the same replicate plant. This implies design involving ASE and total RNA-seq counts contains some sort of subsampling, multiple observations from the same biological sample are measured.

3.2.2 Gene expression patterns

A non-heterotic gene expression is characterized by an additive pattern in the comparison of the hybrid with parental lines, i.e. genes where the transcript abundance in the hybrid is equal

to the average abundance in the inbred. Non-additive gene expression patterns are related to heterosis. When the hybrid expression is similar to the high (low) parent expression is called high (low) dominance, and if the hybrid expression is above the high (below the low) parent it is called overdominance (underdominance) (Swanson-Wagner et al., 2006).

There are 3 gene expression patterns used to characterize gene heterosis. Mid-parent heterosis, MH, corresponds to the expression in the hybrid is different from the average expression of the parental lines. Low-parent (high-parent) heterosis, LH (HH), occurs when the gene expression in the hybrid is lower (higher) than the minimum (maximum) of the two parental lines. The last two patterns can be referred jointly as extreme-parent heterosis (EH).

Usually, models for detect gene heterosis use total expression only (Niemi et al., 2015). In this paper, we complement this analyses including ASE information and studying its relation with heterotic patterns, in distinguishing dominance from overdominance heterotic patterns.

There are two allelic patterns consider in this paper, the differential expression between allele expression in hybrids plants and the so-called allelic imbalance. Allelic difference, AD, refers to the ratio between the two alleles abundances (in log scale) in hybrid plants, while allelic imbalance, AI, compares AD ratio with the corresponding ratio of total expression in parental lines.

Figure 3.2 is a parallel coordinate plot of the total RNA-seq counts of the three varieties, for genes presenting extreme values in AD_g measure. We compute AD_g for every gene and its quantiles 1% and 99%, the plots show only genes with AD_g lower than the first percentile or higher than the 99 percentile. The plot suggests that most genes with very large allelic differences are associated with non-additive dominance expression patterns. In these genes, one of the inbred lines shows low abundances while the other is relatively high and the hybrid abundance is close to the high expressed parent. Then, by having both types of alleles, the hybrid is able to "choose" which allele to use and this gets reflected in allele expression.

Figure 3.3 is another parallel coordinate plot of total RNA-seq counts, but showing only genes with extreme AI_g instead. The plot suggests two main expression patterns in genes with large allelic imbalance. Genes where both parents having similar total abundance but the alleles in the hybrid show a big difference, and genes where one parent is more expressed than the

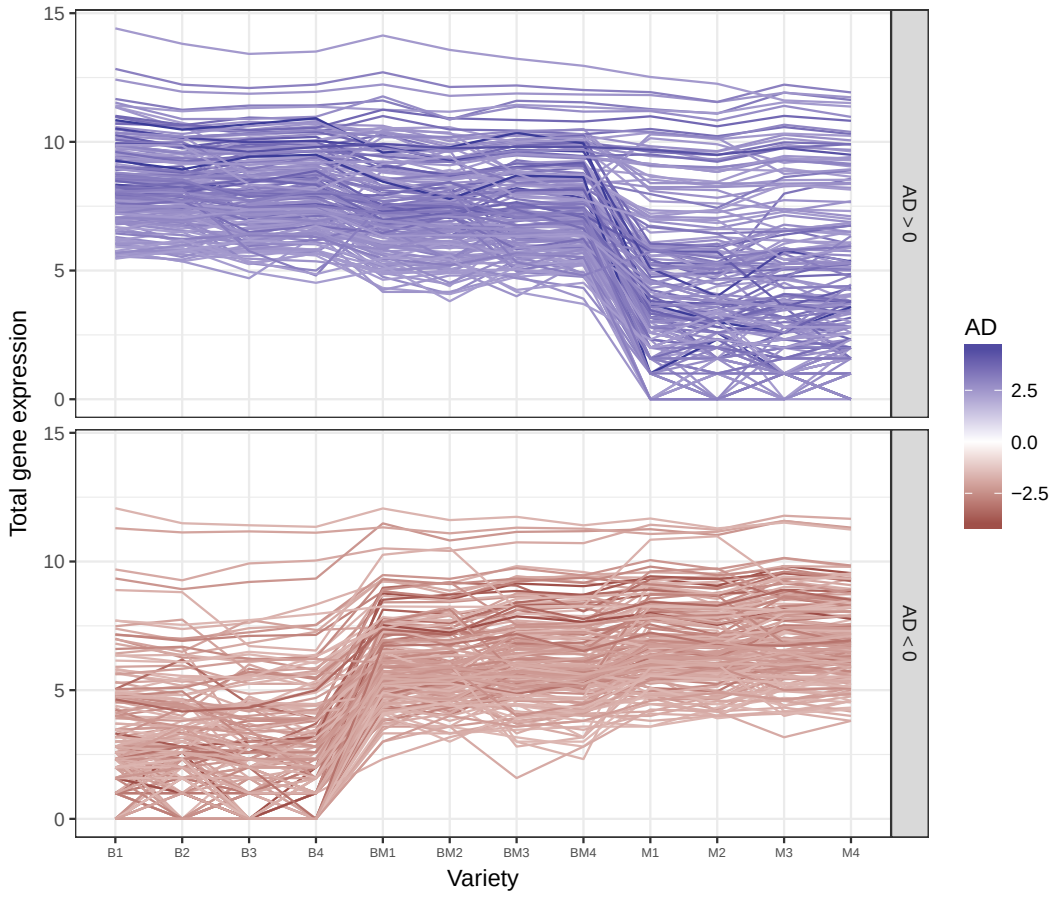

Figure 3.2: Total RNA-seq parallel coordinate plot in genes with large observed allelic difference (AD). The facet and color of the line distinguish genes with allele difference above 99th quantile (left panel, blue lines) or below 1th quantile (right panel, red lines).

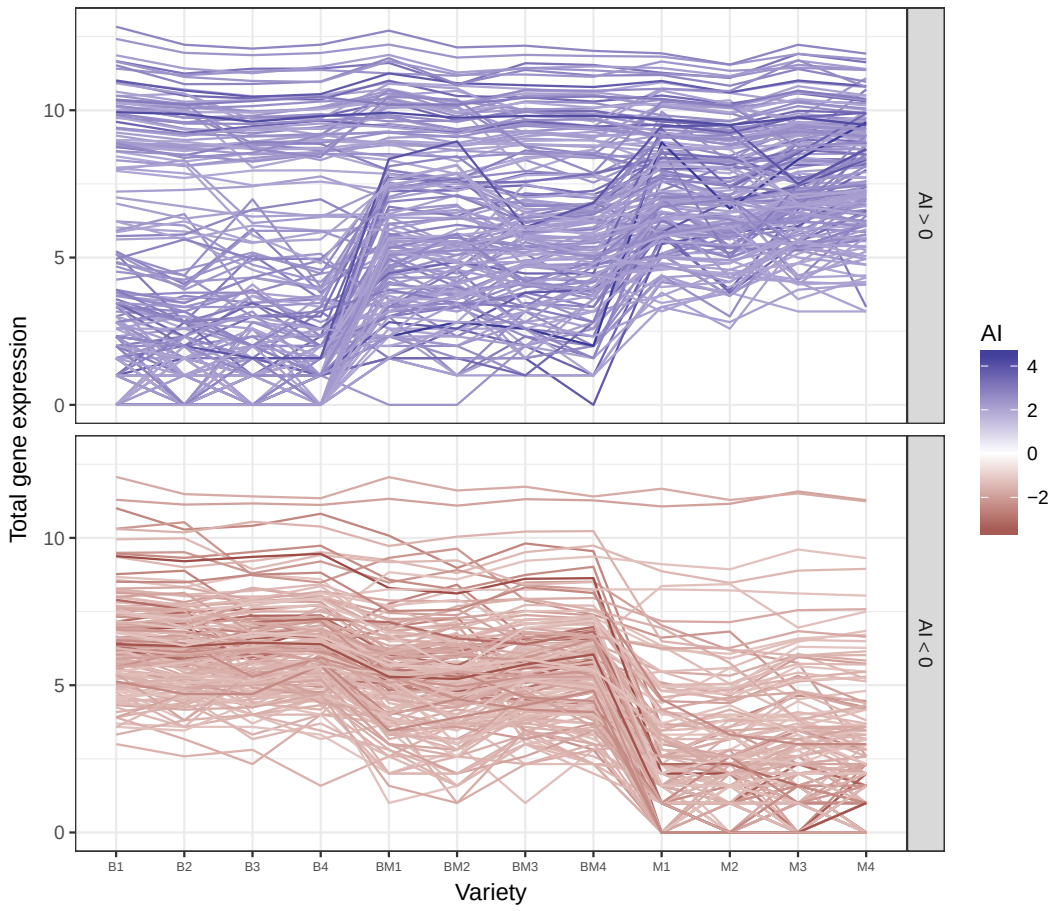

Figure 3.3: Total RNA-seq parallel coordinate plot in genes with large observed allelic imbalance (AI). The facet and color of the line distinguish genes with allele imbalance above 99th quantile (left panel, blue lines) or below 1th quantile (right panel, red lines).

other but this relation is inverted in the hybrid (for instance, blue increasing lines). However, many genes plotted in figure 3.3 present additive expression comparing hybrids to parental lines. The total expression for the hybrids plants, in genes with very large allelic imbalance, is close to the average abundance of the inbred parents.

3.3 Poisson-lognormal hierarchical model

In this section we describe the statistical model proposed to analyze the relationship between gene expression profile of interest. The model presented here is built over the model presented in previous Chapters to deal with ASE from a single hybrid variety, here we enlarge this model to incorporate more varieties and the total RNA-seq count.

The rest of this Section presents several aspects of the proposed method. First, the hierarchical Poison-lognormal mixture model is described. Then we focus in how to design a parametrization that ensures independence among the regression coefficients, a data-driven approach to this end is described. Then, a short description of the method for computing normalization factors is presented. The two final subsections corresponds to present the main contrasts of interest and the association measures among those contrasts.

3.3.1 Data model and hierarchical distributions

Let Y_{gn} the gene expression count of gene g in sample n, corresponding to a total RNA-seq count or an ASE count. Equation (3.1) present the data model, the abundance RNA-seq (total and allele-specific) counts are independently distributed as Poisson while its mean follows a lognormal distribution.

$$\begin{aligned} Y_{gn} &\overset{\text{ind}}{\sim} Po(e^{h_n + x_n^\top \beta_g + \epsilon_{gn}}) \\ \epsilon_{gn} &\overset{\text{ind}}{\sim} N(0, \gamma_g) \end{aligned} \tag{3.1}$$

Overdispersion effects are represented by ϵ_{gn} that allows to accommodate gene-specific mean-variance relationship, h_n corresponds to normalization factors and $x_n^\top \beta_g$ contains effects of genotype, allele, and blocking effects. We divide the $x_n^\top \beta_g$ in two main parts,

$$x_n^\top \beta_g = \mu_{gve} + r_{gn}$$

where μ_{gve} are 9 cell means of potential interest, combining variety and expression type. Variety index $v = \{B, BM, M\}$ indicates BB, BM, MM genotypes respectively, and $e = \{b, m, t\}$ indicates allele B, allele M and total RNA expression, respectively. The term r_{gn} contains all the experiment-specific effects (blocking factors, covariates, subsampling effects, etc).

Equation (3.2) presents the gene-specific layer of the model consisting of a hierarchical distribution for regression coefficients and overdispersion variance. Gene-specific regression coefficients are independent across genes (given hyperparameters) and the component of β_g vector are independent within each gene, similarly overdispersion variances are conditionally independent across genes, and independent from the gene regression parameters.

$$\begin{aligned} \beta_{gk} &\overset{\text{ind}}{\sim} \text{Cauchy}(\theta_k,\ \sigma_k^2) \\ \gamma_g &\overset{\text{ind}}{\sim} IG(\tfrac{\nu}{2}, \tfrac{\nu\tau}{2}) \end{aligned} \tag{3.2}$$

A Cauchy distribution is used since it showed good results in detecting differentially expressed alleles in sparse context (see chapter 2) and is an easy way to avoid possible confounding of mean signals (see chapter 4). Depending on the parametrization of the model matrix (discussed blow) it might be preferable a normal distribution for the intercept term, since this type of effects does not present the same level of sparsity.

Gene-specific overdispersion variances are modeled as independent across genes with an inverse-gamma distribution. Parameter τ affects the central location of the distribution (the median) while ν is related with the right tail weight and the shrinkage around τ.

3.3.2 Model matrix parametrization

Vectors $x_n^\top$ are the rows of the model matrix X. Is convenient to divide X two submatrices corresponding to separate effects that might be relevant in general from effects that are experiment specific. We consider

$$X = [M|R]$$

such that the matrix M has the 9 columns that corresponds to the cell means, μ_{gve}, involved in the relevant contrasts, and R contains all blocking factors and other effects specific to a particular experiment. The hierarchical model needs a parametrization that ensures independence

among the components for the gene-specific regression coefficient vector in the model. In this subsection, we focus in the method to obtain a convenient parametrization for matrix M

We first show that some basic parametrization for M produce correlated coefficient point estimates, and then we propose an empirical strategy to obtain the matrix X. Initially we consider parametrizations correspond to contrast matrices commonly used in linear models for two-factorial designs, with the two factors being variety and expression type. Cell mean parametrization consist in simply an identity matrix where each column indicates one of the combination of variety and expression type. The so-called SAS parametrization includes an intercept column and eliminates some of the columns of variety and expression type in order to get a full rank matrix. Similarly, a zero-sum parametrization includes and intercept and the rest of the columns are set to add up to zero.

In addition to these parametrizations, a so-called Fisher parametrization is considered. Fisher parametrization it is used in Ji et al. (2014); Niemi et al. (2015) to model total expression type counts, having one coefficient for the half difference among the hybrid and the average of parental lines and two coefficients for the half difference of the hybrid expression and each parent. Here we applied the same approach to each expression type separately.

To study the correlation structure among regression coefficients β_g under different parametrizations, we estimate model (3.1) on the data set described later. The model is estimated with `edgeR` library using every different parametrizations described above and then the sample correlation matrix among point estimates of regression coefficients is computed.

Figure 3.4 shows the sample correlation for the different parametrizations. Using a cell means parametrization result in the worst scenario since most of the coefficients pairs present high correlations. It is possible to identify groups of highly correlated cells, for instance, the three variety means with total RNA expression, or the two allele-specific expression means in the hybrid variety. Zero-sum and SAS contrast matrices show improvement respect to the cell means but there are still some pairs of coefficients highly correlated. Fisher parametrization seems to work best in this data with the highest correlation equal to 0.6.

Therefore based on the estimated correlations we could choose to use a Fisher parametrization. However, this result might depend on the particular data set, we prefer to have a method

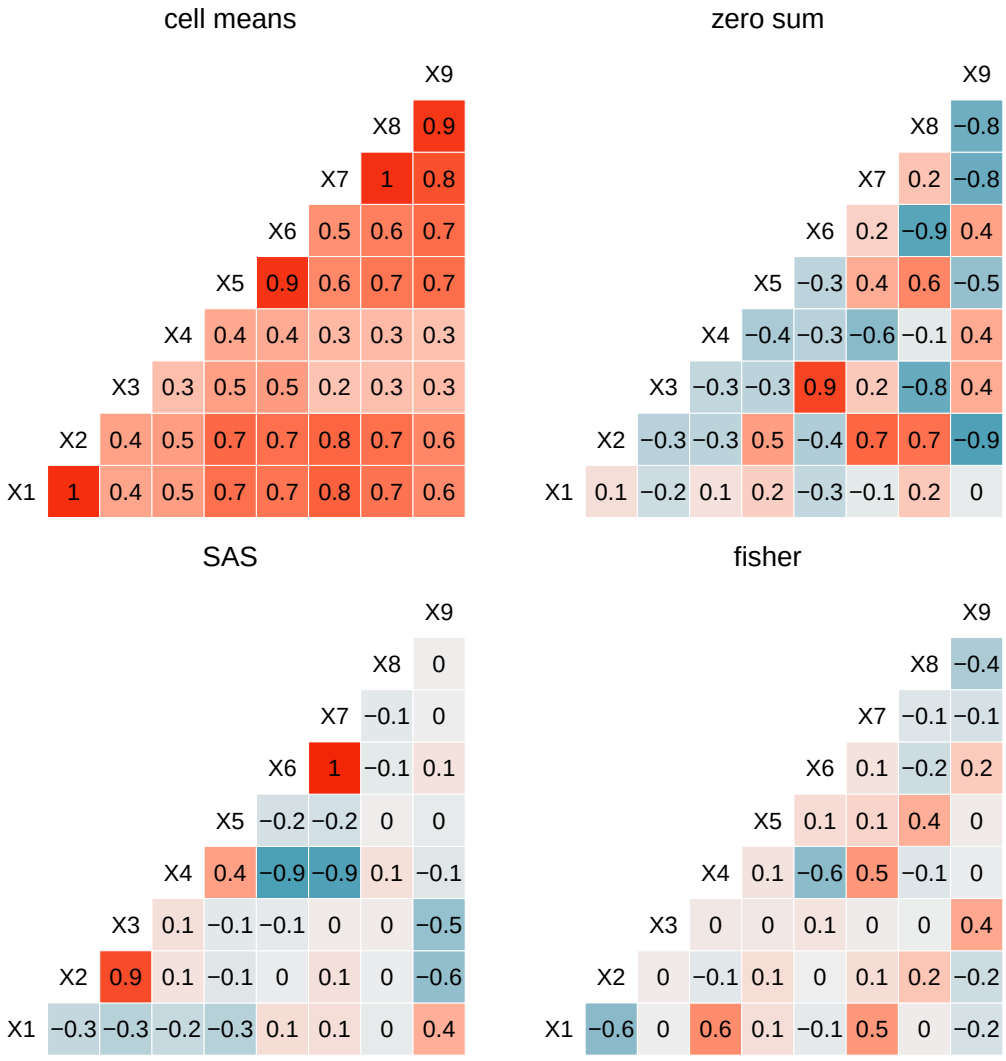

Figure 3.4: Tile plot of correlation matrices. Correlation among regression coefficient point estimates, from `edgeR`, for different parametrizations. Color indicates the direction in the correlation coefficient and color intensity indicates its absolute value.

to design a model matrix that could work in any experimental data set.

We follow an approach suggested by Lithio and Nettleton (2015), who addressed this problem by using a data-driven approach based on computing a singular value decomposition. The model matrix is obtained through the following steps:

1. set $M = I_9 \otimes 1_n$
2. Use `edgeR` to obtain $\hat{\mu}_g$, with $[M|R]$ as model matrix
3. Compute $var(\hat{\mu}_g) = \hat{V}_\mu$
4. Compute $\hat{V}_\mu$ SVD decomposition $\hat{V}_\mu = Q\ D\ Q^\top$
5. Use $X = [M|R]\ Q$ as model matrix in model 3.1

The first two steps corresponds to obtaining cell means estimates, $\hat{\mu}_g$, using `edgeR` package. To this end the matrix $[M|R]$ is set as model matrix (or design matrix in terms of `edgeR` functions), where $M = I_9 \otimes 1_n$ is the cell mean matrix and R has experiment-specific effects. Then we compute the sample covariance matrix of the point estimates, $\hat{V}_\mu$, and its SVD decomposition, $\hat{V}_\mu = Q\ D\ Q^\top$. The Q matrix is orthogonal and then a reparametrization $\beta_g = Q^\top \mu_g$ produce nearly uncorrelated components in the regression coefficient vector. The final model matrix is then computed as $X = [M|R]\ Q$

3.3.3 Normalization factors

The data model described in (3.1) contains normalization factors h_n, these are supposed to remove systematic difference among samples related to measurement technical details. In total RNA abundance data, each sample corresponds to a column in the data set, then models to analyze this data usually compute one normalization factor per sample. Some methods compute normalizations factors separately for each sample Bullard et al. (2009) and others use some column as a reference (Robinson and Oshlack, 2010).

Potential problems arise when dealing with ASE information and total RNA abundance together. Each sample produces more than one type of measures which break the correspondence between samples and data columns. Recently, Hodgkinson et al. (2016) suggest that

specific methods for computing normalization factors are needed when ASE is present. Their normalization method is based on a set of genes for which is known the allele expression is equal among the two alleles.

In this paper, normalization factors h_n are computed using total RNA-seq abundance counts only. Then the same value is used for all 3 counts coming from the same sample. This approach is biologically analogous to what is common practice in models dealing with no ASE. However, we need to keep in mind that some systematic differences might not be capture by this and then some other correction method is needed. For instance, in computing the AD value for each gene, we need to correct the reference allele bias. The hyperparameters in regression coefficients hierarchical distribution are use for this correction.

3.3.4 Contrasts

The gene expression patterns described earlier can be written as a function of the μ_{gev} parameters in model (3.1). Using total RNA-seq expression from hybrid and inbred parents is possible to identify gene heterosis. Instead, only ASE information in the hybrid is needed to detect genes with alleles differentially expressed. Allelic imbalance uses both types of expressions (ASE and total) from all three varieties.

$$\begin{aligned} MH_g = &\quad \mu_{gBMt} - (\mu_{gBt} + \mu_{gMt})/2 \\ LH_g = &\quad \mu_{gBMt} - \min(\mu_{gBt}, \mu_{gMt}) \\ HH_g = &\quad \mu_{gBMt} - \max(\mu_{gBt}, \mu_{gMt}) \end{aligned} \tag{3.3}$$

Equation (3.3) shows the three gene patterns associated with some form of heterosis in term of model parameters. Some form of heterosis occurs where $MH_g = \mu_{gBMt} - (\mu_{gBt} + \mu_{gMt})/2$ is different from zero, if no extreme heterosis occur then that particular gene present some evidence to a dominance explanation for heterosis. When low parent (or high parent) heterosis is present there is evidence in favor to underdominance (overdominance) explanations for heterosis.

Equation 3.4 show two interesting gene expression patterns involving ASE information. The first one corresponds to genes where the alleles are differentially expressed within the hybrid

variety. The mean value of allele difference across all genes is expected to be positive due to systematic bias towards allele B73. We use the hyperparameters θ of model (3.1) to remove this systematic effect (see chapter 2). Let c_{ad} the vector of coefficients to obtain the difference among alleles in the hybrid variety, the systematic bias is approximated by $c_{ad}^{\top} E(\theta|y)$.

$$\begin{aligned} AD_g = &\quad \mu_{gBMb} - \mu_{gBMm} - c_{ad}^{\top} E(\theta|y) \\ AI_g = &\quad AD_g - (\mu_{gBt} - \mu_{gMt}) \end{aligned} \tag{3.4}$$

Allelic imbalance compares the allele difference in the hybrid with the differential expression of that gene among inbred parents. Allelic difference and allelic imbalance are important contrasts to discover allelic-specific expression patterns. A gene with AI_g different from zero represents the hybrid variety was able to adapt in some way by modifying how much of that gene is used in parental lines. It has been proposed as one explanation for the occurrence of hybrid vigor.

Note that neither AI_g nor any of the heterosis contrast need to be corrected for systematic difference (like bias), this is because the normalization factors used in the model are computed using total RNA-seq expression and capture these effects. But a pattern that compares ASE with total RNA-seq directly would need such type of correction. For instance, this would happen in using false positive corrections obtained from ASE expression in inbred lines to correct total RNA-seq expression. Such corrections are not be used in this paper.

3.3.5 Relationship among contrasts

We propose two gene-specific measures to study the relationship among the gene expression pattern described above: a correlation coefficient for linear association, and conditional probability for statistical dependence. Let C_{μ} is a matrix of contrast over cell mean, i.e., a matrix with three rows to compute the linear combinations corresponding to MH_g, AD_g and AI_g from equations (3.3) and (3.4) (for heterosis patterns the only linear combination is H_g). Contrasts can be expressed as linear combination of the regression parameters, β_g, as

$$\lambda_g = (H_g,\ AD_g,\ AI_g) = C_{\mu}\ Q\ \beta$$

Posterior distribution for the gene-specific regression coefficients, β_g, can be approximated by $\beta_g \sim N(E(\beta|y), \hat{\Sigma}_g)$ where $\hat{\Sigma}_g = \text{diag}(Var(\beta_g|y))$. This ignores correlation among regression coefficient in the posterior, however, because of the parametrization is designed to produce independent β_gs so these correlations effects should be small.

Based on the covariance matrix approximation of β_g coefficients, $\hat{\Sigma}_g$, is easy to obtain the correlation among relevant contrasts. Since λ_g are linear combinations of β_g we can compute the corresponding covariance and correlations matrices as follows:

$$\begin{aligned} Var(\lambda_g|y) = &\quad \Lambda_g = C_\mu \, Q\hat{\Sigma}Q^\top C_\mu^\top \\ Cor(\lambda_g|y) = &\quad \text{diag}(\Lambda_g)^{-1/2} \, \Lambda_g \, \text{diag}(\Lambda_g)^{-1/2} \end{aligned}$$

A second way to study the relationship among gene patterns is based on conditional probabilities. Two main conditional probabilities of observing a gene heterosis pattern are of interest. First the heterosis probability given that allele differential expression is present, $P(|MH_g| > c_1||AD_g| > c_2)$, and second the heterosis probability given that allelic imbalance is present $P(|MH_g| > c_1||AI_g| > c_3)$. Also, joint and marginal probabilities of the occurrence of each gene pattern can be of interest.

Samples from a MCMC algorithm are typically used to obtain posterior probabilities, but MCMC samples for the gene-specific parameters are not available. Library `fbseq` allows computing this type of probabilities at each iteration of the chain. The only issue for this application is that c_1, c_2, c_3 need to be fixed and specified. Probabilities need to be expressed in linear combination pieces, so in order to consider absolute values, we need to decompose it into mutually exclusive parts. For instance,

$$\begin{aligned} P(|H_g| > c_1, |AI_g| > c_2) &= P(H_g > c_1, AI_g > c_2) \\ &\quad + P(H_g > c_1, -AI_g > c_2) \\ &\quad + P(-H_g > c_1, AI_g > c_2) \\ &\quad + P(-H_g > c_1, -AI_g > c_2) \end{aligned}$$

Linear correlations are can be computed (easily) among linear contrasts only. Instead, the conditional probabilities measures can be extended in the same way to include extreme heterosis and are computed directly from MCMC samples.

3.4 Data analysis

The methods described in previous sections are applied to a specific gene expression data of maize plants. The data comes from (Paschold et al., 2012) experiment, it consist in 4 replicates biological samples from 3 maize varieties: inbred B73, inbred Mo17 and its hybrid cross B73×Mo17. The 4 replicates are analyzed in 2 flow cell blocks with Illumina technology to obtain RNA abundance data. There are 3 types of RNA abundance obtained for each gene in every replicate sample: the total abundance and the abundance for the two alleles.

There are $n = 36$ samples per gene corresponding to three genotypes or varieties (v), two flow cell blocks (f), two replicates plants (r) and three transcript counts relative to expression types (e) for the total RNA-seq and the two alleles.

3.4.1 Bayesian inference

The model matrix is computed with the data-driven procedure described before, which starts with the matrix $[M|R]$ where M corresponds to the cell means matrix for variety and expression type and R corresponds to the experiment specific effects. In Paschold et al. (2012) data, matrix R is formed by two sets of factors, ϕ_f and $(\phi\delta\gamma)_{fvr}$, which can be set up as 7 columns with zero-sum restriction, as follows

$$\phi_f = \begin{cases} 1 & f = 1 \\ -1 & f = 2 \end{cases} \qquad (\phi\delta\gamma)_{fvr} = \begin{cases} 1 & r = 1 \\ -1 & r = 2 \end{cases} \quad \text{for each}(f, v)$$

The value of ϕ_f represents the half difference among the two flow cell blocks, while $(\phi\delta\gamma)_{fvr}$ represent the half different among two replicate plants within each variety and flow cell combination.

Fully Bayesian inference is obtained with `fbseq` package, running 4 MCMC chains with 15000 iterations after burning and keeping only 1 every 5 simulated values. The potential scale reduction factor statistic for each individual parameter is used to identify lack of convergence

in the model. Prior distribution of the hyperparameters are set as follows:

$$\begin{aligned}
\theta_k &\overset{\text{ind}}{\sim} \text{N}(0, c_k) \\
\sigma_k^2 &\overset{\text{ind}}{\sim} \text{Unif}(0, s_k) \\
\nu &\sim \text{Unif}(0, d) \\
\tau &\sim Ga(a, b)
\end{aligned}$$

These priors are set to be weakly informative so in the high-dimensional context this is being applied the priors effects are dominated by the likelihood.

3.4.2 Data analysis results

Genes showing evidence of mid parent heterosis (MH) consists in genes where the hybrid total RNA-seq abundance is different from the parental average total RNA-seq abundance, high parent heterosis (HH) are genes where the hybrid is total RNA-seq is larger than the maximum among parental lines and low parent heterosis (LH) is the opposites, i.e., genes where the hybrid expression is less than the minimum among parental lines.

In all three patterns a threshold is needed for consider the hybrid expression different, larger or lower, here we use a 25% fold change as the threshold, in terms of the contrasts presented earlier this means that $c = log(1.25)$.

Results indicate there are 1420 (9%) genes where probability of mid-parent heterosis is larger than 0.5, while there are very few genes with LH heterosis pattern (48) or HH heterosis pattern (102) with probability larger than 0.5. Having very few genes with some kind of heterosis differs from previous analysis of this experimental data (Paschold et al., 2012; Niemi et al., 2015). However, in this paper we only use genes with ASE information available which represent 40% of the total genes with total RNA-seq information. This could suggest that genes with extreme-gene heterosis patterns do not contain information about ASE. Additionally, previous analysis do not use a threshold value to define the null region, which might explain these results to be more conservative.

The relationship between heterosis and contrasts involving ASE is studied only using MH pattern. This is because mid-parent heterosis contains the other forms and also due to the

very low proportion of genes with ASE information available that present high probability of extreme heterosis.

The first type of measure to study MH-AD and MH-AI relationship presented above was the gene-specific correlation coefficient. A bivariate histogram of estimated $(\rho(MH, AD), \rho(MH, AI))$ is presented in left panel of Figure 3.5. Most of the genes show a correlation close to zero in both measures, also it seems to be a positive association between the two correlation coefficients.

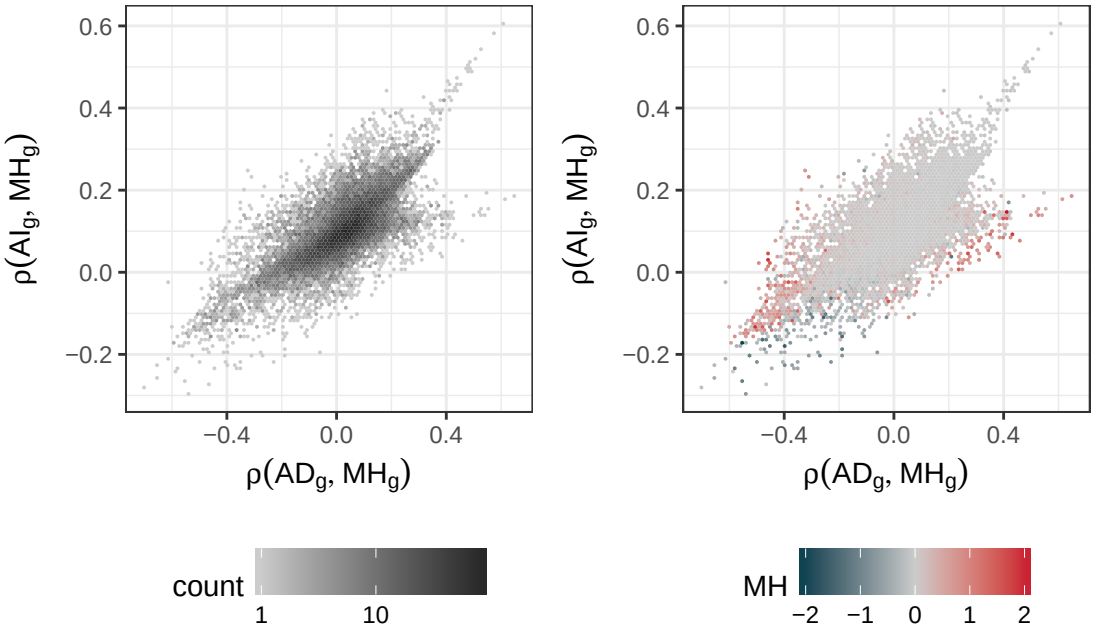

Figure 3.5: Gene-specific correlations between mid-parent heterosis and allele diffence patterns, $\rho(MH, AD)$, against correlation between mid-parent heterosis and allelic imbalance, $\rho(MH, AI)$. The left panel is a bivariate histogram, where darker points representing higher counts. Right panel is an hexagonal heatmap where color represents the mid-parent heterosis contrast (MD).

Left panel of Figure 3.5 does not include information about the MH_g contrast value. Heterosis information is included in the right panel, here the color represents the average value of MH_g in the hexagon cell. The plot suggests a weak pattern, genes where heterosis contrast is positively large (red cells) show high correlation with the allelic difference (in both signs) and low correlation with allelic imbalance. Also, genes where the MH_g value is most negative (blue

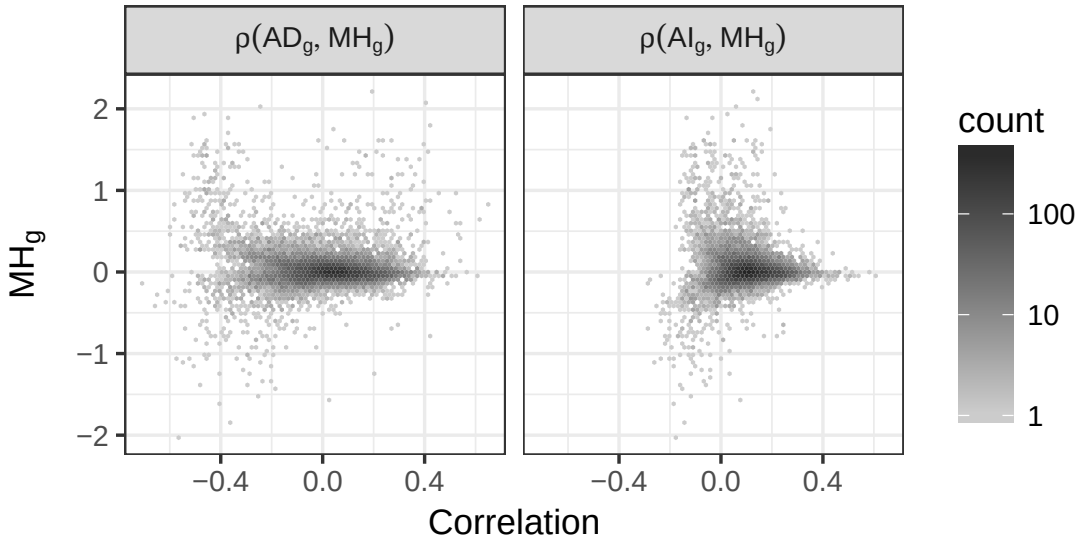

Figure 3.6: Bivariate histograms of gene-specific correlations and mid-parent heterosis contrast (MH_g). Left panel shows correlations between mid-parent heterosis and allele diffence patterns, ($\rho(MH, AD)$, and right panel shows correlation between mid-parent heterosis and allelic imbalance, $\rho(MH, AI)$).

points), are mostly located in a region with both negative correlation.

The same pattern is present in Figure 3.6. The left panel shows how genes with large absolute value in MH_g also shows the largest correlations with AD_g in the data and low correlation with AI_g. On the other hand, genes where the correlation between MH_g and AI_g is large there is no gene heterotic pattern present.

In section 3.2.1 where the maize data characteristics were described, Figures 3.2 and 3.3 indicate a similar history. Genes with extreme values in AD_g have expression patterns consistent with dominance heterosis, but genes with extreme values in AI_g show additive expression in the hybrid compared with the inbred parents. Using probabilities of observing each gene expression pattern are shown in Figure 3.7, the color of each cell indicates the average probability of MH_g in that cell. Genes where the probability of showing heterosis is high, are usually associated with high probabilities of at least one of the allelic patterns.

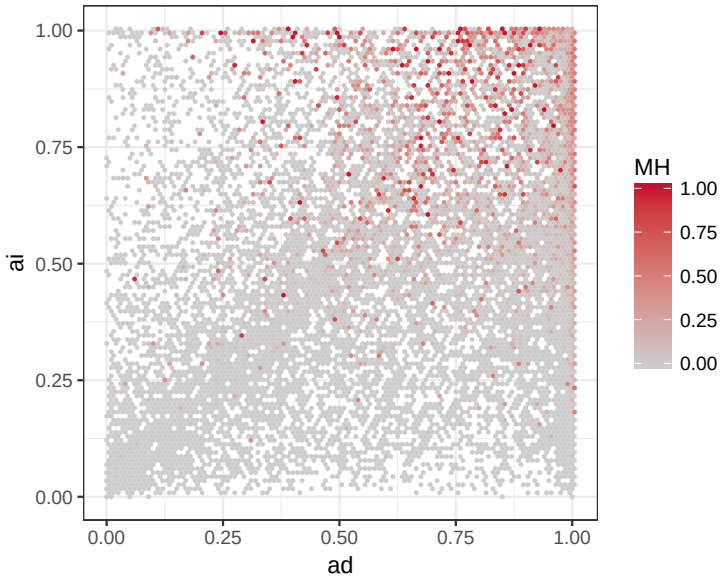

Figure 3.7: Hexagonal heatmap of probability of allelic difference and probability of allelic imbalance, color represents the average probability of mid-parent heterosis in the hexagon cell ($P(MH_g > log(1.25))$

3.5 Discussion

This paper used total RNA-seq and ASE data to characterize the relationship between heterosis and allelic differential expression.

A Poisson-lognormal mixture model is proposed to analyze RNA-seq transcript abundance counts, including total RNA-seq and ASE counts in the same model. A data driven procedure that is used to obtain the model matrix of the regression model. This procedure makes suitable to model the regression coefficients independently.

There are five gene expression patterns of interest, three of those patterns corresponds to mid-parent and extreme-parent heterosis (low or high) and the other two are related to the differential expression of the alleles in hybrids plants. The relationship among these contrasts is explored using parallel coordinate plots, and with association measures. The two gene-specific measures relating heterotic gene expression patterns and allelic differential expression patterns are the linear correlation coefficient and the conditional probability of heterosis given allelic imbalance.

Allelic difference is related with mid-parent heterosis. Genes with large MH_g are associated with large correlation values among MH_g and AD_g, additionally, genes with a large allelic difference can be associated with mid-parent heterosis showing gene expression patterns consistent with dominance explanation of heterosis. The relation between heterosis and allelic imbalance is weaker. Genes with large values in MH_g usually present correlation with AI_g close to zero, and the genes with largest values in AI_g shows additive expression patterns. There are only a few genes with large probability of extreme heterosis among the genes with ASE information available, suggesting the relationship between extreme heterosis and ASE patterns.

There is further work needed in two aspect of this paper. Firstly, the biological relevance of the statistical measures of dependence among heterosis and allelic imbalance are need to be explored deeper. Secondly, it might be worth to study the relation between normalization factors and false positive corrections, in order to include false positive information in the model.

CHAPTER 4. Confounding effects double shrinkage hierarchical models

Abstract

Double shrinkage models are used when there are two set of parameters with hierarchical distributions. For instance, cases in which is needed to model group means and group variances.

In this paper, we deal with hierarchical models for both mean and variances simultaneously in sparse high dimensional context. Specially we focus on the effects of variance hierarchical modeling on the mean vector inference. The inference about group means and variances can be confounded. Resulting in some surprisingly poor estimates of the means when signals are "too" strong.

This confounding effect occurs in both, simple simulated scenarios and in real data set from RNA-seq expression in maize plants. The main reason for the confounding effect is related to the light tails of the normal distribution. When a normal distribution is used as hierarchical distribution for the group means, the few groups with strong signals are mistakenly shrunk towards zero. Changing the hierarchical distribution to Cauchy seems to solve this issue.

4.1 Introduction

An important traditional problem in statistics is to make inference about large dimensional vector of means. Suppose we are interested in a simple normal model $y_{gi} \sim N(\mu_g, \sigma^2)$, where $i = 1, \ldots, n$ and $g = 1, \ldots, G$ and the variance σ^2 is known. Initial papers like James and Stein (1961) and Lindley (1962) point out that hierarchical approach is beneficial when the main goal is to obtain inference about the group means μ_g. James and Stein (1961) propose a shrinkage estimator for μ_g and show it has lower risk than maximum likelihood estimator when $G \geq 3$, which implies the inadmissibility of a maximum likelihood solution. Lindley (1962) described how the proposed shrinkage estimator arises naturally by taking a Bayesian approach to the estimation problem.

The power of a hierarchical approach, is that it make use of latent information between the groups that only can be seen when the estimation problem consider all groups together (Efron, 1992). Within a context of high dimensional problems, where $G \gg n$, the share of information among components of the mean vector is crucial. The importance of this type of problems have increased in recent years, in particular, in the analysis of gene expression data where the information sharing approach has been extremely successful.

Several hierarchical distributions for the means μ_g has been proposed, Lindley (1962) initial model uses a normal distribution, $\mu_g \sim N(0, A)$. In sparse scenarios, i.e., most of the means close to zero, we could use spike and slab prior that consist in a mixture of two components: a point mass at $\mu_g = 0$ and a normal distribution (Berger and Strawderman, 1996). Also, there are continuous approximations to this mixture, using of so-called shrinkage priors that pose a great proportion of the probability mass around 0 while having heavy tails to capture groups with larger means. Popular options for this are Laplace prior (Park and Casella, 2008), the horseshoe prior (Carvalho et al., 2010), and Cauchy distribution.

4.1.1 Hierarchical models for variances

One limitation of the previous approach is the assumption of equal variances. In practice, we are more interested in cases like (4.1) where each group has its own variance parameter.

In applied context, it is common to allow each group to have a different but known variance parameter replacing σ_g^2 with group sample variances (see references in Jing et al. (2016)), although the main inferential interest is still the group means.

$$y_{gi} \overset{ind}{\sim} N(\mu_g, \sigma_g^2) \tag{4.1}$$

In order to build hierarchical models for base on the data model (4.1), we need to set hierarchical distributions for group means and group variances. The modeling of the mean vector had been extensively studied for several decades, some of the main options were described above. On the other hand, the hierarchical modeling of the variances has not receive that much attention.

The inverse gamma distribution is widely use to model variances, mainly because it can be used to set a conjugate prior in the normal model (see Gelman et al. (2013) Section 3.3). Further, inverse gamma distribution can accommodate skewed patterns (few groups with much more variability than the rest) and at same time, it can shrink around a central value. Equivalently is possible to use a gamma as hierarchical distribution of group precision (see Kruschke (2014) Section 15.2).

Gelman (2006) describes some problems with inverse gamma distribution when it is used for latent variables in hierarchical models, and propose to use half-Cauchy distributions peaked at 0. However, inverse-gamma problems appear when it is used to approximate a non-informative distribution, so it is still possible to use it for the purposes of sharing information about variability within groups. Additionally, using Cauchy peaked at 0 it won't shrink towards a central value. Other options to model variances are distributions with support in the positive part of the real line, as lognormal, gamma or Weibull distributions.

Recently some models with shrinkage in means and variances at the same time within an empirical Bayesian framework has been proposed, so-called double-shrinkage estimators. In modeling gene expression data, Cui et al. (2005) uses a James-Stein estimator for log-transformed variances and then transformed back to obtain an exponential shrinkage estimator of the variances components. They use the exponential estimator to develop a test to detect differentially expressed genes. *"A more realistic model is that both means and variances of ex-*

pression values can be modeled with some distributions and these distributions can be estimated by borrowing information across genes" (Hwang and Liu, 2010).

Hwang and Liu (2010) use a normal distribution for means and log-normal for variances to construct a double-shrinkage test to detect differentially expressed genes. The same model, called log-normal model, is used to construct empirical Bayesian confidence intervals (Gene Hwang et al., 2009) and to obtain empirical Bayesian point estimates (Zhao, 2010) for the gene mean expression level. Alternatively, empirical Bayesian estimates using double-shrinkage approach but with inverse-gamma as hierarchical distribution for variances was recently proposed (Jing et al., 2016).

In this paper, we deal with hierarchical models for both mean and variances simultaneously in sparse high dimensional context. Specially we focus on the effects of variance hierarchical modeling on the mean vector inference. The issues we describe are relevant any situation where there is a grouping structure, with few observations per group and many groups, and the effect of interest is known to be sparse. Gene expression data is one interesting scenario with the mentioned characteristics.

Next Section presents the results of a hierarchical normal model on a gene expression data set. Specific details of the data set and the inference method are presented later, but we start showing some of the model results since they motivate the rest of the paper. In Section 4.3, a simple simulation study is presented to illustrate the so-called confounding effect. Section 4.4 presents summary statistics proposed as diagnostic for the confounding problem, and Section 4.5 show modeling options that avoid the problem. Details on the real data set and an appropriate model to analyze the data, i.e. without confounding effect, is presented in Section 4.6.

4.2 An initial model for gene expression data

This Section presents an RNA-seq expression data used as motivating example, and the initial hierarchical Bayesian model to analyze that data set.

A diploid genome has two sets of chromosomes, one from each parent, so every gene has two copies. One of the advantages of next generation sequencing is that makes possible to measure the expression of each gene copy, we call allele-specific expression (ASE) to refer this measure.

ASE can be obtained using single nucleotide polymorphism (SNP) that makes it possible to distinguish the expression of the two alleles (Sun and Hu, 2014). We assume the ASE counts for a single hybrid variety are available, the main interest is to detect genes that present alleles differentially expressed.

Let B_{gi} and M_{gi} the transcript abundance count for gene g in sample i for the reference and non-reference alleles respectively. In this paper we do not model directly the observed transcript abundance for each allele, instead a logarithmic transformation and then normal data model is used.

The log-transformed allele ratio, d_{gi}, and its centered version, y_{gi} are defined as follows:

$$d_{gi} = \log\left(\frac{B_{gi}+1}{M_{gi}+1}\right) \qquad y_{gi} = d_{gi} - \frac{\sum_{gi} d_{gi}}{\sum_g n_g}$$

where the addition of 1 read to each count in the ratio d_{gi} ensure the transformation is well defined.

The response variable is defined as the centered allele expression ratio in logs, y_{gi}. Is important to center the allele ratio to remove any systematic difference affecting all genes, these effects are not relevant to detect interesting genes and are usually contaminated with bias.

It could be argued is better to model directly the ASE counts instead of the log-transformed version. However, the main interest here is to illustrate confounding effects and explore its reasons. Normal hierarchical models are better for this goals since they are more analytically tractable. In addition, it has been suggested that transformation-based methods show competitive performance in detecting differentially expressed genes (Soneson and Delorenzi, 2013). Nevertheless, in Section 4.6 we also perform a data analysis of the ASE counts with a Poisson-lognormal mixture model that shows similar results in terms of confounding effects.

As an initial model, we assume response variable is normally distributed, group means also are normal while group variances distributed as inverse gamma, this model is presented in (4.2). As we mentioned in previous Section, the normal data model with both group-specific means and variances has receive attention in recent years to analyze microarray data (Gene Hwang et al., 2009; Hwang and Liu, 2010; Zhao, 2010). Using hierarchical distributions for

group means and group variance parameters, we can borrow information across groups, which is specially appealing when there are few observations per group and thousands of groups.

$$\begin{aligned} y_{gi} &\overset{\text{ind}}{\sim} \sim N(\mu_g, \sigma_g^2) \\ \mu_g &\overset{\text{ind}}{\sim} N(\mu_0, \sigma_0^2) \\ \sigma_g^2 &\overset{\text{ind}}{\sim} IG(\nu\tau/2, \tau/2) \end{aligned} \tag{4.2}$$

In terms of the log transformed ASE counts described above, the main interest is to identify groups (or genes) where $\mu_g \neq 0$ since it means that gene show different expression level among alleles. Figure 4.1 shows results of full Bayesian inference for model (4.2) (details on how inference is performed are described later), each panel is a bivariate histogram plot of posterior expectations and observed group means. Top row facets show the posterior expectation of the group means and bottom row facets show the square root posterior expectation of the group variances. Genes are declared as differentially expressed (DE) if a 95% credible interval does not contain zero, and declared as non-DE otherwise.

Figure 4.1 suggests that groups (genes) with observed sample mean close to zero are non-differentially expressed while genes with larger observed sample means results in credible intervals that not contains zero value. However, there are a few genes with large sample means which are not flagged as differentially expressed, the reason for this can be found in the bottom row panels, those genes have the largest estimated variances in the data set. This might constitutes a reasonable explanation of the signals present in the data. Genes with weak signals or large signals but too much noise are founded to be non-differentially expressed, while genes with strong signals and low levels of noise are flagged as DE.

However, it seems suspicions that all genes with the largest observed means result in small posterior expectation and large posterior variances. A similar effect is pointed out by Cook et al. (2007), the most interesting genes detected using plots had very large adjusted p-values.

In the rest of this paper, we make a case against the explanation of data provided by model (4.2), and suggest that sometimes strong signals are confounded with noise when we use a normal as a hierarchical distribution for the means in sparse context.

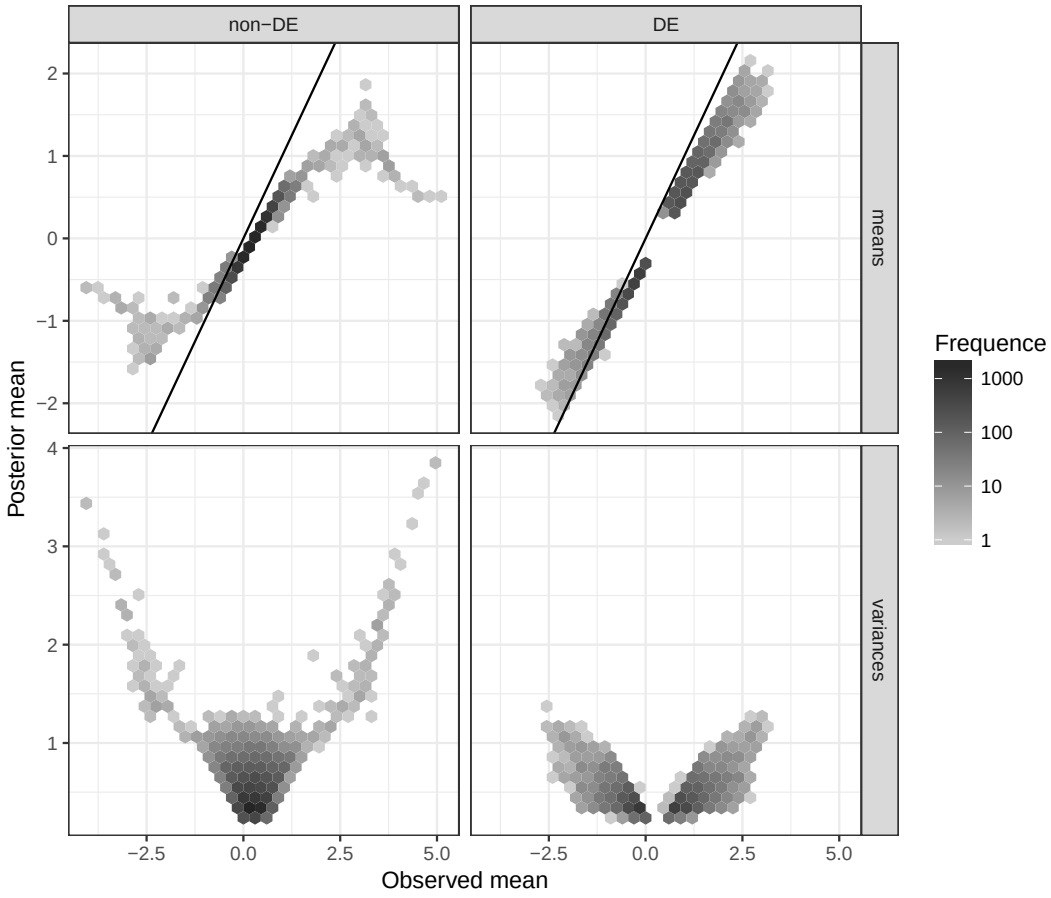

Figure 4.1: Bivariate histograms of model results for ASE counts of Paschold et al. (2012) hybrid data, model with normal and inverse gamma hierarchical distributions. Posterior expectation of group mean and group sample mean (top facets), and square root of posterior expectation of group variance and group sample mean (bottom facets). Column facets indicates genes has its alleles differentially expressed (DE) or not (non-DE).

4.3 The confounding problem

This Section presents a simulation study designed to illustrate the mean-variance inference confounding that can occur if we use the initial model presented in equation (4.2). The goal is to design a particularly simple scenario, where we would expect to get perfect results from the initial normal model, but instead the model performance is poor.

This Section describes the simulated data scenarios, the method to make inference for model (4.2), and the model performance results on the simulated data sets.

4.3.1 Simulated data scenarios

Data are simulated from a normal data model, $y_{gi} \sim N(m_g, s_g^2)$, setting values for m_g and s_g^2. We use $G = 1000$ groups and $i = 1 \ldots, 4$ observations per group. Group specific parameters, (m_g, s_g^2), are set as

$$m_g = \begin{cases} -m & g \in (1, 50) \\ m & g \in (51, 100) \\ 0 & g > 100 \end{cases} \qquad s_g^2 = T \quad \forall\, g$$

The group means, m_g, can take one of 3 values $(-m, 0, m)$ and 95% of them are set equals to zero. The design parameter m controls true signal strength, when m is large it should be easier to detect the groups with $m_g = m$. Meanwhile, all group variances are equal, $s_g^2 = T$, then design parameter T controls the level of noise in the simulated data. In terms of the double shrinkage procedures commented earlier, this scenario in variances is particularly simple, the models should gain a lot by modeling variances hierarchically in this case.

Table 4.1 presents the values for each design parameter. We combine 3 values signal strength and two values for noise level, to obtain six simulation scenarios for which we simulate one two replicate datasets.

Table 4.1: Simulation scenarios

	Parameter	Values
Signal Strength	m	1, 2, 4
Noise level	T	1, 4

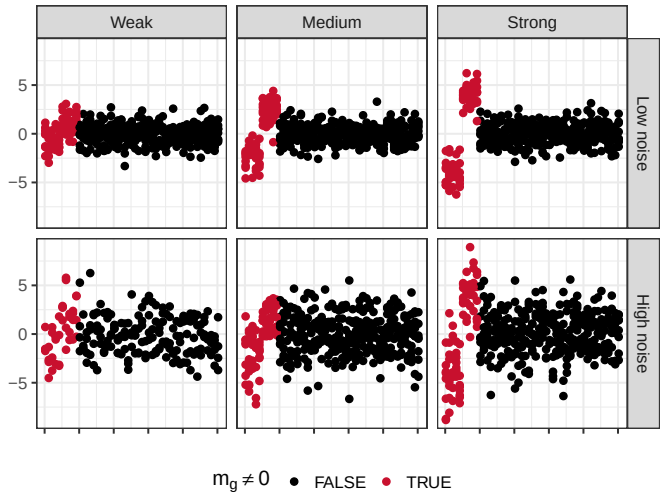

Figure 4.2: Scatter plot of simulated datasets for selected groups. The color indicates if the true mean group is different from zero (red) or not (black). Column facets correspond to signal strength level (m) and row facets correspond to noise level (T)

Figure 4.2 shows a scatter plot for selected groups in every simulated data set, color indicates if the group mean is zero or not. Rows facets correspond to the noise level and column facets correspond to signal strength level. The overall structure of the simulated data seem to be clear, which is expected since the simulation model we use is relatively simple.

The goal is to identify the groups with no zero mean, $m_g \neq 0$, i.e. the groups with red points in from Figure 4.2. Better detection is expected when signals are larger and noise is lower. However, as we show below, it might be the case that $m = 4$ produce group where the signal is "too big" and confounded with noise.

4.3.2 Initial model results

Each data set is analyzed with the initial model (4.2), to perform a fully Bayesian analysis of this model priors for the hyperparameters are needed, equation (4.3) shows these priors.

The two hyperparameters related to dispersion, σ_0 and ν have uniform priors, the mean across all group means, μ_0, is normally distributed while the group variances location related parameter τ is gamma distributed. All these priors are set to be weakly informative and have little influence the posterior inference. The hierarchical model uses directly all data to estimate

the hyperparameter posterior distribution, therefore in high-dimensional problems is expected that the prior is dominated by the data model.

$$\mu_0 \sim N(0, 100) \quad \sigma_0 \sim Unif(0, 100) \quad \tau \sim G(1, 1) \quad \nu \sim Unif(0, 100) \tag{4.3}$$

We obtain inference for each data set using STAN software (Carpenter et al., 2017), we run 4 MCMC chains with 5000 iterations after burning and monitor convergence with potential scale reduction factor statistic.

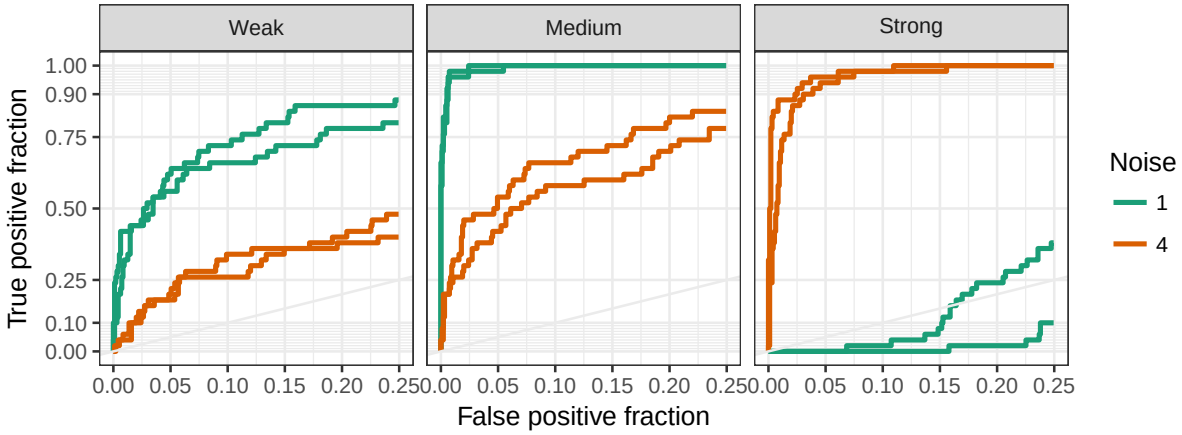

Figure 4.3: ROC curves for Normal-IG model. Column facets represent signal strength and color indicates the noise level.

When diagnostics show no signs of lack-of-convergence we construct receiver operating characteristic (ROC) using the ratio of posterior expectation of the mean (in absolute value) over the posterior standard deviation of the mean, as a continuous measure.

Figure 4.3 presents ROC curves for each simulated data using model (4.2), facets corresponds to signal strength and line color represents the noise level. Weak signals ($m = 1$) have low detection rates, and show even lower detection rates when the noise level is large. When signal strength increase to $m = 2$, ROC curves show clear improvements, in the scenarios with low noise level, $\sigma_g^2 = T = 1$, the true mean for groups with $m_g = 2$ is two standard deviations

from the zero group, so it is easily detected. Once again, adding noise makes ROC curves decrease. The counter-intuitive results occur when signals get stronger. In the scenario with $m = 4$ the model has opposite results to what is expected. In the low level noise case, ROC curve is flat out at zero detection rate, but with more noise is just fine. It seems that adding noise in the data improves model performance.

ROC curves describe the ordering of group means. Looking at group-specific parameter inference is possible to describe more explicitly the problems with model (4.2) when signals are too strong. Figure 4.4 shows posterior expectation for the group-specific parameters. The top panel shows the posterior expected values of each group mean, $E(\mu_g|y)$, against the group sample means. The bottom panel shows the posterior expectation of group each group variances $E(\sigma_g^2|y)$ against the group sample means in square value. In both panels, facets represent the signal strength, color represents the noise level, and shape indicates the replicate.

Overall, posterior expectations of μ_g shows more shrinkage towards 0 when the signal is weaker or the variability is larger. Shrinkage towards zero is a well-known and expected effect, particularly using a normal as a hierarchical distribution. However, this shrinkage should decrease when the true signal gets stronger or noise level gets smaller. Top panel in the figure 4.4 shows the opposite result, posterior means are really close to zero if the variability in the data is low.

The σ_g^2 posterior expectations in the bottom panel of Figure 4.4 bring some light on what might be happened with the mean inference. Scenarios with $m \leq 2$ show no relation among posterior expectation of variance parameter and group sample mean. However, in the paradoxically pathological case (strong signals with low variability), there is a clear overestimation of σ_g^2 in non-noisy context, moreover, groups variance are capturing the observed group mean effect. This suggests the model confounding the few true signals in a sparse environment with low noise.

4.4 Phase shift boundary

This Section explores, through simulations, the existence of a signal strength boundary where this sort of phase shift take place. This boundary is expressed in terms of some statistic,

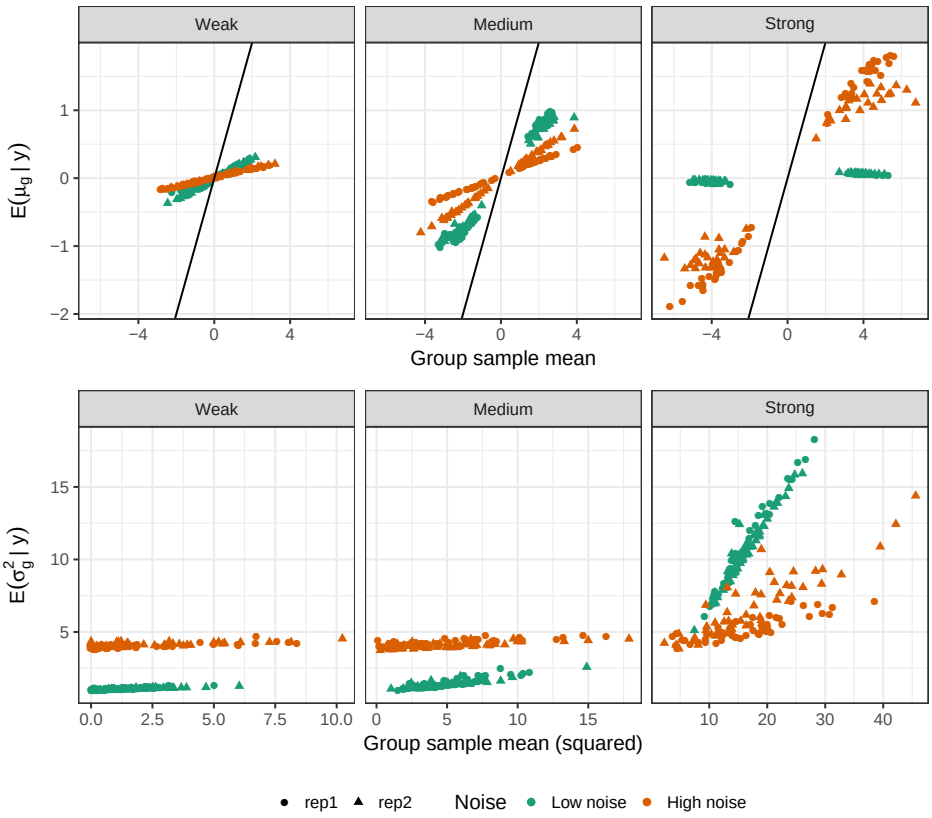

Figure 4.4: Results for Normal-IG model in simulated data. Scatter plot of μ_g posterior expectation against sample group mean (top panel) and scatter plot of γ_g posterior expectation against the group sample mean squared (bottom panel). Column facets represent signal strength and color indicates the noise level.

i.e. a function of the data, that correlates with the model performance.

First, we describe how the statistic to predict if model (4.2) going to fail in a particular data set is constructed, based on the notion of the mean-variance confounded issue described above. Next we simulate more data set using the same design than Section 4.3 but focused on signal strength with $m \in (2, 4)$, the proposed statistic is computed for each data set.

4.4.1 Alternative explanation for signals

A hierarchical model can be thought as a compromise between having a unique model for all groups (complete shrinkage) and having an independent model for each group (no shrinkage). In the case of model (4.2) there is a hierarchical layer in both means and variances.

Table 4.2 presents the possible results in terms of shrinkage level for each set of parameters. For instance, with respect to the group means, no shrinkage situation is characterized for each group having its own μ_g unrelated to the rest. While total shrinkage is the opposite, $mu_g = \mu_0$, all groups have the same group mean value. The partial shrinkage is between these two scenarios, the group means are different among groups but all share the same probability distribution. A similar construction can be made with the variances because the IG distribution can shrink towards a common value.

Table 4.2: Shrinkage levels in model (4.2)

Shrinkage	Means	Variances
None	μ_g	σ_g^2
Partial	$\mu_g \sim N(\mu_0, \sigma_0^2)$	$\sigma_g^2 \sim IG(\nu/2, \nu\tau/2)$
Total	$\mu_g = \mu_0$	$\sigma_g^2 = \tau$

Combine shrinkage for means and variances may result then in one of nine combinations. However, datasets simulated in the previous Section are really simple, all group variances are equals and the means has only three possible values. Because of the simplicity in the data scenarios, it is suitable to think the resulting shrinkage its only one of two possible options.

When the model correctly identify there are some groups with large means, then all variances will be pulled towards its common value. This corresponds then to a case with partial shrinkage in means and total shrinkage in variances. On the other hand, when the model confounds the

strong signal with variability, all means will be pulled towards $\mu_g \approx 0$, then corresponding to partial shrinkage in variances and total shrinkage in means.

Then, there are two possibilities results for shrinkage level in the simulated cases. As the data model is normal, is possible to integrate out the group-specific parameters under each of the possible shrinkage levels. When the model completely shrinks the variance to one value, the mu_g parameter can be integrate out to obtain $y_{gi} \sim N(\mu_0, \sigma_0^2 + \tau)$, the marginal distribution of the response variable only depends of the hyperparameters in the model. A similar argument can be made for the alternative model in which all means are shrunk to 0, here the marginal likelihood becomes $y_{gi} \sim t_\nu(0, \tau)$.

Table 4.3: Alternative shrinkage effects on simulated data

		Means	Variances	Observed data
M_μ	Signal as μ_g	$\mu_g \sim N(\mu_0, \sigma_0^2)$	$\sigma_g^2 = \tau$	$y_{gi} \sim N(\mu_0, \sigma_0^2 + \tau)$
M_σ	Signal as σ_g^2	$\mu_g = 0$	$\sigma_g^2 \sim IG(\nu/2, \nu\tau/2$	$y_{gi} \sim t_\nu(0, \tau)$

Table 4.3 presents the two alternative models resulting from different levels of shrinkage, and the implied marginal distribution of the data in each case. The fact that in both alternative models is possible to marginalize out the hierarchical parameters to get the marginal distribution of the data, is key to developed a useful statistic. We refer as M_μ to the model that treats the signals as means, and M_σ to the model that treats the signals as variances.

We can obtain the maximum likelihood estimates for the parameters in each model options, then with point estimates for the hyperparameters we can compute the observed likelihood function under each model. Finally, we use the ratio of the observed likelihood between the models as a statistic to distinguish when model (4.2) is going to fail. This lr statistic is described in equation (4.4).

$$lr = \frac{1}{\sum_g n_g} \sum_g \sum_i \log t_{\hat{\nu}}(y_{gi} \ ; \ 0, \sqrt{\hat{\tau}}) - \log N(y_{gi} \ ; \ \hat{\mu}_0, \hat{\gamma}) \tag{4.4}$$

where $\hat{x}$ represents the MLE of x and $\gamma = \sigma_0^2 + \tau$.

Clearly lr is not a likelihood ratio *test* statistic, since there is no comparison of an unrestricted estimation with estimates under a null hypothesis. Neither a Bayes factor, since we do not integrate with respect the hyperparameter priors. However, the interpretation might work

in similar way, when lr is large this means the model that treat the signals as variances and completely shrink the group means have more likelihood than the alternative.

4.4.2 Identify performance boundary

An extension of the previous simulation study is performed, with more scenarios for signal strength around the values where the model performance seems to drop dramatically. We define 33 combinations of values for m, w and T. The signal strength is control by m taking values between 2 and 4, the sparsity by w taking values between .9 and .99, finally the noise by T taking values 1, 4, and 9. We use almost the same data generation scheme than in Section 4.3, with a small change in the way we set the signals mean value. Here we use $m_g = m\sqrt{T}$, so m can be interpreted as the signal strength measured in terms of standard deviation.

Figure 4.5 presents ROC curves when the sparsity level is set as 95%, facets represent noise level and line color the signal strength level. In the previous simulations only have $m = 2$ and $m = 4$ scenarios, now this plot suggest a more clear relationship among the signal strength (line color) and the model performance. It seems to be a phase change around $m = 3$, when signal strength get bigger than that level the model performance decay very quickly. This effects occurs for any noise level.

We want to determine if the lr measure from equation (4.4) is informative to detect the confounding problem described in section 4.3. High values of lr implies the likelihood of M_σ is bigger than M_μ, i.e. treating the true signals as variances is more likely than treating the signals as means. As a measure of model performance, we summarize each ROC curve with the true positive rate (TPF) fixing the false negative rate at 0.1, this gives us an index from 0 to 1 on how good the model performs.

In addition to lr, two other statistics are computed in each scenario. The proportion of groups with small value of the sample mean, and the ratio between standard deviation in group means over mean of group standard deviations, i.e.

$$pr = \frac{\sum_g \bar{I}(Z_g<1)}{G} \qquad rt = \frac{\text{sd}(\bar{Y}_g)}{\sqrt{\sum_g S_g^2}} = \sqrt{\frac{\sum_g(\bar{Y}_g-\bar{Y})^2}{\sum_g S_g^2}}$$

where $(\bar{Y}_g, S_g^2)$ are the group sample mean and variance, and $Z_g = \bar{Y}_g/\sqrt{S_g^2}$.

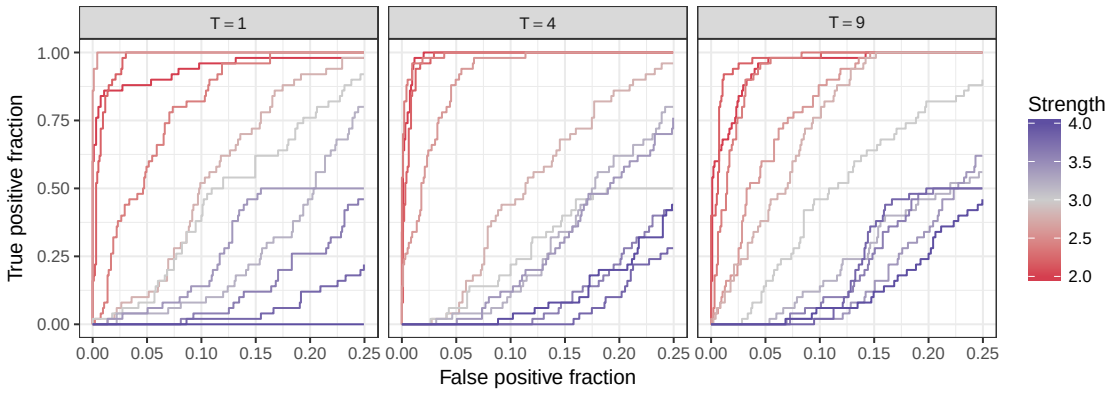

Figure 4.5: Results for Normal-IG model in simulations close to the phase shift boundary. ROC curves when 5% have true mean different from zero. $w = 0.95$, color is signal strength and facets are noise level (T)

The pr statistic is related to the sparsity in the data set, the normal model should perform worse when this statistic is large.

The rt corresponds to the ratio of between group variance over within group variance in square root. It indicates how variable the group means are, with respect to the internal variability within each group. When rt is large a normal distribution might have problems to accommodate the mean signal information.

Figure 4.6 shows scatter plots of lr value and rt value against pr for each scenario, color represents the model performance. The left panel suggests an association between lr and model performance, when $lr > 0.03$ model (4.2) results in very poor detection rates in every scenario (lower to 25%), except for 3 points close to this border. On the other hand, it seems that very small values of lr are associated with good detection rates.

The boundary seems to be relatively sharp, suggesting a phase shift at some cut point. There are some darker points that correspond to scenarios with really high sparsity. Those scenarios imply very few groups with means truly different from zero making the problem more variable. For instance, a 99% sparsity level imply that only 10 groups have $m_g = m$, each with 4 observations.

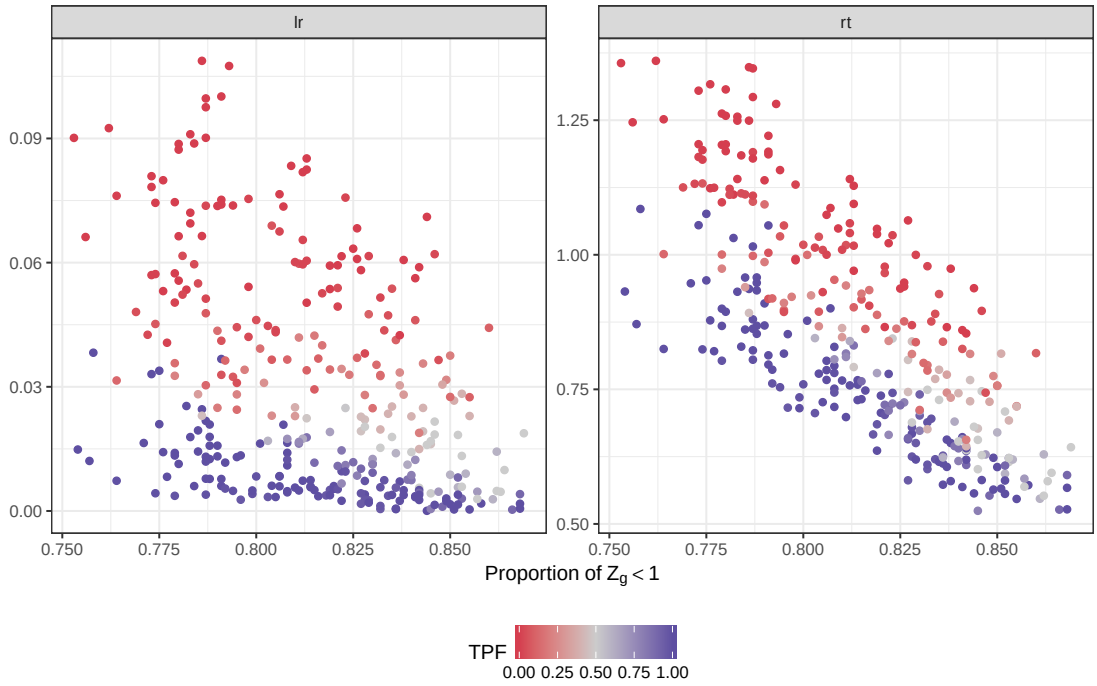

Figure 4.6: Scatter plots of lr statistic (left panel) and mean-variance ratio statistic, rt, (right panel) against proportion of scaled sample means lower than 1 (pr). Color indicates the true positive rate (TPF) when the false positive fraction is 10%.

Right panel presents a similar picture, there is a boundary between high and low performance scenarios. However, there is no horizontal level of rt, the boundary corresponds to a line with negative slope. For lower sparsity values, the between group variability increase since more groups would have sample means apart from zero, but within group variability does not change since the noise level in the data is the same.

4.5 Heavy tails to avoid confounding

This Section explores alternative models to understand when strong mean signals are confounded with variability. In every case the data model is $y_{gi} \sim N(\mu_g, \sigma_g^2)$, what distinguish the different options are the hierarchical distributions for μ_g and σ_g^2.

Cauchy distribution for the group means is used because its heavy tails. In scenarios where there are only few groups with means different from 0, but those group means are large in standard deviation units, the normal distribution would be dominated for groups with zero mean and its light tails cannot accommodate the non-zero groups, therefore the signal for the

non-zero groups might be interpreted as a variance signal instead. If this explanation is right, then using a hierarchical distribution with heavier tails should avoid the confounding issue.

We also include an alternative hierarchical distribution for variances, to test the possibility that the confounding is drive by the inverse-gamma choice. Log-normal distribution is another popular distribution for variances.

Table 4.4: Combinations of hierarchical distributions for (μ_g, σ_g^2)

Model name	Means (μ_g)	Variances (σ_g^2)
Nr-IG	$N(\mu_0, \sigma_0^2)$	$IG(\frac{\nu\tau}{2}, \frac{\tau}{2})$
Nr-LN	$N(\mu_0, \sigma_0^2)$	$LN(\tau, \nu)$
Ca-IG	$Ca(\mu_0, \sigma_0)$	$IG(\frac{\nu\tau}{2}, \frac{\tau}{2})$
Ca-IG	$Ca(\mu_0, \sigma_0)$	$LN(\tau, \nu)$

Table 4.4 shows the alternative models used for simulated datasets. There are four combinations for the hierarchical mean and variance distribution, combining normal or Cauchy distribution for mean signals, with inverse-gamma or log-normal for variances. Additionally, we estimate with a conjugate normal-inverse gamma model, i.e., having $\mu_g \sim N(\mu_0, \frac{\sigma_g^2}{\sigma_0^2})$ and $\sigma_g^2 \sim IG(\nu\tau/2, \tau/2)$.

Prior distributions for the hyperparameters μ_0, σ_0, and ν are the same in all models and equals to model (4.2), similarly τ prior is the same when variances are modeled with an IG hierarchical distribution and $\tau \sim N(0, 10)$ for lognormal models. This is due to the different role that τ plays in each distribution, related to the overall mean of the group-specific variances in IG and related with the log of that mean in LN distribution. The rest of the method to obtain inference from all four models are the same as the described earlier, except that models using Cauchy as hierarchical distribution need a thinning value of 5 in order to keep all potential reduction factor statistic below 1.1 threshold.

Figure 4.7 shows ROC curves for the two additional models using normal hierarchical distribution for the mean signal. Column facets represent the signal strength, the top rows shows results for the conjugate normal model, cjNr-IG, and bottom row shows results for normal - lognormal combination, Nr-LN. This complements the normal-inverse gamma results already presented in Figure 4.3. The results suggest that confounding effect of strong signals is present

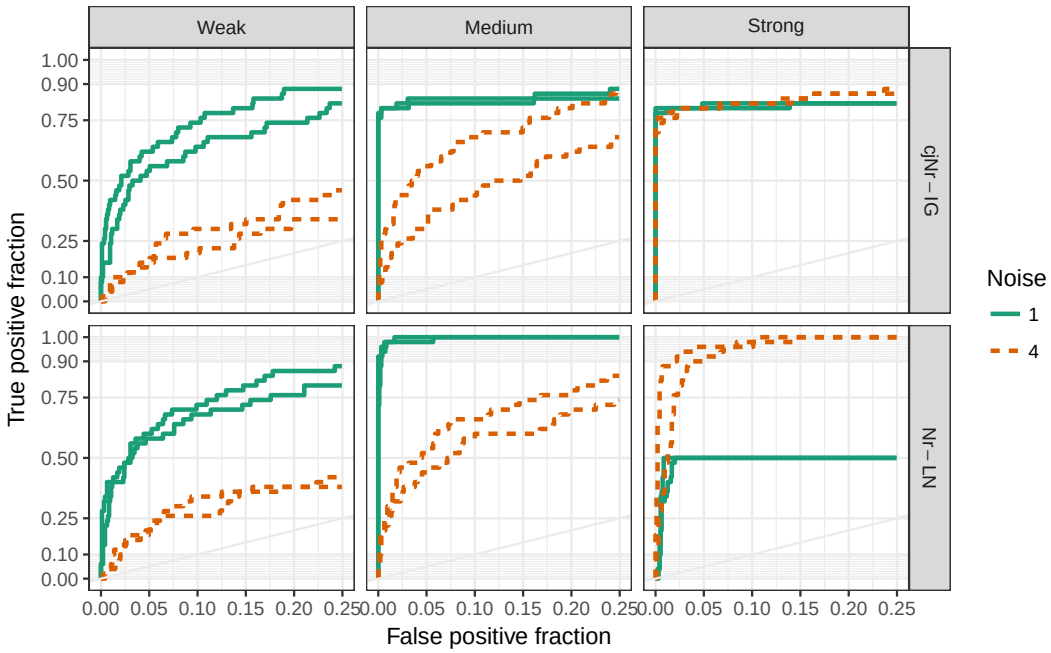

Figure 4.7: ROC curves for normal-lognormal, Nr-LN, and conjugate normal-inverse gamma (cjNr-IG) models in simulated data. Column facets represent signal strength (m) and row facets correspond to the hierarchical model, line color represents noise level.

in the log-normal case. Using a conjugate normal - inverse gamma model does not present very good results in any case, but it seem the confounding is somewhat mitigated. In this option, the hierarchical model for means is conditioned to its group variance, so it could be possible this makes harder to accommodate mean signals as variances. Figure 4.8 shows ROC curves of models using Cauchy as hierarchical distribution for the mean signals. Here there are no confounding effects among mean signals as variability, the tails of the Cauchy distribution are heavy enough to accommodate very strong means in sparsity context. None of the panels suggest the model gets better performance when more variability is present in the data.

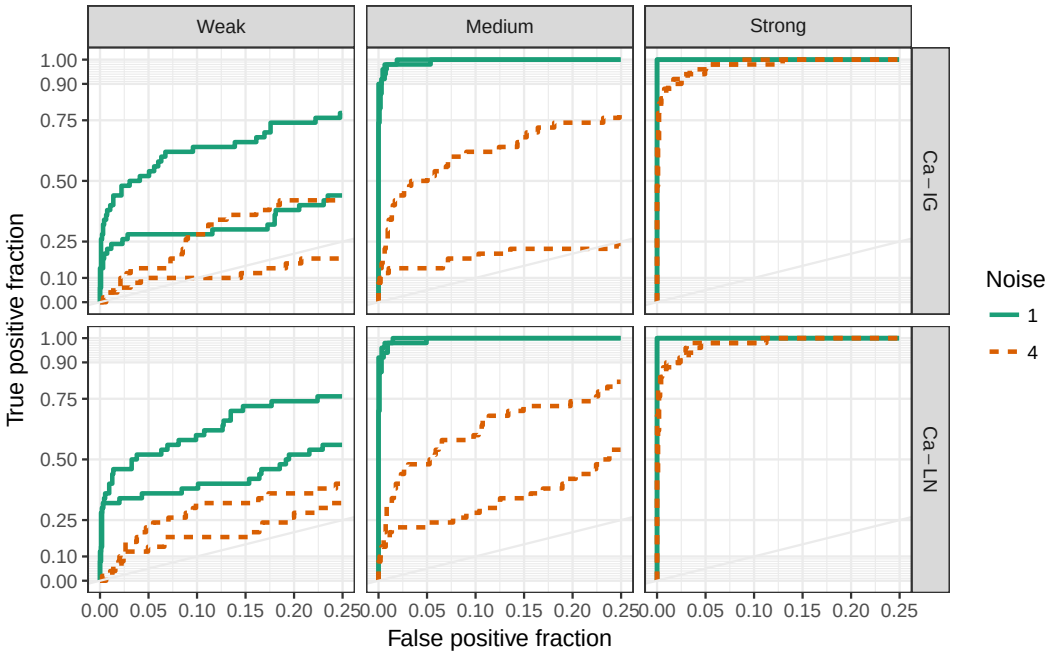

Figure 4.8: ROC curves for Cauchy-lognormal, Ca-LN, and Cauchy-inverse gamma (Ca-IG) models in simulated data. Column facets represent signal strength (m) and row facets correspond to the hierarchical model, line color represents noise level. ROC curves for models using Cauchy distribution for means

4.6 An appropriate analysis of maize experimental data

This Section presents some specific details of the data set used as motivational example, and repeat the data analysis presented in Section 4.2 but using a statistical model without the confounding problem.

The RNA-seq data set with allele-specific counts used in this paper constitute a portion of the experimental data obtained by Paschold et al. (2012). Data set includes four replicate plants of a hybrid genotype (B73xMo17) distributed in 2 flow cell blocks and two allele count measures per plant. With Iliumna® technology the ASE information for approximately 16 thousand genes is obtained.

Recall figure 4.1 describing model (4.2) results and maize data set to detect genes with differentially expressed alleles. Those results have some indication of confounding effects that later were illustrated using simulated data. The genes with the largest allele differences (in absolute value) present also the largest shrinkage of its means towards 0 and also the largest posterior variance.

The two statistics proposed above to diagnostic when a particular data set is susceptible of mean-variance confounding are lr and rt statistics. Computed using observed allele differential data, we obtain $lr = .06$ and $rt = 1.42$, both values in regions where using normal hierarchical distribution result in mean-variance confounding and poor detection rates (see Figure 4.6). This seems to confirm that we should not use a normal as μ_g hierarchical distribution. It need to be taken carefully though, since observed data are not similar to the simple simulated data set.

Previous Section suggested that use a Cauchy as hierarchical distribution for mean would alleviate the confounding effects. Therefore, a new hierarchical model with shrinkage in means and variances is used solely changing the μ_gs hierarchical distribution from normal to Cauchy.

Figure 4.9 shows the results of using Ca-IG model on allele differences data. Top row panels show posterior expectation against observed means while bottom row facets show the posterior standard deviation in vertical axis instead. Column facets split genes in non-DE and DE groups corresponding to a 95% credible interval does contain zero value or not. As in

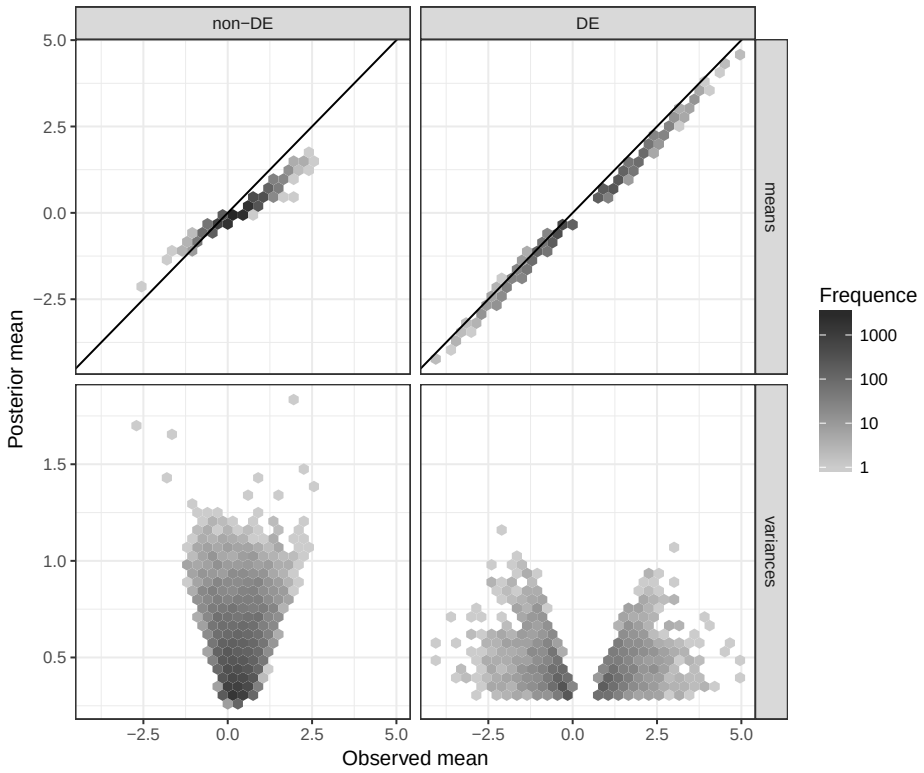

Figure 4.9: Cauchy - inverse gamma model results for ASE counts of Paschold et al. (2012) hybrid data. Bivariate histograms of posterior expectation of group mean against group sample mean (top facets), and square root of posterior expectation of group variance against group sample mean (bottom facets). Column facets indicates genes has its alleles differentially expressed (DE) or not (non-DE).

simulations, using Cauchy seems to remove confounding or unexpected shrinkage effects, all genes classified as non-DE corresponds to observed sample means close to zero.

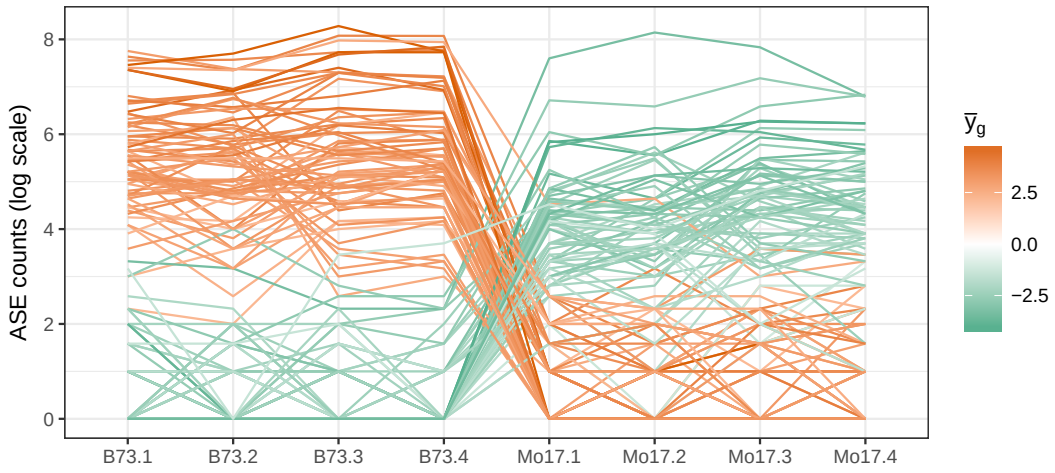

Figure 4.10: Parallel coordinate plot of ASE count profile for genes identified as DE using Cauchy hierarchical distribution but non-DE with normal. Color represents the sample mean difference between alleles.

There are around 500 genes flagged as DE by Cauchy model and non-DE by the normal model. Figure 4.10 shows a parallel coordinate plot ASE patterns of those genes. Each line in the plot represent a gene, the color intensity indicates the difference among alleles is large for these genes, which is also visible in the plotted count patterns. Most of the genes flagged only by the Cauchy distribution present a really low ASE count in the less expressed allele, having at least some replicates with zero read count. One advantage of RNA-seq technology is the detection of genes with low expression levels that could be biologically relevant (Paschold et al., 2012). Is then important ensure the statistical model does not miss comparisons involving these low expressed genes.

As it was mentioned in Section 4.2 it might be better to directly model the observed counts for each allele, instead of the log transformed allele ratio. In chapter 2 a Poisson-lognormal mixture model is proposed to model this data set. This model can be described as a generalized

linear model with Poisson data model, log link function and a normally distributed linear predictor term. The response variable is the observed allele abundance and linear predictor expectation contains blocks and replicate effects. Applying the model described in chapter 2, is possible to compute an allele difference for each gene and detect genes with differentially expressed alleles.

Inference for the Poisson-lognormal mixture model is obtain with `fbseq` package, the priors are the same described in the previous Chapter. The model is run two different times, using normal and Cauchy as alternatives hierarchical distribution for the gene specific regression coefficients

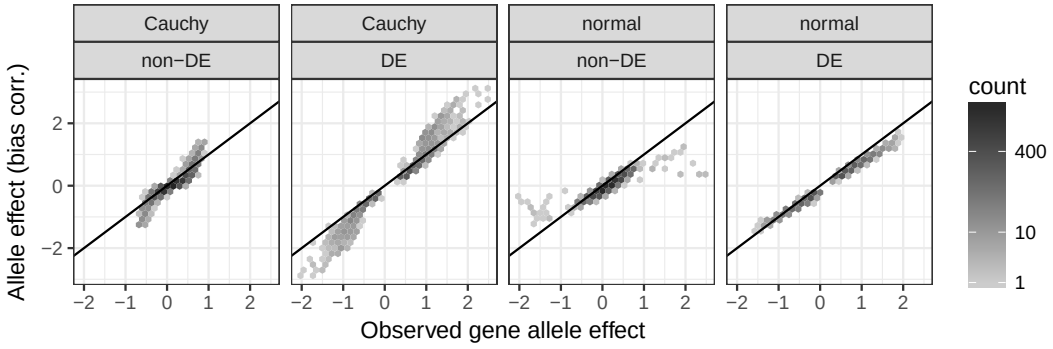

Figure 4.11: Bivariate histograms of posterior expectation of allele difference and sample allele difference using Poison-lognormal mixture model for Paschold et al. (2012) data. Facets represent a combination of the hierarchical distribution of regression coefficients (normal or Cauchy) and genes consider DE or not-DE.

Figure 4.11 the results from Poisson-lognormal mixture model with Paschold et al. (2012) data. Each panel in the Figure shows a bivariate histogram of the posterior expectation for

the gene-specific regression coefficient associated with the allele difference and the sample gene-specific mean difference between alleles. The facets combines the hierarchical distribution of the regression parameters and if the genes are flagged as DE or not (based on a credible interval).

Figure 4.11 suggests that again with this model, confounding effects seems to be present when a normal distribution is used. A similar patterns than observed in figures 4.1 and 4.9, the normal distribution shrink towards zero the genes with largest sample mean difference, it can be checked the overdispersion effects in these genes are really large (not shown). The extreme shrinkage disappear when a Cauchy is used as hierarchical distribution for the gene-specific regression coefficients.

4.7 Discussion

Double shrinkage models are used when there are two set of parameters with hierarchical distributions. For instance, cases in which is needed to model group means and group variances, or group specific regression coefficients with group specific overdispersion levels. Typically, only one set of parameters is relevant while the rest are nuisance parameters.

Hierarchical modeling of nuisance parameters, like group variances, can affect the inference about the parameters of interest, the group means. In sparse scenarios, with a large number of groups but only a few of those with true non-zero means, the double shrinkage Bayesian model might perform poorly. The bad performance is due because the mean signal might be treated as variability, groups with true large means present posterior means are shrunk towards zero and large posterior group variances.

This confounding effect occurs not only in the simulated data set with a very simple structure of means and variances vectors. It seems to be present in a much more complex real RNA-seq data set, and also when a more sophisticated model is used.

The main reason for the confounding effect is related to the light tails of the normal distribution. When a normal distribution is used as hierarchical distribution for the group means, the few groups with strong signals are mistakenly shrunk towards zero. Changing the hierarchical distribution to Cauchy seems to solve this issue.

Further work is needed to analytically obtain the form of the model performance boundary

at which the mean signal strength starts being treated as variance, also a simulation studies in more complex scenarios such as more groups or simulations closer to the RNA-seq data example, might help to characterize these confounding effect.

CHAPTER 5. An approximate Bayesian estimation of a Poisson Markov random field model for crash data

Abstract

In this paper we propose Winsorized Poisson Markov random field (WPMRF) approach to model crash frequencies. WPMRF model is used to introduce spatial dependence at observations level, instead of using latent spatial random effects. The neighborhood structure in our approach is anisotropic, to allow for the spatial dependence parameters to change for different directions. We propose a measure of crash risk for each individual intersection, based on the posterior predictive distribution. Departments of Transportation may use this measure to identify intersections with high risk, as complementary to more commonly used indicators as traffic volume or crash history.

WPMRF model introduces computational challenges since the associated posterior distribution is doubly intractable. We use approximate Bayesian computation (ABC) to perform parameter estimation in a WPMRF model. We calibrate the ABC algorithm via a simulation study and apply it to make inference of the WPMRF model using crash frequency data in three cities of Iowa (Ankeny, Marshalltown, and Tipton).

In Marshalltown, we find the spatial dependency of the number of crashes is significative in both directions, North-South and East-West, even after conditioning by covariates. In all three cities we found high risk intersections with low traffic volume or crash history. These intersections would not be consider for intervention using information from traditional indicators.

5.1 Introduction

The modeling of crash frequency data is a field of extensive research. At the national level, motor vehicle fatalities account for approximately 30% of all injury deaths in the United States every year. As a consequence, the cost to society due to years of life lost are enormous, estimated to be approximately $150 billion per year by the National Highway Traffic Safety Administration (NHTSA). Only cancer and heart disease are comparable in being responsible for loses of similar magnitude. In Iowa, about 400 lives are lost annually in traffic accidents and crashes represent a total cost of 1 billion dollars per year (McDonald, 2012). Therefore, preventing crashes or at least minimizing the loss of life and major injuries due to crashes is critically important.

Due to budgetary constraints in Departments of Transportation nationwide, one of the main tasks of traffic engineers is the identification of hot spots, i.e., the sites with potential risk.

In some cases, sites are ordered according to the mean crash frequencies observed over a few years. This can lead to candidate lists that are highly dependent on traffic volume and do not account for the uncertainty inherent in sites' expected crash numbers (Zegeer, 1986).

The main subject of this paper is to model crash frequency at the intersection level while introducing spatial correlation among intersections. Poisson Markov random fields (PMRF) have been used to model spatially correlated data and Winsorized PMRF for count data. (Kaiser, 2002; Kaiser and Cressie, 1997) The winsorization is introduced in order to allow positive correlation among the observed counts. There is a tractability issue with the Winsorized PMRF model since the joint probability function is only known up to a normalizing constant that depends on the parameter set to estimate.

Approximate Bayesian computation (ABC) is a field of Bayesian research that has gained much popularity in recent years. This constitutes a powerful estimation technique based on simulations, these methods are designed for complex problems where the likelihood is computationally or analytically intractable.

The use of ABC in crash frequency models is attractive in at least two aspects. Firstly, it allows to fit a Poisson MRF, which makes possible to model crashes at each intersections.

Commonly used methods, as Song et al. (2006) use a CAR (*conditionally autoregressive*) model with spatially correlated random effects. However this approach does not allow for zero counts so it could not be directly applied to intersection level data. Secondly, it extends the fields where ABC inference is applied. Despite the active work in ABC in recent years, so far its application is still focused in a few topics as genetics, biology or epidemiology.

Next section presents the real spatial data we are using, section 5.3 presents the Poisson model using the Winsorization approach, and section 5.4 describes the approximate Bayesian computation methodology used for inference. Results from simulation study and real data example are shown in sections 5.5 and 5.6 respectively. Finally, some discussion is presented in section 5.7.

5.2 Spatial data

We present methods to work areal-referenced count data. The response variable is a discrete variable which is available in a set of locations, while the covariates are continuous variables also available at location level. Spatial dependence is introduced through the neighborhood structure, locations might be connected or not, according the neighbor structure.

We applied the proposed model to real data examples, Iowa Department of Transportation provide crash and traffic data for Tipton, Marshalltown and Ankeny on a ten years period (2004 - 2013). The response variable represents the number of crashes in cities intersections, and the continuous covariate represents the total traffic volume at that intersection. Figure 5.1 shows the intersections layout in each city, the color pattern represent the traffic volume. The main roads in Tipton and Marshalltown runs in North-South direction while Ankeny shows roads with large traffic volume in both directions.

Connectivity information was also provided by Iowa DOT. Two intersections are considered as neighbors if there is a direct link between them, this corresponds with a rook structure in a regular lattice. Neighbors are classify in two directions according to its orientation. Any pair of neighbors intersections, if the distance in latitude is smaller than distance in longitude then are classified as East-West (EW) neighbors, and North-South (NS) otherwise.

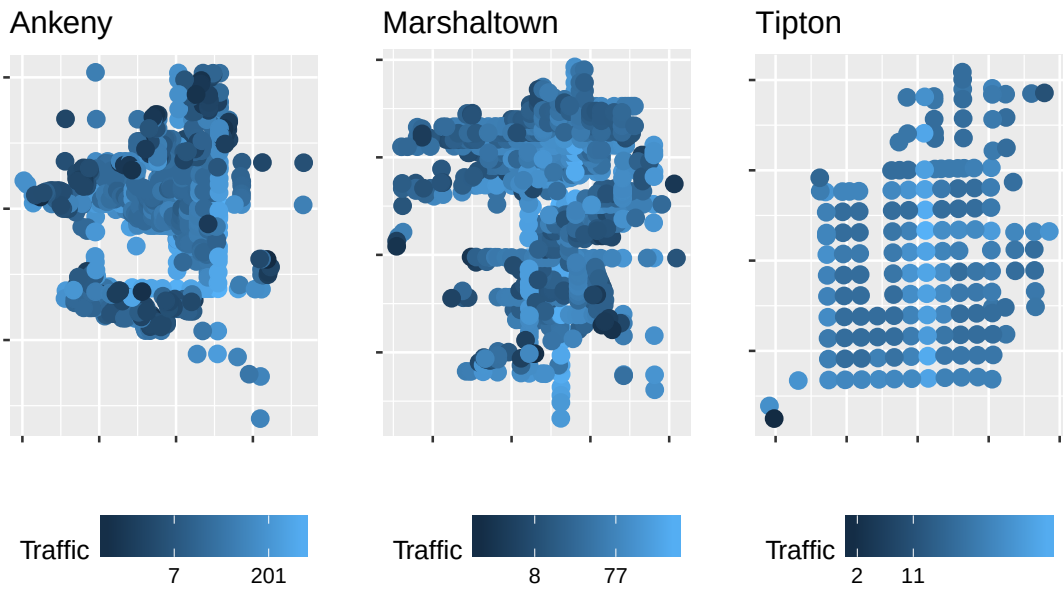

Figure 5.1: Traffic volume at each intersections in Ankeny, Marshalltown and Tipton

We use the Moran I statistic as a measure of spatial dependence,

$$I(Z) = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{n}(z_i - \bar{z})^2}$$

where Z is the variable of interest and $w_{ij} = 1$ if z_i and z_j are neighbors and 0 otherwise. We modify slightly this definition to obtain a statistic sensitive to neighbors direction. We obtain $I_{NS}(Z)$ by setting $w_{ij} = 1$ when z_i and z_j are neighbors in the NS direction and 0 otherwise, similarly with $I_{EW}(Z)$.

Table 5.1: Summary statistics of Iowa crash data

Town	Crashes				Total traffic	
	Mean	Q90	I_{EW}	I_{NS}	I_{EW}	I_{NS}
AK	3.54	20	0.19	0.28	0.42	0.57
MR	3.27	12	0.18	0.46	0.16	0.78
TN	0.7	3	0.09	0.53	0.09	0.79

Table 5.1 shows some descriptive statistics, the mean number of crashes, the 90% quantile of the number of crashes among intersection with crashes, and the modified Moran statistic in each direction of the number of crashes and traffic volume. As was suggested by Figure 5.1 the traffic volume dependence is larger in the North-South direction in Tipton and Marshalltown, while Ankeny shows moderate dependence in traffic volume in both directions.

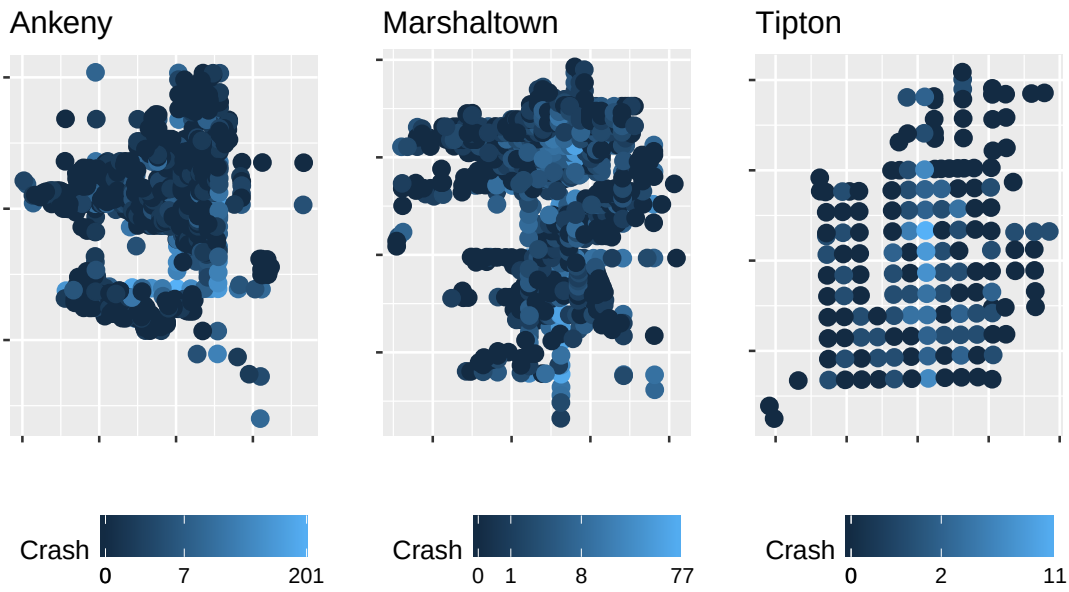

Figure 5.2: Number of crashes at each intersections in Ankeny, Marshalltown and Tipton

Figure 5.2 shows the observed crashes at each intersection. Spatial dependence in number of crashes seem to be moderate and positive in all cases except for Tipton in EW direction, in all cases dependence in the North-South direction is larger than East-West. This is consistent with the Moran statistic presented in Table 5.1. The mean number of crashes is relatively low, also comparing the maximum in Figure 5.2 with Q90 we see there are a few intersection with much larger number of crashes than the rest, and this happens in the three cities.

5.3 The (Winsorized) Poisson Markov random field model

We propose a Markov random fields (MRF) approach to model the spatial data presented earlier. MRF are the basis for several statistical models that include spatial (or spatiotemporal) dependence. This section present MRF for Poisson random variables and the Winsorization approach to allow positive spatial correlations. The section ends with a description of the specific model we use for Iowa DOT crashes data.

5.3.1 Markov random fields

Let $y(s_i)$ the response variable for location s_i, N_i the set of neighbors of s_i, and $y(N_i) \equiv \{y(s_j) : j \in N_i\}$, i.e., the set of response values in the neighbors of s_i. The Markov property consist in that the distribution of the response at one location only depends of its neighbors, i.e., $p(y(s_i)|\theta, y) = p(y(s_i)|\theta, y(N_i))$.

The joint distribution is constructed based on the *neg-potential function* $Q(y|\theta)$ (Besag, 1974; Kaiser and Cressie, 1997). Following the notation of Kaiser and Cressie (1997) we write $Q(y|\theta) \equiv log\left[\frac{p(y|\theta)}{p(y_0|\theta)}\right]$, for y_0 being any arbitrary (fixed) value and $y, y_0 \in \Omega$. The joint density is obtained as $p(y|\theta) = \frac{\exp[Q(y|\theta)]}{\int_\Omega \exp[Q(t|\theta)]dt} = \frac{\exp[Q(y|\theta)]}{k(\theta)}$, and then the posterior distribution for θ can be written as

$$\pi\left(\theta|y\right) = \frac{1}{k(\theta)} \frac{e^{Q(y|\theta)}\pi(\theta)}{p(y)} \tag{5.1}$$

where $\pi(\theta)$, $\theta \in \Theta$ is the prior distribution of θ and $p(y) = \int_\Theta \left[exp\left[Q(t)\right]\pi(\theta)/k(\theta)\right] d\theta$. Standard MCMC methods are designed to deal with the intractability of $p(y)$, a second intractable integral, $k(\theta) = \int_\Omega e^{Q(t|\theta)}dt$, makes these models to be known as doubly intractable (Murray et al., 2006).

There are several approaches to deal with the intractability within the likelihood function like the pseudo-likelihood approach (Besag, 1974) or auxiliary variable methods (Møller et al., 2006; Murray et al., 2006) and more recently approximate Bayesian computation (Grelaud et al., 2009; Everitt, 2012). We use an approximate Bayesian computation which is described in the next section.

5.3.2 Poisson Markov random fields

For single parameter exponential families, Besag (1974) proposed the so-called auto-models where the natural parameter is expressed as a function of the response values in the neighborhood of a location and thus introduce (spatial) dependence.

$$\begin{aligned} p\left(y(s_i)|y(N_i)\right) &= exp\left\{\sum_{k=1}^{s} A_{i,k}[z(N_i)]T_k\left(y(s_i)\right) - B_i[z(N_i)] + C_i[z(s_i)]\right\} \\ A_i\left[y\left(N_i\right)\right] &= \alpha_i + \sum_{z(N_i)} \eta_{i,j} y(s_j) \end{aligned} \tag{5.2}$$

where $A_i\left[y\left(N_i\right)\right]$ is the natural function that depends on the neighbors and this dependence is captured in the $\eta_{i,j}$ parameters. Based on the conditionals distribution is possible to construct the neg-potential function and then have a valid MRF model.

Kaiser and Cressie (1997) presents the Poisson auto-model i.e. when all conditionals are Poisson distributed, obtained by setting $T\left[y(s_i)\right] = y(s_i)$, $B_i\left[y(N_i)\right] = exp\left\{A_i\left[y(N_i)\right]\right\}$, $C_i\left[y(s_i)\right] = -log\left[y(s_i)\right]$, and $\log(\lambda_i) = A_i\left[y\left(N_i\right)\right]$. An explicit construction of $Q(y|\theta)$ functions shows there are two restrictions for the dependence parameter: symmetry ($\eta_{i,j} = \eta_{j,i}$) and non-positivity ($\eta_{i,j} \leq 0,\ \forall i,\ j$) (Kaiser and Cressie, 2000). This means that the model, as it is, does not allow for positive spatial dependence.

5.3.3 Winsorization approach

One approach to overcome the non-positivity restriction just mentioned is to use Winsorization (Kaiser and Cressie, 1997).

Starting with a random variable with Poisson distribution, $\tilde{Y} \sim Poi(\lambda)$, a Winsorized version of $\tilde{Y}$ is defined by combining all the mass beyond a Winsorization point R into a point mass at that particular point: $Y \equiv \tilde{Y} \cdot I(\tilde{Y} \leq R) + R \cdot I(\tilde{Y} > R)$, with $R < \infty$ and $I(.)$ denoting the indicator function.

The probability mass function of the Winsorized variable Y can be written as the sum of the regular Poisson variable $\tilde{Y}$ limited to $Y < R$ and has the point mass in the Winsorization point R:

$$P(Y = y|\lambda, R) = \left[\frac{\lambda^y}{y!}exp(-\lambda)\right] \cdot I_{y \leq R} + \left[1 - \sum_{t=0}^{R-1} \frac{\lambda^t}{t!}exp(-\lambda)\right] \cdot I_{y=R}, \quad y \in \{0, 1, ..., R\}\,. \quad (5.3)$$

Kaiser and Cressie (1997) derive two significant results. First, the expected value $E\left(Y|\lambda, R\right)$ is strictly increasing in λ, and (more importantly) $E(Y) \approx E(\tilde{Y})$ when R is large ($R \geq 3\lambda$).

Putting all of this into the context of this paper and using $\log(\lambda_i) = A_i\left[y\left(N_i\right)\right]$, we can

write a spatial formulation as presented in (5.4)

$$\begin{aligned} p(y(s_i)|y(N_i)) = &\quad \exp\left\{\log(\lambda_i)y(s_i) - D_i(y(N_i)) - \log(y(s_i))\right\} \\ D(y(N_i)) = &\quad \begin{cases} \lambda_i & \text{if } y(N_i) \leq R-1 \\ \lambda_i e^{-\psi_i} & \text{if } y(N_i) = R \end{cases} \end{aligned} \tag{5.4}$$

We will refer to model (5.4) by writing $Y_i|N_i \sim WP((\lambda_i, R)$, i.e., response at each location is conditionally distributed as Winsorized Poisson given the response on its neighbors locations. The upper bound R is supposed to be known.

5.3.4 Modelling crashes data

We describe the specific model used for intersection crash data. Each location represents one intersection, and $y(s_i)$ represents the number of crashes occurred at intersection s_i, we use the model (5.4), $Y_i|N_i \sim WP((\lambda_i, R)$, with the natural parameter as function of the traffic covariate and the response in the neighbors locations. The covariate X_i represents the total traffic in the intersection s_i. We assume anisotropic dependence among intersections, allowing different effects of East-West (EW) and North-South (EW) type of neighbors.

Equation (5.5) shows the model used for analyzing crash data. Conditionally on neighbors, the total intersection crashes are modeled with a Winsorized Poisson distribution and its expected value with an identity link parametrization.

$$\begin{aligned} &y(s_i)|N_i \sim WP(\lambda_i, R) \\ &\lambda_i = \beta_0 + \beta_1 X_i + \textstyle\sum_k \sum_{j \in N_{i,k}} \eta_k \left[y(s_j) - \beta_0 - \beta_1 X_j\right] \end{aligned} \tag{5.5}$$

Log-link and centered parametrization have been proposed (Besag, 1974; Kaiser et al., 2012). These options are unstable; they only work when η_k are restricted to a small neighborhood around zero and even then, they can only account for small positive dependence in the data (see section 4.3.1 of Griffith and Paelinck (2011)).

The ultimate goal is to obtain a risk measure for each individual intersection, such measure should be based in the posterior predictive distribution $p(y(s_i)|y_{obs})$ where y_{obs} represent the observed crash data.

A natural measure could be the posterior predictive expectation. However, the effect of the traffic covariate could be very influential in a measure like the posterior expectation (at least in this model with only one covariate). In addition the scale of this measure would be relative to each city, it would be nice to have a measure with the same scale for all cities. Due to dependence among the intersections, an intersection might show high posterior expected value because is located in a neighborhood with high level of crashes.

Complementary to the posterior expectation, we propose to use the right tail posterior predictive probability of the observed count as a risk measure, i.e.,

$$R_i = P(y(s_i) > y_{obs}(s_i)|y_{obs}) \tag{5.6}$$

Intersections where this probability is high can be considered of high risk and then signaled for intervention. Risk measure R_i scale is not city dependent, might be uncorrelated with traffic volume since is relative to the observed count, and flag individual intersections that present larger number of crashes than expected.

5.4 Approximate Bayesian Computation (ABC)

Approximate Bayesian Computation is a family of techniques to obtain samples from a posterior distribution in complex models, where the likelihood function is not analytically or computationally tractable. As we mention in the previous section, the posterior distribution in (5.1) is double intractable due to the presence of the computationally intractable integrals $k(\theta)$ and $p(y)$. This creates serious challenges to make Bayesian inference in dealing with a Winsorized Poisson MRF model. We take an ABC approach to tackle down the double intractability.

One of the first ABC method was proposed by Pritchard et al. (2000) adapting a Rubin ideas into a rejection algorithm. The heuristic of this rejection algorithm consist in to obtain simulated data from the likelihood and to keep the parameter values that produce simulated data close to the observed data. Marin et al. (2012) presents a review of several variants and improvements that has being proposed in the last 15 years.

The discrepancy between observed and simulated data is typically measured trough a summary statistic $S()$ and a distance function $\rho()$. We can write the rejection algorithm for ABC in a nearest neighbor fashion, as in 1.

Algorithm 1 ABC-Rejection sampler

1. Compute $s_0 = S(y_{obs})$
2. For $(i \in 1:N)$
   - generate $\theta_i \sim \pi(\theta)$, and $y_i^* \sim f(y|\theta_i)$
   - compute $s_i = S(y_i^*)$ and $d_i = \rho(s_i, s_0)$
3. Return $\left\{\theta_i : d_i < d_{(k_N)}\right\}$, the k_N-nearest neighbors of s_0

Where y_{obs} represents the observed data set, N is the total of simulated parameter values, k_N represents the samples to approximate the posterior distribution. Usually, k_N is defined relative to N, for instance $k_N = 0.01N$ keep the 1% simulations closest to the observed data. The output of algorithm 1 is $\{(\theta_i, s_i)\}_{i=1}^{k_N}$ a random sample from the joint density of $(\theta, S(y))$ restricted to a neighborhood of s_0 (Biau et al., 2015). Based on this sample, we might estimate the posterior $p(\theta|s_0)$ using kernel smoothing tools.

5.4.1 Introductory ABC examples

This section presents some simple examples of how a posterior distribution can be obtained without evaluating the likelihood function. There are two basic ABC issues treated in this Section that might help to understand how ABC works. First, a rejection ABC example shows the relevance of using a summary statistic instead of the observed data to decide witch simulations are rejected. Secondly we illustrate how modelling the relationship between simulated parameters and summary statistic improve the posterior approximation.

Let suppose we observe only one data point, $y_0 = 7$ from a binomial data model, $y \sim Bin(20, \theta)$. Bayesian inference for this model could be obtained using the conjugate prior, $\theta \sim Beta(1, 1)$, in this case the posterior is simply

$$\theta|y_0 \sim Beta(7 + 1, 20 - 7 + 1).$$

A rejection algorithm to obtain samples from $p(\theta|y_0)$ can be written as

1. generate $\theta_k \sim Beta(1,1)$ and $y_k \sim Bin(20, \theta_k$ for $k = 1, \ldots, K$
2. keep θ_k if $y_k = y_0 = 7$

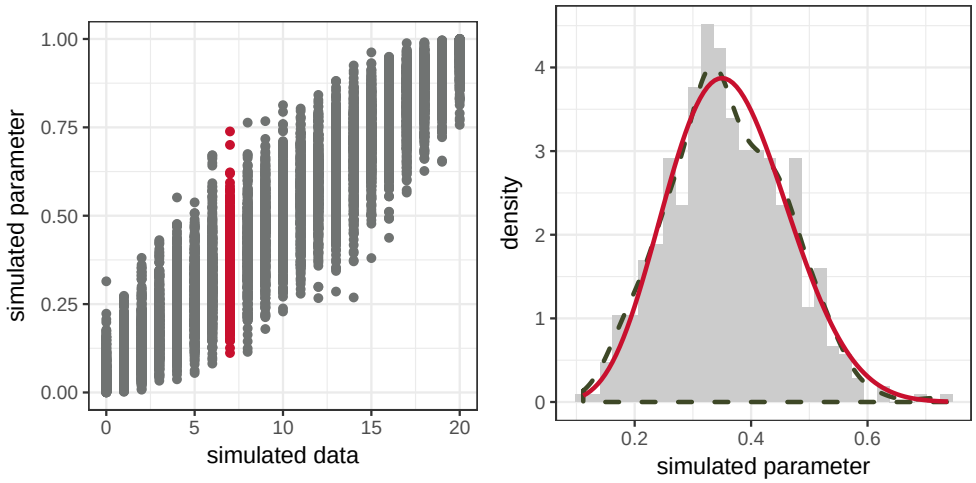

Figure 5.3: Left panel: Scatter plot of the simulated parameter θ_k against the simulated observation, y_k. The highlighted points corresponds to simulations equal to observed data point, y_0. Right panel: histogram of θ_k values that result in simulated data $y_k = y_0 = 7$. The red curve is the true posterior distribution and the dashed line is a kernel estimate.

Figure 5.3 presents the results of apply the rejection strategy with $K = 10000$. The left panel shows the simulated parameters and simulated observations, there are 501 values of θ_k for which $y_k = 7$. These 501 points are samples from θ posterior distribution. The right panel show an histogram of the selected values of θ_k and the true posterior is being approximated. Certainly the algorithm works properly, although more than 501 points are needed.

The previous example does not correspond to an example of ABC, the θ_k values come from the actual posterior distribution. However this approach is unpractical even in a slightly harder problem, as is illustrated with the next example. Lets continue with the same data model and prior as before, but observing 15 data points,

$$y = (3,6,5,3,3,2,4,6,5,2,5,8,4,3,2) \quad \hat{p} = \frac{1}{15}\sum_i \frac{y_i}{20} = 0.2033$$

where $\hat{p}$ represent the observed proportion in the observed data (the data were simulated independently from $Bin(20, 0.2)$ distribution). Again in this simple problem the posterior

distribution can be easily obtained

$$\theta|y \sim Beta(61+1, 239+1)$$

Simulating 100000 values from $Beta(1,1)$ and its corresponding simulated vector of length 15, there is no simulations matching exactly the observed data vector. Then the approximate solution is to accept simulated vectors y^k *similar* to the observed data. The following algorithm corresponds to an ABC rejection algorithm, we include two variants for computing the distance: between simulated and observed data, and between simulated and observed summary statistic.

1. For $k = 1 \ldots, K$:: Generate $\theta_k \sim Beta(1,1)$, and $y_i^k \sim Bin(20, \theta_k)$, for $1, \ldots, 15$
2. Compute distance:
   (a) $d_k^1 = \sum |y_i^k - y_i|/15$
   (b) $d_k^2 = |\hat{p}_k - \hat{p}|$
3. Keep θ_k corresponding to the smallest ϵK values of d_k

The tolerance ϵ represent the proportion of simulated values we keep to use as an approximate posterior sample. Figure 5.4 shows the results from the rejection ABC algorithm for different levels of tolerance and for the two ways of computing the distance among simulated and observed values.

There are two clear patterns illustrated by Figure 5.4. Reducing the tolerance level improves the approximation,for instance, keeping the 50% closest observations result in a poor estimate of the true posterior, ϵ value need to get as low as 1% in order to obtain a reasonable result. The second pattern is the effect is that better results are obtained when the summary statistic is used to compute the distance among simulated and observed values.

The final step in this list of introductory examples consists in how is possible to use a statistical model to improve the ABC result. The target of the approximation is the conditional distribution of the parameter θ given the observe statistic $\hat{p}$, then a model relating the simulated parameter with simulated statistic in each iteration is a way to correct for the discrepancy between simulated and observed statistic.

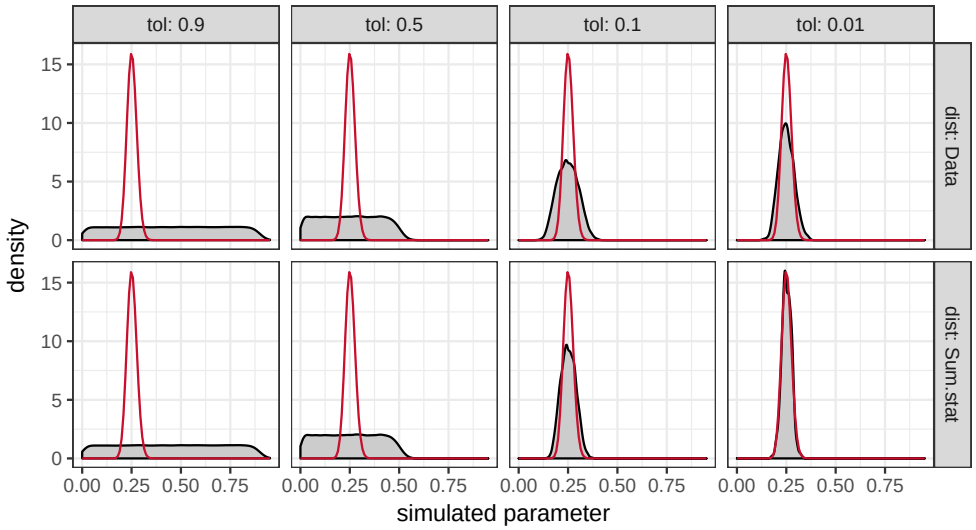

Figure 5.4: Results from rejection ABC algorithm in a binomial sample of 15 observations. In each facet panel, the red curve represents the true posterior distribution, and the grey area is the estimated density. Row facets indicate if distance d_k is computed between data or between summary statistic, the column facets represent the tolerance level.

Figure 5.5 present a scatter plot of the simulated parameters, θ_k against the simulated summary statistic, $\hat{p}$. Left panel show all simulations, the red points correspond to accepted simulations with rejection ABC method using $\epsilon = .1$ as tolerance, and the vertical line indicates the observed $\hat{p} = 0.203$ value. There is a strong linear relationship between θ_k and p_k in this example, we can correct the difference $|p_k - \hat{p}|$ by projecting the points along the regression line.

Assume a regression model $\theta_k = \beta_0 + \beta_1 p_k + e_k$, so the empirical residuals from this model are

$$\hat{e}_k = \theta_k - \hat{\beta}_0 - \hat{\beta}_1 p_k$$

and based on this empirical residuals is possible to construct an corrected simulated value as conditional expectation (given $\hat{p}$) plus the residual from the model.

$$\begin{aligned} \tilde{\theta}_k &= E(\theta_k|\hat{p}) + \hat{e}_k \\ &= (\hat{\beta}_0 + \hat{\beta}_1\hat{p}) + \hat{e}_k \\ &= (\hat{\beta}_0 + \hat{\beta}_1\hat{p}) + (\theta_k - \hat{\beta}_0 - \hat{\beta}_1 p_k) \\ &= \theta_k + \hat{\beta}_1(\hat{p} - p_k) \end{aligned} \tag{5.7}$$

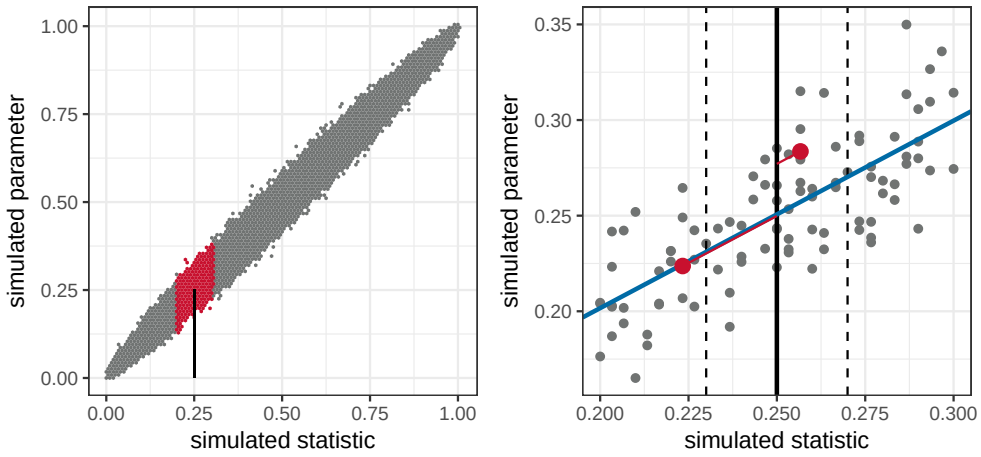

Figure 5.5: Left panel: relationship between simulated parameter and simulated proportion, red points corresponds to the acceptation region with $\epsilon = 0.1$. Right panel: some selected points from the left panel. Vertical line is the observed statistic, vertical dashed line is the rejection region, blue line is the regression line.

Last expression in equation (5.7) show the correction came from the difference between the observed summary statistic and each simulated statistic, and uses the relationship between simulated parameter and statistic trough $\hat{\beta}_1$. In this simple example a simple regression model works fine, but more flexible models are needed for other than toy examples.

Right panel of Figure 5.5 show a few selected points to illustrate the projection. The two red points are projected using the model. The main advantage of using the model is shown in Figure 5.6, this is analogous to Figure 5.4 but using the results from the regression model.

The sensitivity to the tolerance level is greatly reduced by using a model, even for $\epsilon = .9$ the approximation to the true posterior is not bad. The practical consequences are that it is possible to reduce the number of initial simulations and that the choice of the tolerance level is not so critical.

5.4.2 Kernel density estimation and ABC

The ABC approach we use in this paper rely in the kernel density estimation (KDE), therefore we comment the basics of KDE and its relation with ABC. Here we describe two approaches, first using a univariate KDE and secondly using a conditional KDE. In the next

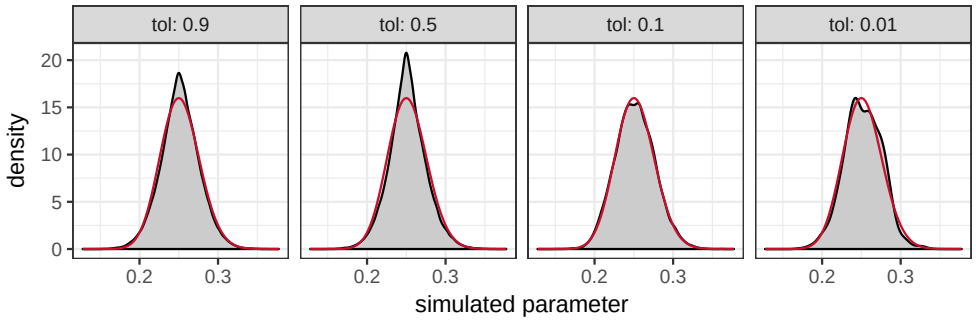

Figure 5.6: Results from rejection ABC algorithm plus a linear model correction, in a binomial sample of 15 observations. In each facet panel, the red curve represents the true posterior distribution, and the grey area is the estimated density. Column facets represent the tolerance level

subsection we describe the specific ABC methods we use in this paper to obtain the results.

Let $X_i \sim f(x)$ with $i = 1 \ldots, n$, the kernel estimate of f is $\hat{f}(x) = (1/nh) \sum_i K\left(\frac{|x_i - x|}{h}\right)$ where h is a pre-specified bandwidth parameter and $K()$ is a 2nd order, non-negative, symmetric kernel function (This means that $K(u) \geq 0$, $\int K(u)du = 1$, $\int uK(u)du = 0$, $\int u^2K(u)du = k_2 < \infty$ and $K(u) = -K(-u)$. Popular choices for $K()$ are Gaussian or Epanechnikov kernels). Applying this method to the output of algorithm 1, $\{(\theta_i, s_i)\}_{i=1}^{k_N}$, Biau et al. (2015) propose to approximate the posterior $p(\theta|s_0)$ as in equation 5.8 and derive convergence rates for that estimate.

$$\hat{p}_h(\theta|s_0) = \frac{1}{k_N h} \sum_{i=1}^{k_N} K\left(\frac{|\theta_i - \theta|}{h}\right) \tag{5.8}$$

The approximation 5.8 does not take into account the distance between s_i and s_0 after the k_N nearest neighbors are selected. Then is important that all k_N samples are close to s_0, in other words k_N/N needs to be small enough for the method to work. This restriction can be relaxed using KDE for the conditional density instead.

Let $(X_i, Y_i) \sim f(x, y)$ with $i, \ldots, n$ and we want to estimate $f(y|x)$. Hyndman et al. (1996) shows the classic KDE estimator might be improved using a regression model

First, we assume the model $y = m(x) + e$, where $m(x) = E(y|X_i = x)$ and $e_i \overset{ind}{\sim} g()$ with $E(e|X) = 0$. Next, using the fact $f(y|x) = g(y - m(x)|x)$ and an estimate for the conditional mean, say $\hat{m}(x)$, we can compute $e_i = Y_i - \hat{m}(X_i)$ to obtain a kernel estimate of $g(e|x)$ based

on the observations $((e_1, X_1), \dots (e_n, X_n))$ as $\hat{g}(e|x) = \sum_i w_i \frac{1}{h_2} K_{h_2}(|e_i - e|)$, then replacing the residuals by we obtain

$$\begin{aligned} \hat{f}(y|x) &= \sum_i w_i \tfrac{1}{h_2} K_{h_2}(|Y_i - \hat{m}(X_i) - y + \hat{m}(x)|) \\ &= \sum_i w_i \tfrac{1}{h_2} K_{h_2}(|\tilde{Y}_i - y|) \end{aligned}$$

where $\tilde{Y}_i = Y_i - \hat{m}(X_i) + \hat{m}(x)$, $K_h(u) = K(u/h)$ and $w_i = K(\frac{|X_i - x|}{h_1}) / \sum_i K(\frac{|X_i - x|}{h_1})$. So $\hat{f}(y|x)$ might be viewed as the kernel estimator using the definition but using the 'adjusted' values $\tilde{Y}_i$ instead of the Y_i, different ways for estimating the expectation function determines different adjustments.

Once again, let $\{(\theta_i, s_i)\}_{i=1}^{k_N}$ to be the output of algorithm (1) and we want to approximate the posterior $p(\theta|s_0)$. Beaumont et al. (2002) propose a local-linear model

$$\theta(d) = \alpha + d\beta(d) + \zeta$$

where $d = \rho(s, s_0)$. Applying the procedure described earlier, a sample from this conditional distribution is obtained by computing $\tilde{\theta}_i = \theta_i - \hat{\alpha} - d_i\hat{\beta}(d_i) + \hat{\alpha} + 0\hat{\beta}(0) = \theta_i - d_i\hat{\beta}(d_i)$, where $(\hat{\alpha}, \hat{\beta})$ are obtained minimizing $\sum_i (\theta_i - \alpha - d_i\beta(d_i))^2 K_\delta(d_i)$. Finally, based on the sample of $\tilde{\theta}$ the posterior distribution is estimated by

$$\hat{p}_{\delta_1, \delta_2}(\theta|s_0)(\theta|s_0) = \frac{\sum K_{\delta_2}(\tilde{\theta} - \theta) K_{\delta_1}(s_i - s_0)}{\sum K_{\delta_1}(s_i - s_0)} \tag{5.9}$$

With this approach simulated θ_i are weighted according to $\rho(s_i, s_0)$ and the local-linear regression corrects for the difference due to $E[\theta|s_i] \neq E[\theta|s_0]$. This corrections reduce the sensitivity of the choice of k_N. For instance, Marin et al. (2012) perform an exercise in a MA(2) model and obtain that a rejection sampler with 20% tolerance ($k_N = 0.2N$) plus the local-linear density estimation recovers the results of the rejection sampler at 0.1% tolerance ($k_N = 0.001N$).

5.4.3 Specific ABC approach

Here we describe the specific ABC method we use to obtain inference from the WP-MRF model.

The first step is to apply a nearest neighbor approach with the algorithm (1) using a relative tolerance of 20% ($k_N = 0.2N$). Is important to note that is not possible directly simulate from the model (5.4). We use a 500 iteration Gibbs sampler to obtain this simulations. Convergence was previously studied using Gibbs sampler with 4 chains of 100 iterations after burn-in, convergence is monitored with scale reduction factor (Gelman et al., 2013).

A similar approach appears in Grelaud et al. (2009); Everitt (2012) where a MCMC run is used to obtain a simulations from the likelihood needed to set an ABC-MCMC algorithm (Marjoram et al., 2003). In that case, Everitt (2012) show that after substituting the simulation step by a MCMC run, the ABC-MCMC algorithm maintain its convergence properties.

We run an initial set of 20000 simulations with the rejection ABC algorithm (1) and keep 20% of the simulations with closest summary statistics to the observed in the data, distance is measured with euclidean norm.

The second step is to apply the neural network approach, proposed by Blum and François (2010), to obtain the posterior distribution. The local-linear approximation described in 5.4.2 suffer from the course of dimensionality and it hard to use this method when many parameters or summary statistics are included. Blum and François (2010) generalize this method by allowing the conditional expectation to have any form and also adding a variance function in the model, as follows

$$\begin{aligned} \theta(s) &= m(s) + \sigma(s)\zeta \\ \log(\theta(s) - \hat{m}(s))^2 &= \log \sigma^2(s) + \xi \end{aligned} \tag{5.10}$$

then a sample from the posterior is obtained by $\tilde{\theta}_i = \theta_i - \hat{m}(s_0) - \hat{\sigma}(s_0)(\frac{\theta_i - \hat{m}(s_i)}{\sigma(s_i)})$ and the posterior density is estimated as in (5.9). The functions $m()$ and $\sigma()$ are estimated using Neural Networks, since it accommodates non-linear relationships and still works when the number of summary statistics is large. This approach is implemented by the `abc` function from the `abc` package in `R` (Csillery et al., 2012). In order to apply the two step procedure described above we need a summary statistic to compare simulated and observed data. We use a four dimensions summary statistic in the ABC algorithm. Table 5.2 shows the summary statistic for all four parameters, NS_i and EW_i represent the set of neighbors in North-South and East-West respectively. Letting y^* to be the simulated values we define $S_k(y^*)$ as the summary statistic

designed to capture the kth parameter. The intercept is captured with the overall mean, and the slope with the correlation between response and covariate (fisher transformed). The last two statistics use residuals from a linear fit of the simulated response over the covariate to compute Moran statistic in NS and EW directions.

Table 5.2: Summary statistics for approximate Bayesian computation.

Parameter	Statistic	Description
β_0	$S_1(y^*) = (1/n)\sum_{i=1}^{n} z_i$	Overall mean
β_1	$S_2(y^*) = \frac{1}{2}\frac{1+\rho_{zx}}{1-\rho_{zx}}$	Fisher transform of $\rho_{zx} = \text{cor}(log(y^*), x)$
η_{ns}	$S_3(y^*) = \frac{n}{n_1}\frac{\sum_{i=1}^{n}\sum_{j\in NS_i} e_i e_j}{\sum_{i=1}^{n} e_i^2}$	$I_{NS}(e)$, $e_i = y_i^* - \hat{y}_i^*$
η_{ns}	$S_4(y^*) = \frac{n}{n_1}\frac{\sum_{i=1}^{n}\sum_{j\in EW_i} e_i e_j}{\sum_{i=1}^{n} e_i^2}$	$I_{EW}(e)$, $e_i = y_i^* - \hat{y}_i^*$

Intercept β_0 represents the expected number of crashes in an intersection with average traffic volume and with all its neighbors in the mean value. It should be a positive small value, we use $\beta_0 \sim t_3^+(1,1)$ as prior distribution. The prior distribution for the slope is $\beta_1 \sim N(0, 2^2)$. Finally, both dependency parameters has uniform prior in $[-1, 1]$, since they acts as spatial correlations within the model.

5.5 Simulation Study

Simulation scenarios are set to match the real data set characteristics. We set the intercept $\beta_0 \in (1, 5)$, and keeping the slope and the Winsorization bound constant, $\beta_1 = 0.3$ and $R = 300$. The total traffic covariate is computed in two ways. First we use the actual traffic data from each of the three cities. Secondly we simulate the total traffic from a negative binomial distribution independently for each intersection, we match the mean and variance of each city observed traffic data.

We use several dependency patterns; combining medium dependence in NS with medium and no dependence in EW, large dependence in NS with small and no dependence in EW, and an independence scenario. There are 156 simulated scenarios provided by 13 (η_{NS}, η_{EW}) combinations, two values for β_0 and six traffic covariate. One data set is generated for each scenario.

Figure 5.7 shows the Moran statistic used to measure the NS and EW dependence against

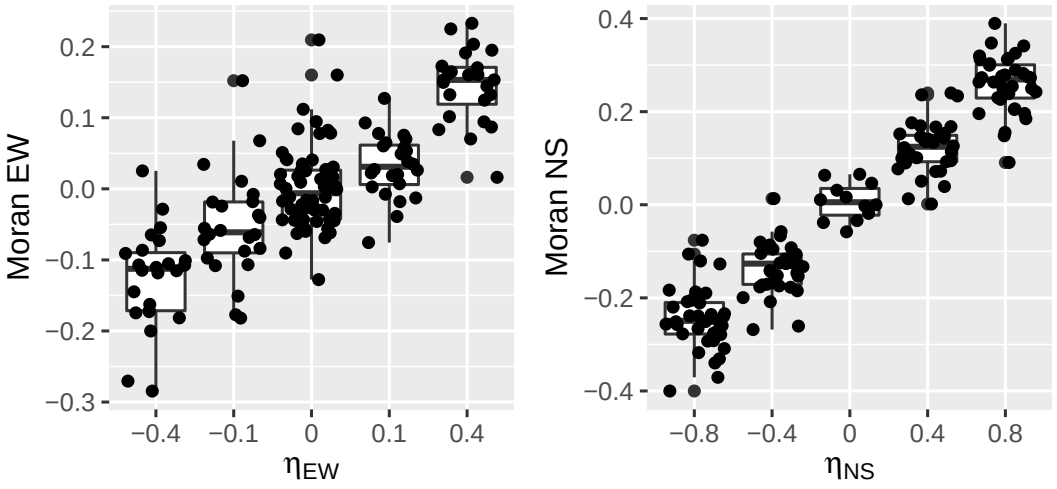

Figure 5.7: Boxplots of the observed Moran statistic for simulated data. Each panel represent a direction, EW on the right and NS on the left.

the true value of the correspondent dependency parameter used in the simulation. There is a positive relation among the true parameter and the Moran statistic in both directions. The range of the Moran statistic is about a half of the correspondent true parameters, this may be due the numerator only takes into account neighbors in one direction while the denominator compute variance considering all data points.

Figure 5.8 shows 95% credible intervals for the intercept β_1 in the 52 simulated scenarios that use Mrshalltown traffic information. In all cases the true value of the parameter is cover by the credible interval while the zero value is never included. It seems the credible intervals corresponding to positive dependence are somewhat wider.

Figure 5.9 shows 95% credible intervals for β_1 in the 52 simulated scenarios that use Mrshalltown traffic information. Credible intervals cover the true value of the slope in all but two cases, corresponding $(\beta_0, \eta_{NS}, \eta_{EW}) = (1, -0.8, 0.1)$ with simulated traffic covariate and $(\beta_0, \eta_{NS}, \eta_{EW}) = (1, 0, 0)$ with actual traffic covariate. The only parameter in common is that both cases happen to be scenarios with small intercept. The length of the intervals tend to be larger in the scenarios with larger intercept. Specially when $\eta_{NS} = 0.8$ many credible interval

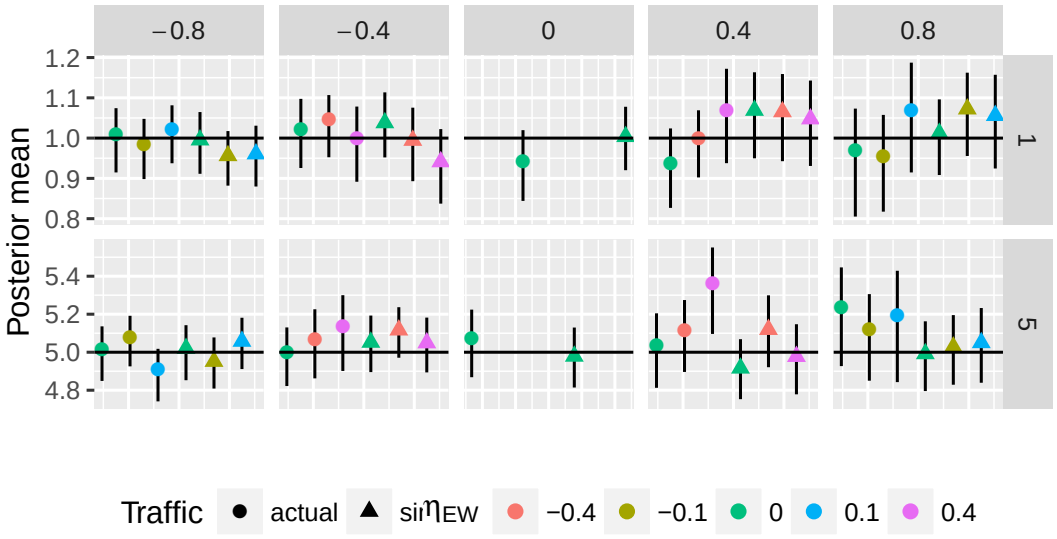

Figure 5.8: Credible intervals for β_0. Facets represent the true values for (β_0, η_{NS}), the color represents η_{EW} value, and the shape indicates if the total traffic is simulated from a negative binomial or is the actual traffic data.

include the zero value.

Figure 5.10 shows 95% credible intervals for η_{NS} in the 52 simulated scenarios that use Mrshalltown traffic information. Again, credible intervals cover its true value most of the times, the combination of $(\eta_{NS}, \eta_{EW}) = (0.4, -0.4)$ with simulated traffic covariate are the two scenarios which true values is not covered by the interval. Figure 5.11 presents credible intervals for η_{EW}, in all cases intervals cover the true value. For small dependence the zero is included in the credible intervals most of the time, and some cases of moderate dependency also cover the zero value.

Overall the figures 5.8,5.9,5.10, and 5.11 suggest the ABC approach is doing a good work in capturing the true values of the parameters. In the appendix Section B we presents similar figures but including the scenarios produced with all three cities traffic information. In all cases the parameters of the model are recover by the ABC approach, also the parameters are correctly distinguish from zero except for small values of dependence parameters.

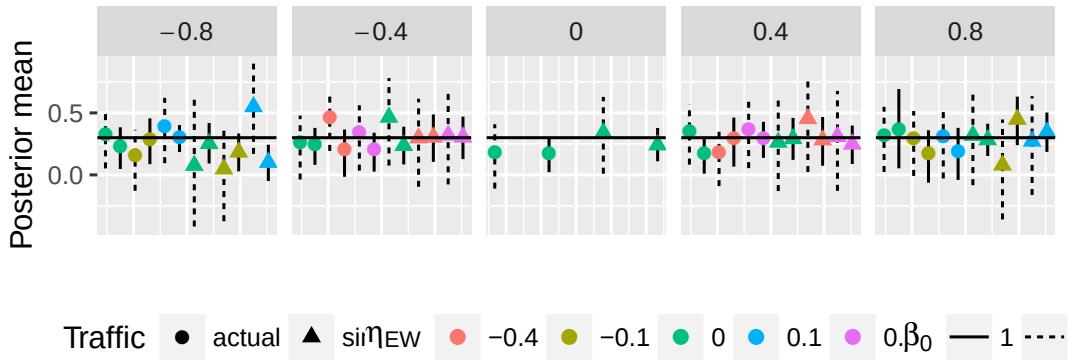

Figure 5.9: Credible intervals for β_1. Facets correspond to η_{NS} value, color corresponds to η_{EW} value, the shape indicates if the total traffic is simulated from a negative binomial or is the actual traffic data, and the line type represents β_0 value.

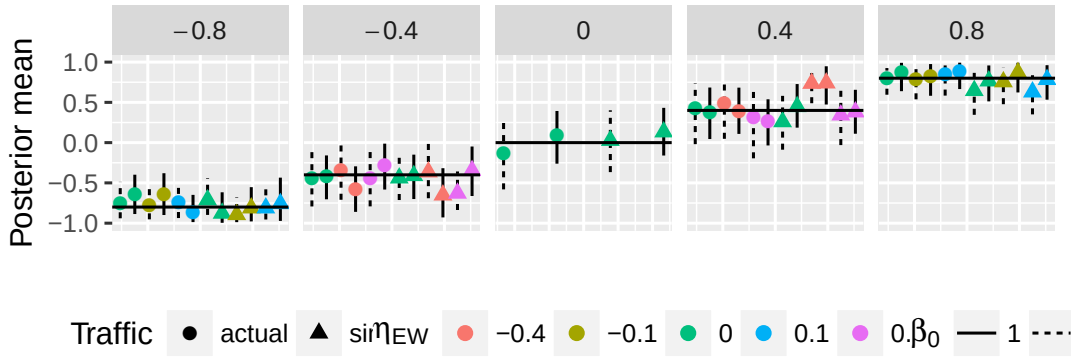

Figure 5.10: Credible intervals for η_{NS}.Facets corresponds to η_{NS} value, color corresponds to η_{EW} value, the shape indicates if the total traffic is simulated from a negative binomial or is the actual traffic data, and the line type represents β_0 value.

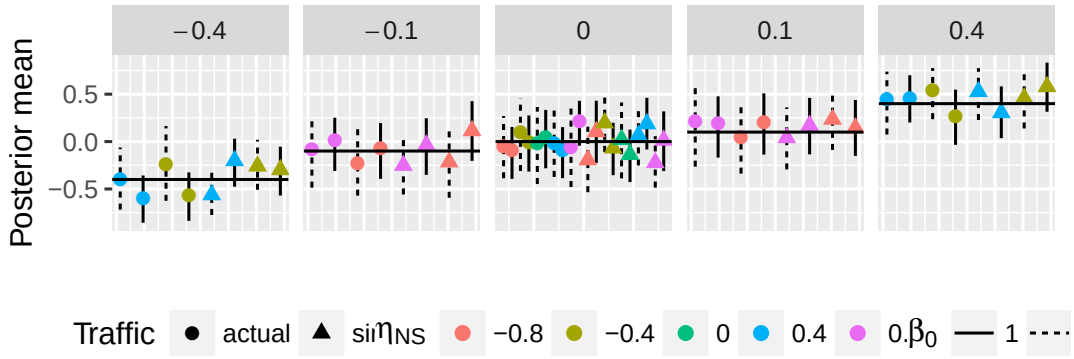

Figure 5.11: Credible intervals for η_{EW}. Facets correspond to η_{EW} value, color corresponds to η_{NS} value, the shape indicates if the total traffic is simulated from a negative binomial or is the actual traffic data, and the line type represents β_0 value.

5.6 Analysis of Iowa crash data

The Iowa DOT provided information of crashes per intersection in Tipton, Marshalltown and Ankeny for a ten years period (2004 - 2013). We fit the WPMRF model to each data set using ABC methodology described in 5.4.3. We presents inference results for all parameters in the model, and two measures that can be used to asses the danger of each individual intersection: posterior predictive mean, and the risk measure described in 5.3.4.

Model (5.5) has two regression parameters: an intercept, β_0 that represents the expected number of crashes in an intersection with the average traffic and all its neighbors at the mean level, and the traffic slope, β_1 that represents the effects of the traffic covariate over the expected number of crashes. In addition there are two dependence parameters, (η_{NS}, η_{EW}), that control the spatial correlation among intersections.

Figure 5.12 shows credible interval for every parameter in the model. The two regression coefficients are smaller for Tipton than the other two cities, which is consistent with the fact that Tipton is the smallest city out of the three with less crash accidents and smaller traffic volume. The effect of the covariate is positive in all 3 cities, as expected a large traffic volume increase the number of crashes at one particular intersection. In terms of neighbor dependence, only in Marshalltown there is evidence of positive dependence among intersections.

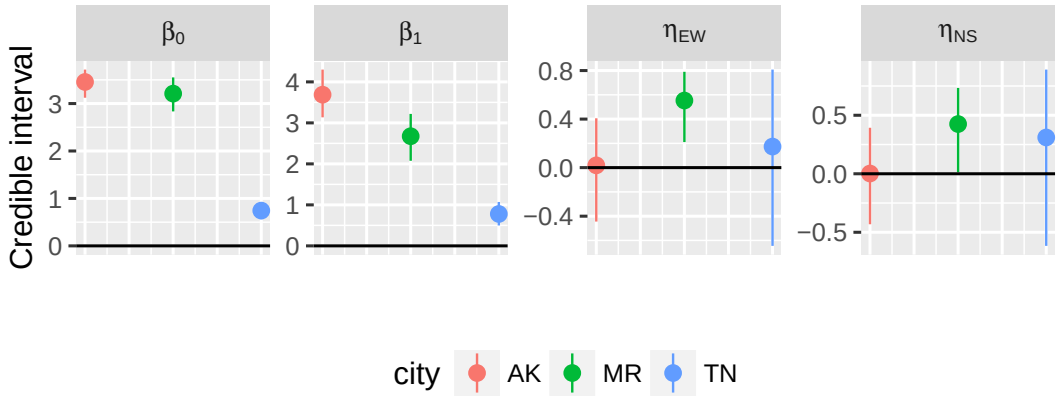

Figure 5.12: Parameter posterior median and posterior credible intervals. Each facet corresponds to one of the four parameters in the model, the color represent the city.

Figure 5.13 shows the posterior expected number of crashes at each intersection for all the three cities. This represents a smoothed version of the data presented in Figure 5.2. All the intersection with no observed crashes appears having a small but positive expectation, on the other extreme intersections with number of crashes larger than the Q90 statistic, present a posterior predictive mean much smaller than its observed value.

Figure 5.14 shows the intersections location for all three cities, red points indicates that the intersection's risk is high and the size of the dots is related with the total traffic in that intersection. Intersections with highest risk appears align on the main road. However, there are some intersections outside the main road showing risk between 50-75% range. This pattern is specially clear in Tipton, but also is present in Ankeny and Marshalltown. This may suggest the risk measurement is not dominated by the total traffic and has the potential to flag intersections with high risk of accidents even if those are not too much transited.

5.7 Discussion

In this paper, we have presented a strategy to model areal spatial data with covariate information and spatial correlation at the observation level. We propose to use an ABC approach to deal with intractability challenge the proposed model presents. A simulation study was conducted to determine the calibration of the ABC algorithm and, in the previous section, an

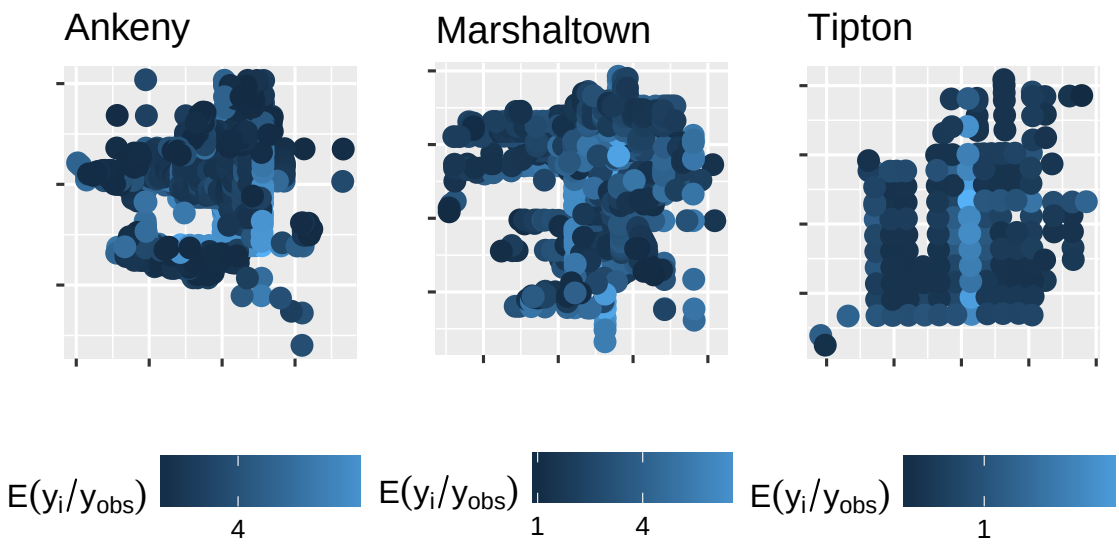

Figure 5.13: Posterior predictive expectation

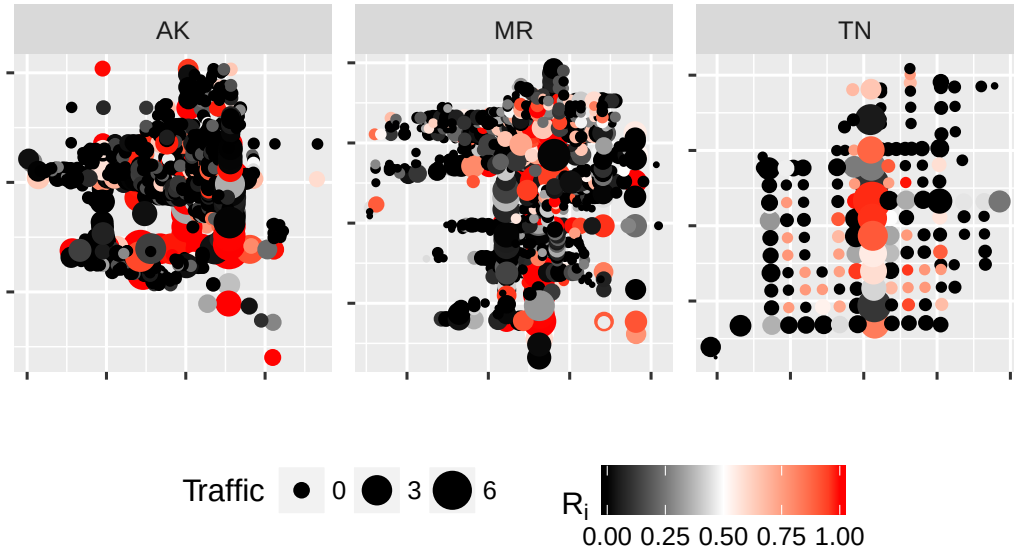

Figure 5.14: Intersections of Ankeny, Marshalltown and Tipton. Intersection's risk is represented by color of the points while intersection total traffic is represented by size of the point.

analysis of the number of crashes at intersections in three cities of Iowa was presented.

In the literature, many methods of modeling crash frequency data can be found. While most methods include different covariate information, relatively few models make use of spatial dependencies (with the exception of papers such as Song et al. (2006)). We argue that information about the location of intersections can be valuable in modeling some of the otherwise unexplained variation in crash numbers, as it seems likely that the crash number for one site is not completely independent from those of its neighbors, even after conditioning on other covariates.

We use a Markov random field model that puts the dependence structure directly on the crash numbers, which is a crucial difference from Song et al. (2006), who introduce spatial correlation via latent random effects. Another option would be data augmentation, a method that has been used by Wolpert and Ickstadt (1998). However, Markov random field models are a very promising option, especially since they allow for directly modeling of intersection level counts and then develop some model based measure of intersection's risk. This feature can be used to construct an index that help to decide which intersections should be closely monitored or reformed.

We make use of approximate Bayesian computation as a solution for the problem WP-MRF pose through the intractable nature of the joint posterior distribution. ABC makes possible to consider WP-MRF models as an alternative to others type of models (like latent hierarchical random effects models) and, at the same time, the crash frequencies modeling expand the horizon where ABC techniques has been applied in recent years.

Further research is warranted with respect to defining the neighborhoods used in the MRF model – while an anisotropic model as the one we propose allows for potentially different effects of North/South neighbors than of East/West neighbors, one can think about introducing distance-based weights or defining more sophisticated neighborhood structures that includes edge effects. Other model choices that might be tested might be the mean parametrization choice and zero-inflated models. All these options might be thought as part of a Bayesian model choice which within ABC methods is an active field of research at this moment.

CHAPTER 6. Summary and discussion

This dissertation presents severals methods for conducting Bayesian analysis of count data in high-dimensional cotexts. This Section presents a summary of the main finding from each Chapter and some suggestions for future work.

6.1 Summary

Chapter 2 presents a modeling strategy for ASE information, in the case of having data from one single hybrid genotype. The proposed hierarchical Poisson-lognormal mixture model address the main characteristics of ASE data. A measure of allele effect corrected for reference allele bias and a method to obtain credible intervals for this measure are proposed. The sparsity in the allele effects is handle with a Cauchy distribution, which showed better results than other shrinkage distribution in simulation studies.

The model proposed in Chapter 2 for a single hybrid is extended in Chapter 3 to include inbred genotypes and total RNA-seq. Gene-specific association measures between them heterotic patterns and allelic imbalance are proposed. The connection between gene heterosis patterns and ASE is explored using parallel coordinate plots, a gene-specific correlation coefficient, and conditional probability of heterosis given ASE observed patern.

Hierarchical models for both mean and variances simultaneously in sparse high dimensional context are studied in Chapter 4. The main focus of this Chapter are the effects of variance hierarchical modeling on the mean vector inference. We show the hierarchical Bayesian inference of the mean vector can be extremely biased in some cases, where the shrinkage level is learned for two sets of group-specific parameters. Measures to diagnostic this issue and modeling options that overcome this problem are presented.

In Chapter 5, models for areal spatial data with covariate information and spatial correlation at the observation level are presented. In particular, Chapter 5 proposes a Winsorized Markov random field (WP-MRF) model that puts the dependence structure directly on the crash numbers, and an approximate Bayesian computation (ABC) approach to deal with intractability challenge the proposed model presents. ABC makes possible to consider WP-MRF models as an alternative to others type of models and, at the same time, the crash frequencies modeling expand the horizon where ABC techniques has been applied in recent years. A measure of intersection's risk is proposed, which can be used as an index that help to decide which intersections should be closely monitored or reformed.

6.2 Future work

The horseshoe distribution was included among the hierarchical distributions considered in Chapter 2, since it has been suggested as a good default shrinkage distribution to use in high-dimensional contexts (Hahn and Carvalho, 2015). However, horseshoe results showed lack-of-convergence problems so it was excluded from the proposed models. Recently, Hahn and He (2016) point out the poor mixing of a horseshoe implementation based on a scale mixture of normals (which is the one used in this work) and propose to use an elliptical slice sampler instead. We would like to continue working analyzing the effect of the elliptical sample for the horseshoe distribution in the proposed models. Moreover, it could be interesting to study the horseshoe properties in different scenarios from what are usually considered, e.g. non-sparse situations, or non-centered set of parameters.

In the context of gene expression models with ASE counts, it is possible to continue working to compare the methods to analyze ASE proposed in this work with other methods to obtain inference for gene-specific allele effects. Additionally, it is desirable to include the false positive counts in the computation of the relevant contrasts. A gene-specific correction measure can potentially be developed in this way. To this end, deeper study of the connection among model parametrization and normalization factors is needed. Additionally, new measures relating heterosis and ASE information can be proposed, based on new biological basis for heterosis other than the dominance and overdominance hypothesis.

Relative to the confounding effects in hierarchical models, more work is needed to obtain an analytical form of the model performance boundary at which the mean signal strength starts being treated as variance, also simulations in more complex scenarios such as more groups or with similar structure to the used RNA-seq data example, might help to characterize these confounding effect.

With respect to methods presented in Chapter 5. Bayesian model choice in ABC context is an active field of research at this moment. Incorporating different aspects of the WP-MRF model (neighborhoods structure, edge effects, mean parametrization, etc) as part of a model choice strategy could be a line of research to continue exploring.

BIBLIOGRAPHY

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106.

Bar, H. Y., Booth, J. G., and Wells, M. T. (2014). A bivariate model for simultaneous testing in bioinformatics data. *Journal of the American Statistical Association*, 109(June):537 – 547.

Beaumont, M. A., Zhang, W., Balding, D. J., Beaumont, M. A., Rienzo, A. D., Donnelly, P., Toomajian, C., Sisk, B., Hill, A., Estoup, A., Wilson, I. J., Sullivan, C., Cornuet, J.-M., Moritz, C., Fan, J., Gijbels, I., Fan, J., Zhang, W., Fu, Y.-X., Li, W.-H., Hudson, R. R., Ihaka, R., Gentleman, R., King, J. P., Kimmel, M., Chakraborty, R., Loader, C. R., Nordborg, M., Ohta, T., Kimura, M., Perez-Lezaun, A., Calafell, F., Seielstad, M. T., Mateu, E., Comas, D., Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., Feldman, M. W., Reich, D. E., Feldman, M. W., Goldstein, D. B., Seielstad, M. T., Minch, E., Cavalli-Sforza, L. L. C., Shoemaker, J. S., Painter, I. S., Weir, B. S., Slatkin, M., Tavaré, S., Balding, D. J., Griffiths, R. C., Donnelly, P., Tishkoff, S. A., Varkonyi, R., Cahinhinan, N., Abbes, S., Argyropoulos, G., Wall, J. D., Weiss, G., von Haeseler, A., Wilson, I. J., Balding, D. J., Wilson, I. J., Weale, M. E., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

Bell, G. D. M., Kane, N. C., Rieseberg, L. H., and Adams, K. L. (2013). RNA-Seq analysis of allele-specific expression, hybrid effects, and regulatory divergence in hybrids compared with their parents from natural populations. *Genome Biology and Evolution*, 5(7):1309–1323.

Berger, J. O. and Strawderman, W. E. (1996). Choice of hierarchical priors: Admissibility in estimation of normal means. *Annals of Statistics*, 24(3):931–951.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236.

Biau, G., Cérou, F., Guyader, A., and Others (2015). New insights into approximate bayesian computation. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 51, pages 376–403. Institut Henri Poincaré, Institut Henri Poincaré.

Blum, M. G. B. and François, O. (2010). Non-linear regression models for approximate Bayesian computation. *Statistics and Computing*, 20(1):63–73.

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2009). Evaluation of statistical methods for normalization and differential expression in mRNA-weq experiments. *BMC Bioinformatics*, 11(1):94.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan : A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling Sparsity via the Horseshoe. *Journal of Machine Learning Research WCP*, 5(73-80):73–80.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Chen, Y., Lun, A. T. L., and Smyth, G. K. (2014). Differential expression analysis of complex RNA-seq experiments using edgeR. In *Statistical Analysis of Next Generation Sequencing Data*, chapter 3, pages 25–50.

Cook, D., Hofmann, H., Lee, E.-K., Yang, H., Nikolau, B., and Wurtele, E. (2007). Exploring gene expression data, using plots. *Journal of Data Science*, 5(2):151–182.

Csillery, K., Francois, O., and Blum, M. G. B. (2012). abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*.

Cui, X., Hwang, J. T. G., Qiu, J., Blades, N. J., and Churchill, G. A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, 6(1):59–75.

Darwin, C. (1876). *The effects of cross and self fertilisation in the vegetable kingdom*. J. Murray.

Datta, S. and Nettleton, D. (2014). *Statistical Analysis of Next Generation Sequencing Data*. Springer.

Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., and Pritchard, J. K. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, 25(24):3207–3212.

Efron, B. (1992). Introduction to James and Stein (1961) estimation with quadratic loss. In Kotz, S. and Jhonson, N., editors, *Breakthroughs in Statistics. Volume 1*, pages 437–442. Springer-Verlag.

Everitt, R. G. (2012). Bayesian parameter estimation for latent Markov random fields and social networks. *Journal of Computational and graphical Statistics*, 21(4):940–960.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC press.

Gene Hwang, J. T., Qiu, J., and Zhao, Z. (2009). Empirical Bayes confidence intervals shrinking both means and variances. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):265–285.

Ghosh, J. K., Delampady, M., and Samanta, T. (2006). *An Introduction to Bayesian Analysis*. Springer.

Grelaud, A., Robert, C. P., Marin, J.-M., Rodolphe, F., Taly, J.-F., and Others (2009). ABC likelihood-free methods for model choice in Gibbs random fields. *Bayesian Analysis*, 4(2):317–335.

Griffith, D. A. and Paelinck, J. H. P. (2011). Frequency Distributions for Simulated Spatially Autocorrelated Random Variables. In *Non-standard Spatial Statistics and Spatial Econometrics*, pages 37–73. Springer.

Hahn, P. R. and Carvalho, C. M. (2015). Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association*, 110(509):435–448.

Hahn, P. R. and He, J. (2016). Elliptical slice sampling for Bayesian shrinkage regression with applications to causal inference.

Hallauer, A. R., Carena, M. J., and Filho, J. B. M. (2010). Heterosis. In *Quantitative Genetics in Maize Breeding*, chapter Heterosis, pages 477–529. Springer New York, New York, NY.

Hodgkinson, A., Grenier, J.-C., Gbeha, E., and Awadalla, P. (2016). A haplotype-based normalization technique for the analysis and detection of allele specific expression. *BMC Bioinformatics*, 17(364).

Hu, Y.-J., Sun, W., Tzeng, J.-Y., and Perou, C. M. (2015). Proper use of allele-specific expression improves statistical power for cis -eQTL mapping with RNA-seq data. *Journal of the American Statistical Association*, 1459(December):0–0.

Hwang, J. T. G. and Liu, P. (2010). Statistical applications in genetics and molecular biology optimal tests shrinking both means and variances applicable to microarray data analysis. *Article Statistical Applications in Genetics and Molecular Biology*, 9(1).

Hyndman, R. J., Bashtannyk, D. M., and Grunwald, G. K. (1996). Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics*, 5(4):315–336.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 361–379.

Ji, T., Liu, P., and Nettleton, D. (2014). Estimation and testing of gene expression heterosis. *Journal of agricultural, biological, and environmental statistics*, 19(3):319–337.

Jing, B.-Y., Li, Z., Pan, G., and Zhou, W. (2016). On SURE-type double shrinkage estimation. *Journal of the American Statistical Association*, 111(516):1696–1704.

Kaiser, M. S. (2002). Markov random field models. In *Encyclopedia of Environmetrics*, pages 1213–1225. Wiley John & Sons.

Kaiser, M. S., Caragea, P. C., and Furukawa, K. (2012). Centered parameterizations and dependence limitations in Markov random field models. *Journal of Statistical Planning and Inference*, 142(7):1855–1863.

Kaiser, M. S. and Cressie, N. (1997). Modeling Poisson variables with positive spatial dependence. *Statistics & Probability Letters*, 35(4):423–432.

Kaiser, M. S. and Cressie, N. (2000). The construction of multivariate distributions from Markov random fields. *Journal of Multivariate Analysis*, 73(2):199–220.

Kruschke, J. K. (2014). *Doing Bayesian Data Analysis : a Tutorial with R, JAGS, and Stan*. Academic Press, 2nd edition.

Landau, W. *fbseq: fbseq*. R package version 0.0.

Landau, W. and Niemi, J. (2016). A fully Bayesian strategy for high-dimensional hierarchical modeling using massively parallel computing.

Landau, W., Niemi, J., and Nettleton, D. (2016). Fully Bayesian analysis of RNA-seq counts for the detection of gene expression heterosis.

León-Novelo, L. G., McIntyre, L. M., Fear, J. M., and Graze, R. M. (2014). A flexible Bayesian method for detecting allelic imbalance in RNA-seq data. *BMC Genomics*, 15(1):920.

Lindley, D. (1962). Discussion on professor Stein's paper: 'Confidence Sets for the Mean of a Multivariate Normal Distribution'. *Journal of the Royal Statistical Society. Series B (Methodological)*, 24(2):285 – 288.

Lithio, A. and Nettleton, D. (2015). Hierarchical modeling and differential expression analysis for RNA-seq experiments with inbred and hybrid genotypes. *Journal of Agricultural, Biological, and Environmental Statistics*, 20(4):598–613.

Lorenz, D. J., Gill, R. S., Mitra, R., and Datta, S. (2014). Using RNA-seq Data to Detect Differentially Expressed Genes. In Datta, S. and Nettleton, D., editors, *Statistical Analysis of Next Generation Sequencing Data*, chapter 2, pages 25–49.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328.

McDonald, T. (2012). Traffic Safety Analysis for Local Agencies.

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006). An efficient Markov chain monte carlo method for distributions with intractable normalising constants. *Biometrika*, 93:451–458.

Muller, P., Parmigiani, G., and Rice, K. (2006). FDR and Bayesian multiple comparisons ules.

Müller, P., Parmigiani, G., Robert, C., Rousseau, J., and Ler, P. M. (2004). Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association*, 99(46):990–1001.

Murray, I., Ghahramani, Z., and MacKay, D. (2006). MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Inteligence*, pages 359–366.

Niemi, J., Mittman, E., Landau, W., and Nettleton, D. (2015). Empirical Bayes analysis of RNA-seq data for detection of gene expression heterosis. *Journal of Agricultural, Biological, and Environmental Statistics*, 20(4):614–628.

Panousis, N. I., Gutierrez-Arcelus, M., Dermitzakis, E. T., and Lappalainen, T. (2014). Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies. *Genome Biol*, 15(9):467.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Paschold, A., Jia, Y., Marcon, C., Lund, S., Larson, N. B., Yeh, C.-T., Ossowski, S., Lanz, C., Nettleton, D., Schnable, P. S., and Hochholdinger, F. (2012). Complementation contributes to transcriptome complexity in maize (Zea mays L.) hybrids relative to their inbred parents. *Genome Research*, 22(12):2445–54.

Pirinen, M., Lappalainen, T., Zaitlen, N. a., Consortium, G., Dermitzakis, E. T., Donnelly, P., McCarthy, M. I., and Rivas, M. a. (2014). Assessing allele specific expression across multiple tissues from RNA-seq read data. *bioRxiv*, 31(March):007211.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1):139–40.

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):25.

Ryder, P., McKeown, P. C., Fort, A., and Spillane, C. (2014). Epigenetics and Heterosis in Crop Plants. In *Epigenetics in Plants of Agronomic Importance: Fundamentals and Applications*, pages 13–31. Springer International Publishing, Cham.

Sachs, M. C. (2016). *plotROC: Generate Useful ROC Curve Charts for Print and Interactive Use*. R package version 2.0.3.

Schnable, P. S. and Springer, N. M. (2013). Progress toward understanding heterosis in crop plants. *Annu. Rev. Plant Biol*, 64:71–88.

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., Minx, P., Reily, A. D., Courtney, L., Kruchowski, S. S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S. M., Belter, E., Du, F., Kim, K., Abbott, R. M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S. M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M. J., McMahan, L., Van Buren, P., Vaughn, M. W., Ying, K., Yeh, C.-T., Emrich, S. J., Jia, Y., Kalyanaraman, A., Hsia, A.-P., Barbazuk, W. B., Baucom, R. S., Brutnell, T. P., Carpita, N. C., Chaparro, C., Chia, J.-M., Deragon, J.-M., Estill, J. C., Fu, Y., Jeddeloh, J. A., Han, Y., Lee, H., Li, P., Lisch, D. R., Liu, S., Liu, Z., Nagel, D. H., McCann, M. C., SanMiguel, P., Myers, A. M., Nettleton, D., Nguyen, J., Penning, B. W., Ponnala, L., Schneider, K. L., Schwartz, D. C., Sharma, A., Soderlund, C., Springer, N. M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T. K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J. L., Dawe, R. K., Jiang, J., Jiang, N., Presting, G. G., Wessler, S. R., Aluru, S., Martienssen, R. A., Clifton, S. W., McCombie, W. R., Wing, R. A., Wilson, R. K., Doebley, J. F., Gaut, B. S., Smith, B. D., Troyer, A. F., Duvick, D. N., Paterson, A. H., Bowers, J. E., Chapman, B. A., Blanc, G., Wolfe, K. H., Swigonova, Z.,

Paterson, A. H., SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y., Bennetzen, J. L., Wei, F., SanMiguel, P., McClintock, B., Feschotte, C., Jiang, N., Wessler, S. R., Kumar, A., Bennetzen, J. L., Kapitonov, V. V., Jurka, J., Yang, L., Bennetzen, J. L., Liang, C., Mao, L., Ware, D., Stein, L., Haberer, G., Alexandrov, N. N., Zhong, C. X., Sharma, A., Presting, G. G., Sharma, A., Schneider, K. L., Presting, G. G., Lisch, D., Alleman, M., Fu, Y., Palmer, L. E., Zhang, W., Lee, H. R., Koo, D. H., Jiang, J., Tian, C. G., Seoighe, C., Gehring, C., Shaked, H., Kashkush, K., Ozkan, H., Feldman, M., Levy, A. A., Song, K., Lu, P., Tang, K., Osborn, T. C., Tate, J. A., Joshi, P., Soltis, K. A., Soltis, P. S., Soltis, D. E., Thomas, B. C., Pedersen, B., Freeling, M., Emrich, S. J., and McClintock, B. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science (New York, N.Y.)*, 326(5956):1112–5.

Shull, G. H. (1908). The composition of a field of maize. *Journal of Heredity*, os-4(1):296–301.

Soneson, C. and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, 14:91.

Song, J. J., Ghosh, M., Miaou, S., and Mallick, B. (2006). Bayesian multivariate spatial models for roadway traffic crash mapping. *Journal of Multivariate Analysis*, 97:246–273.

Srivastava, S. and Chen, L. (2010). A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic acids research*, 38(17):e170.

Stevenson, K. R., Coolon, J. D., and Wittkopp, P. J. (2013). Sources of bias in measures of allele-specific expression derived from RNA-sequence data aligned to a single reference genome. *BMC genomics*, 14(1):536.

Sun, W. and Hu, Y. (2013). eQTL Mapping Using RNA-seq Data. *Statistics in Biosciences*, 5Sun, W.,(1):198–219.

Sun, W. and Hu, Y. (2014). Mapping of expression quantitative trait loci using RNA-seq data. In Nettleton, D. and Datta, S., editors, *Statistical Analysis of Next Generation Sequencing Data*, pages 25–50.

Swanson-Wagner, R. A., Jia, Y., DeCook, R., Borsuk, L. A., Nettleton, D., and Schnable, P. S. (2006). All possible modes of gene action are observed in a global comparison of gene expression in a maize F1 hybrid and its inbred parents. *Proceedings of the National Academy of Sciences of the United States of America*, 103(18):6805–10.

Van De Wiel, M. A., Leday, G. G. R., Pardo, L., Rue, H., Van Der Vaart, A. W., and Van Wieringen, W. N. (2013). Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics*, 14(1):113–128.

Ventrucci, M., Scott, E. M., and Cocchi, D. (2011). Multiple testing on standardized mortality ratios: a Bayesian hierarchical model for FDR estimation. *Biostatistics*, 12(1):51–67.

Vijaya Satya, R., Zavaljevski, N., and Reifman, J. (2012). A new strategy to reduce allelic bias in RNA-Seq readmapping. *Nucleic Acids Research*, 40(16):1–9.

Wei, X. and Wang, X. (2013). A computational workflow to identify allele-specific expression and epigenetic modification in maize. *Genomics, proteomics & bioinformatics*, 11(4):247–52.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H. (2017). *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*. R package version 0.6.3.

Wickham, H. and Francois, R. (2016). dplyr: A grammar of data manipulation. R package version 0.5.0.

Wolpert, R. L. and Ickstadt, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika*, 85:251–267.

Zegeer, C. V. (1986). *Methods for identifying hazardous highway elements*. Number 128.

Zhao, Z. (2010). Double shrinkage empirical Bayesian estimation for unknown and unequal variances. *Statistics and Its Interface*, 3:533–541.

APPENDIX A. Initial simulation study for single hybrid model

A.1 Simulation study

We perform a simulation study to explore how the model capture several characteristics of interest in the data. We use data model (2.1) as the base for simulations, and specify different scenarios. First we describe the datasets simulation scenarios, then we describe the analyses we perform on each simulated data. Finally, we present the simulation study results.

We investigate four topics: overdispersion effects, sparsity in allele effect, distribution of the true signals (allele effects truly different from zero), and bias toward reference allele. We designed three scenarios for each topic but the sparsity (2 scenarios) and combine them in a full factorial way, having 54 scenarios in total. Each scenario is replicated 2 times. In all scenarios we simulate the ASE count for 8 observations per gene, with $G = 5000$ total genes, we do not include block effects, $\beta_{g3} = \beta_{g4} = \beta_{g5} = 0$, and simulate intercept effects as $\beta_{1g} \sim N(3, 1)$.

There are eight analyses performed on each simulated dataset. First every data is analyzed using model 2.1 with each of the three shrinkage distributions, from equation (2.2), for β_{g2} and also a normal distribution. The main reason for this is to asses the impact of the hierarchical distribution of the regression coefficients on the posterior inference. Secondly we fit a non-hierarchical version of each model, this is, fixing $\theta = 0$, $\sigma^2 = 3$, $\tau = .1$, $\nu = 8$.

We run 4 MCMC chains with 50×10^3 iterations, using `fbseq` package, convergence is assessed using potential scale reduction factor statistic.

A.1.1 Data scenarios

Overdispersion effects Overdispersion effects allows for mean-variance quadratic relationship, the analysis of maize data in section 2.5 suggests there is little or none overdispersion

present, then we would like to explore how the model adjust different levels of overdispersion. Overdispersion is controlled by (ν, τ) the two hyperparameters of the distribution of γ_g.

We set the $\tau = (.01, .1, 1)$ and $\nu = (4, 8, 12)$ in order to produce scenarios with little, medium and high overdispersion effects but controlling the right tail of the γ_g distribution.

Table A.1: Quantiles of Inverse Gamma

nu	tau	Q1	Q50	Q90	Q99
4.0	0.01	0.003	0.01	0.04	0.13
8.0	0.10	0.040	0.11	0.23	0.49
12.0	1.00	0.458	1.06	1.90	3.36

Table A.1 shows quantile values of an $IG(\nu/2, \nu\tau/2)$ distribution for selected pairs of (ν, τ). We can see as τ is increase by a factor of 10, 1% and 50% quantiles increase in the same amount, τ controls the overall scale of the overdispersion effects. However, the large quantiles does not increment that much due to the effect of an increasing ν. The chosen scenarios consist in: none overdispersion for almost all genes, a moderate overdispersion effect for most genes and some with high levels of over dispersion, and a third scenario with high overdispersion in many genes.

Reference allele bias

As we explained in section 2.2, genes reads truly corresponding to the non-reference allele (Mo17) are less likely to be assigned, since they are compare again the reference genome table from B73 maize. Importantly, this effect is the same across all genes, which make difficult to deal with if the statistical strategy fit individual models per gene.

We simulate data sets with different amount of reference allele bias by first simulating Y^*_{gM} according to model (2.1) and then simulate $Y_{gM}|Y^*_{gM} \sim Binomial(Y^*_{gM}, p)$. Parameter p controls the proportion of true reads that are actually mapped, we set three scenarios corresponding to large, medium and none bias as $p = (.5, .75, 1)$.

Sparsity and True signal distribution

Another important feature of genetic data is the sparsity of the true signal effects. Allele differences per gene, β_{g2}, are expected to be sparse effects with many genes showing no difference among the two alleles.

Carvalho et al. (2010) uses a t distribution with a point mass at zero mixture, controlling the sparsity by fixing the mixing proportion. This makes hard to control the true signals since simulations from the t distribution might be too close to zero. Instead, Van De Wiel et al. (2013) use symmetric mixture of two normals and a point mass.

We simulate values for β_{g2} using a mixture of a point mass at zero and a true signal distribution. We use three different distributions to simulate the allele effects different from zero: a mixture of normals (MixNr), a mixture of point masses (MixPm) and a normal distribution. We can write the densities we use to obtain simulated values for β_{g2} as follows:

$$
\begin{array}{lll}
\text{MixNr} & \beta_{g2} \overset{\text{ind}}{\sim} & w\delta_0 + (1-w)N(X, s^2) \quad \text{where } X \sim 2Ber(0.5) - 1 \\
\text{MixPm} & \beta_{g2} \overset{\text{ind}}{\sim} & w\delta_0 + (1-w)\delta_X \quad \text{where } X \sim 2Ber(0.5) - 1 \\
\text{Nr} & \beta_{g2} \overset{\text{ind}}{\sim} & w\delta_0 + (1-w)N(0, s^2)
\end{array}
$$

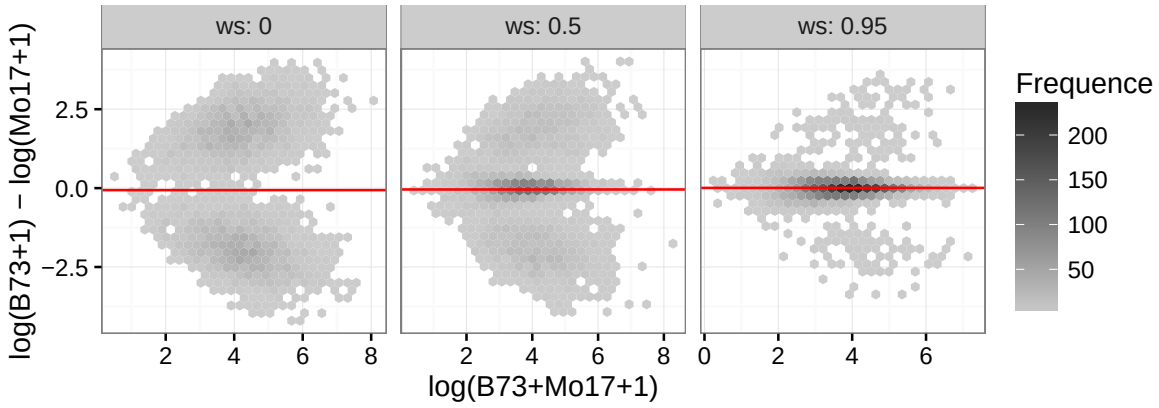

Figure A.1: Hexagon binning plot of means per gene and allele, for simulated data sets with no overdispersion nor bias toward reference allele ($\tau = .01$, $p = 1$)

Figure A.1 shows hexagon binning plots for one replicate data set for each value of the sparsity parameter. We set $w = (.95, .5)$ and $s = 1/3$, the scenarios cover highly and medium sparse signals and assure the coefficients different zero are not too small with no expected bias for any of the two alleles.

A.1.2 Simulation results

In order to describe model performance to detect genes with differential allele expression, we construct receiver operating characteristic (ROC) curves for each simulation scenario. We use $z_g = \frac{|E(\Delta_g|y)|}{sd(\Delta_g|y)}$ as a continuous score to compute the ROC curves.

Figure A.2 presents ROC curves of model 2.1 in most of the scenarios, within each facet there are 18 ROC curves. Overall, the model does a good job in capturing the genes with true allele effects.

Overdispersion level affect the model performance only when is really large. Small and moderate levels of overdispersion for most genes with a few overdispersed genes is nicely handle by the model. Reference allele bias does not seem to have impact in signal detection, even when half of the non-reference allele reads are lost we still be able to identify genes with allele effects correctly. The distribution of true signal indicates a normal distribution (NR) makes detection slightly harder while the mixture of point masses (MxPm) makes it slightly easier. Figure A.2 suggest all three hierarchical distribution from equation (2.2) have the same performance when 50% of the genes has zero allele effect.

Figure A.3 is analogous to the previous figure but reporting ROC curves when 95% of the genes has zero allele effect. As in the scenarios with medium level of sparsity we commented above, model (2.1) shows an overall good performance in detecting true signals for the shrinkage distribution. The major issue appears to be when the overdispersion level is very large and to a lesser extent when there are many weak true signals. When a normal distribution is used as hierarchical distribution of β_{g2} the performance is really bad, indicating normal distribution cannot handle high sparse signals. In addition, the effect of the overdispersion in the normal distribution case is counterintuitive. It seems that when the overdispersion level increase, signal detection rates for the normal distribution increase.

ROC curves for non-hierarchical model are shown in Figure A.4 for cases with 50% of non-zero effects and figure A.5 for datasets with only 5% of non-zero effects. Overall, the non-hierarchical models perform worse than the fully Bayesian model, having lower true positive detection rates in most scenarios.

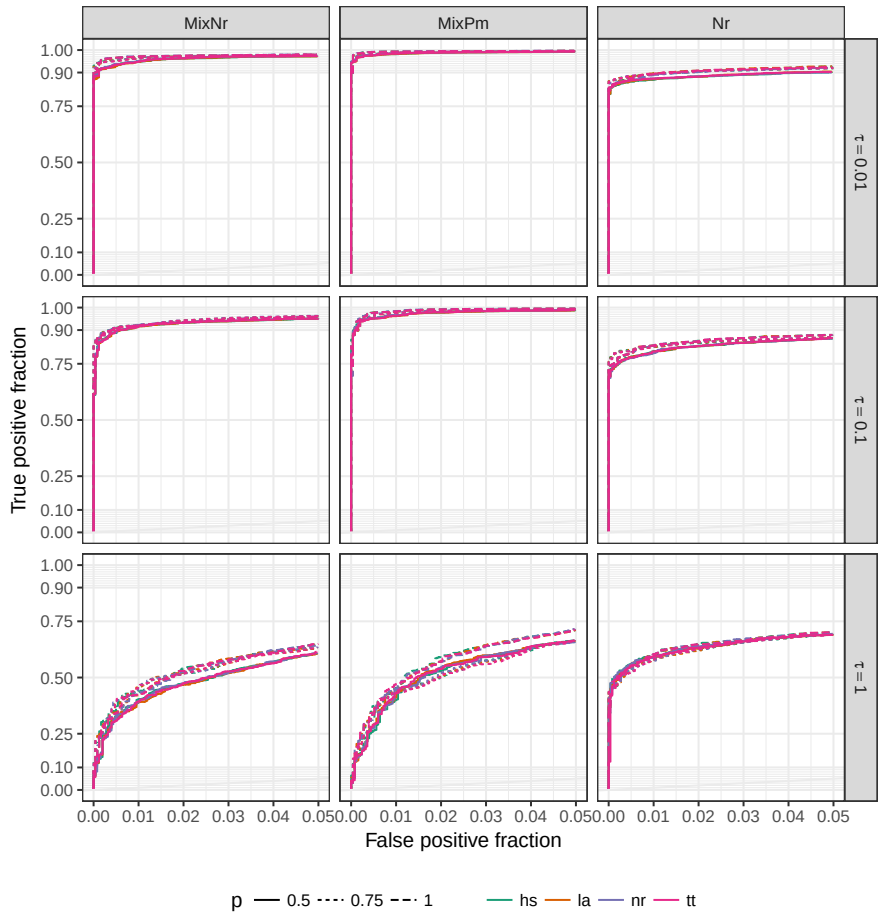

Figure A.2: ROC curves for false positive rates lower than 0.25 in scenarios with 50% of genes with no allele effect. Facets combine the true signal distribution and the overdispersion level, color represent the hierarchical distribution and line type the reference allele bias.

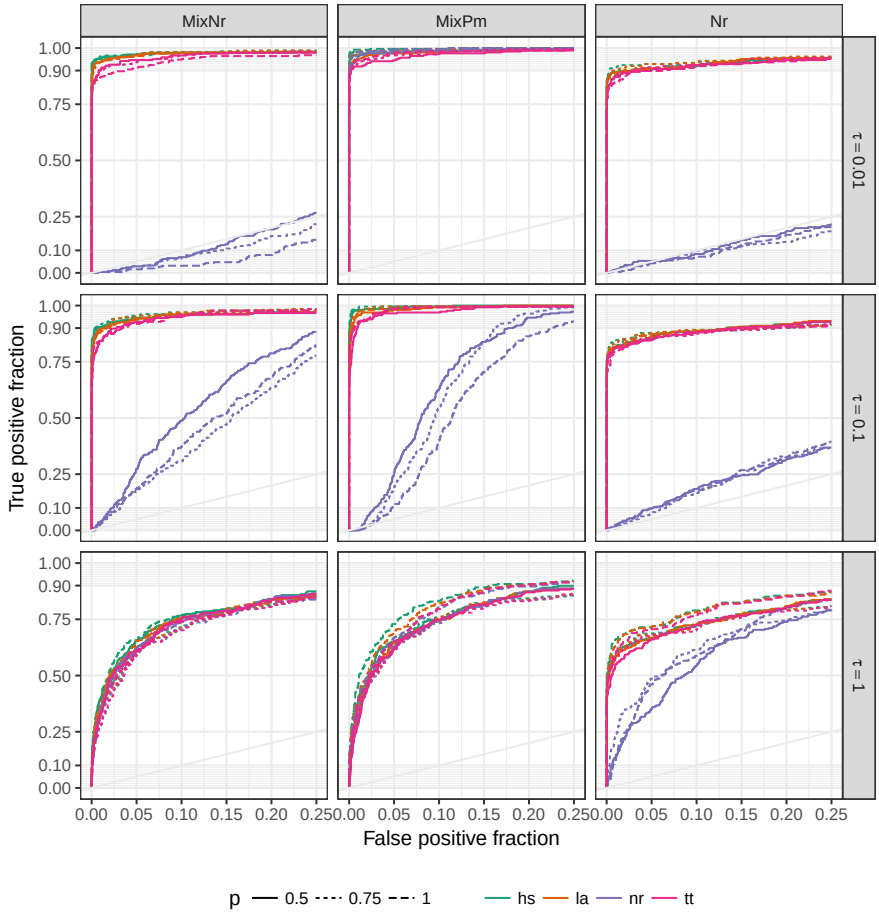

Figure A.3: ROC curves for false positive rates lower than 0.25 in scenarios with 95% of genes with no allele effect. Facets combine the true signal distribution and the overdispersion level, color represent the hierarchical distribution and line type the reference allele bias.

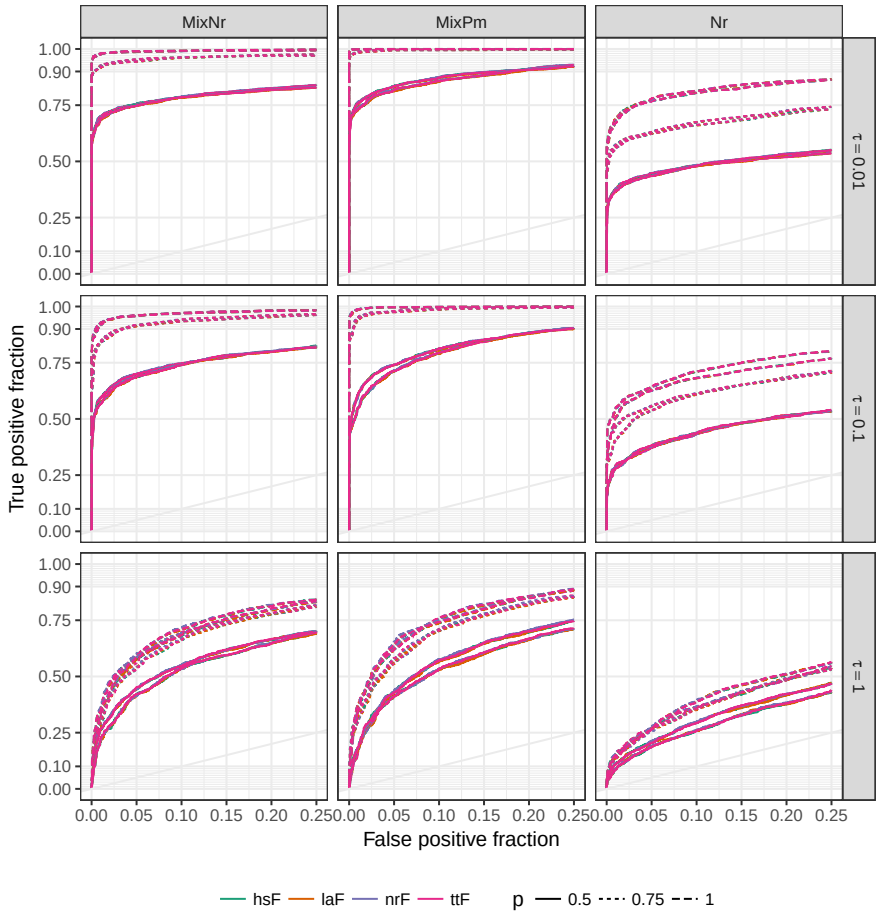

Figure A.4: Non-hierarchical model ROC curves for false positive rates lower than 0.25 in scenarios with 50% of genes with no allele effect. Facets combine the true signal distribution and the overdispersion level, color represent the prior distribution and line type the reference allele bias

Non-hierarchical model results are much more sensitive to the data patterns we include in simulated datasets. When the true signals are weak, i.e., simulated from a normal distribution, the true signal detection rates are very low for all cases. When true signals are stronger but there is a large overdispersion level or there is a large bias towards reference allele, we see bad detection rates again. The sparsity level and the prior distribution for regression coefficients does not seem to have a big effect in the results.

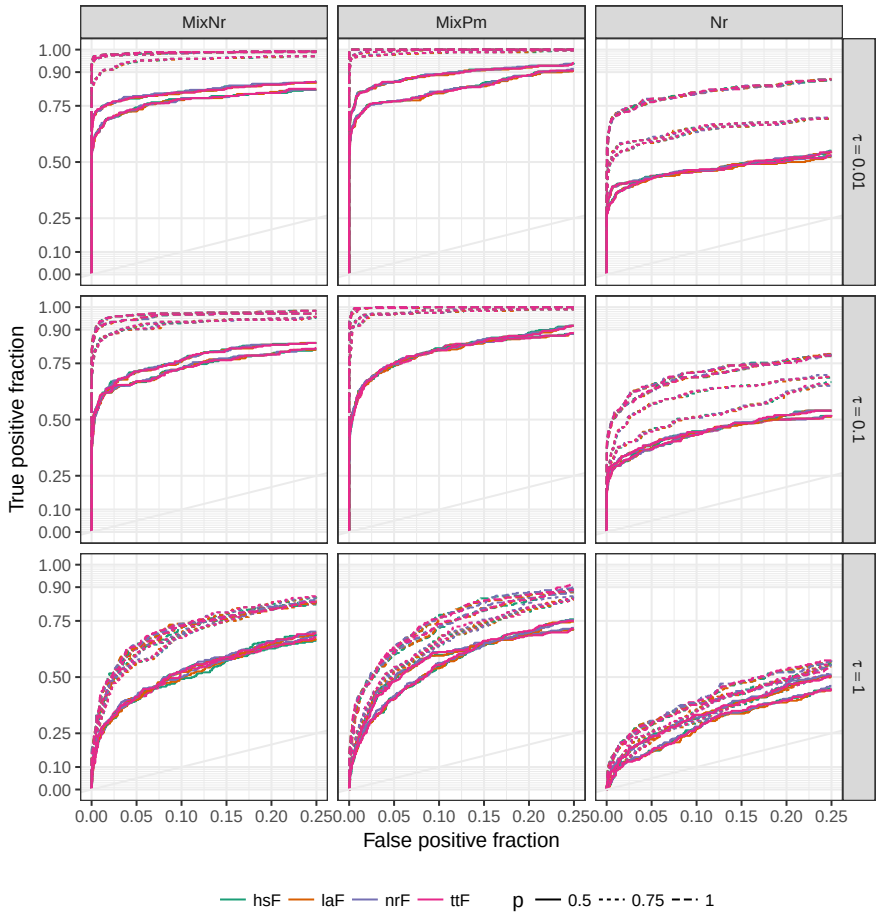

Figure A.5: Non-hierarchical model ROC curves for false positive rates lower than 0.25 in scenarios with 95% of genes with no allele effect. Facets combine the true signal distribution and the overdispersion level, color represent the prior distribution and line type the reference allele bias

APPENDIX B. Supplementary Figures from methods in Chapter 5

This appendix includes some supplementary Figures from the simulation study described in Section 5.5, and alternative Figures to present the model results over the intersection crashes in Tipton, Marshalltown and Ankeny cities.

B.1 Figures from simulation study

Credible intervals for each of the 4 parameters in the model are presented in Figures B.1, B.2, B.3 and B.4

Figure B.1: Credible intervals of β_0. Column facets correspond to traffic covariate information city, row facets represent the true value of β_0 to simulate data, color indicates if the traffic covariate is simulated or not.

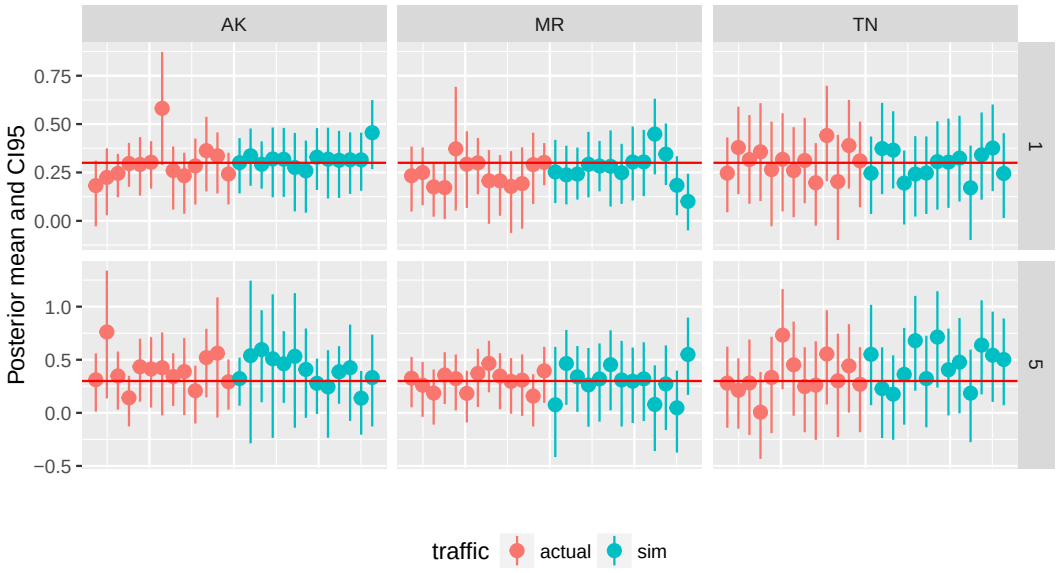

Figure B.2: Credible intervals of β_1. Column facets correspond to traffic covariate information city, row facets represent the true value of β_0 to simulate data, color indicates if the traffic covariate is simulated or not.

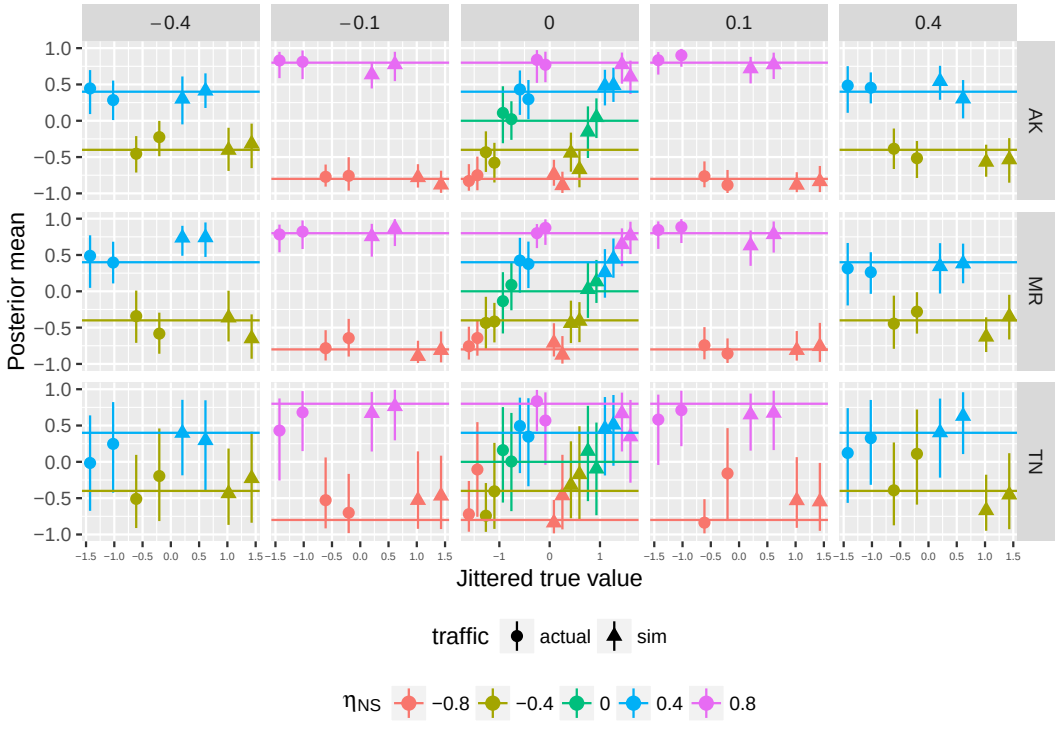

Figure B.3: Credible intervals for η_{NS}. Columns facets corresponds to the true value of η_{EW}, row facets correspond to traffic covariate town, color indicates the true value of η_{NS}.

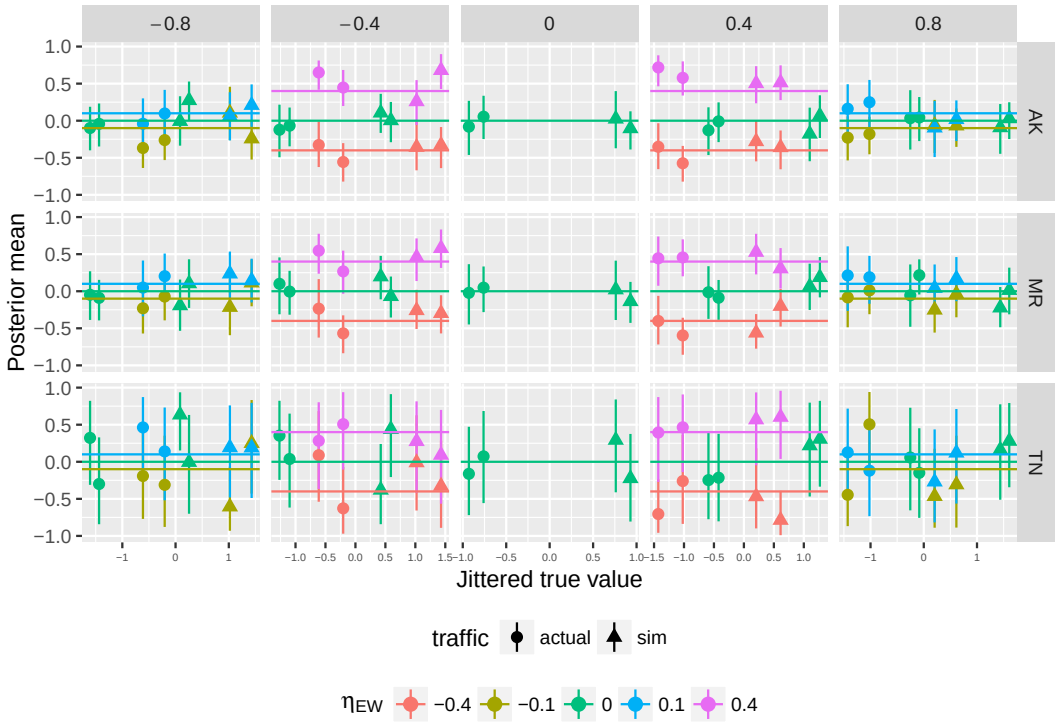

Figure B.4: Credible intervals for η_{EW} Columns facets corresponds to the true value of η_{NS}, row facets correspond to traffic covariate town, color indicates the true value of η_{EW}.

B.2 Figures from restuls of Iowa crash data model

Here we present alternative visualizations to the Figures included in Section 5.6. Figure B.5 shows the intersection's risk measure as a binary variable, we flag intersection with risk index larger than 0.8. Figures B.6, B.7 and B.8 show the posterior prediced expected value of number of crashes for each intersection side by side with the actual observed data, each Figure corresponds to one of the 3 cities analyzed.

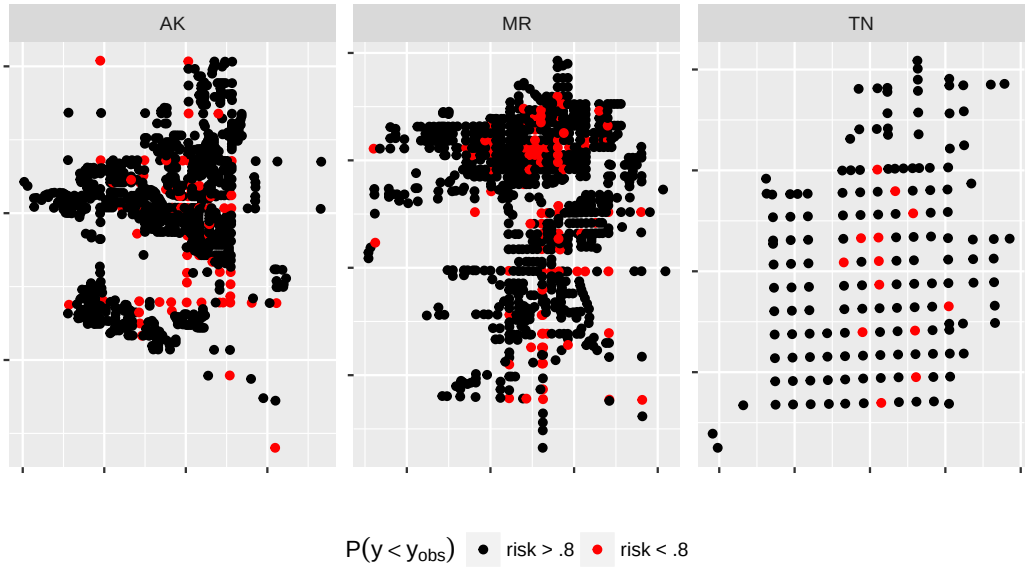

Figure B.5: Intersection's risk bynary index

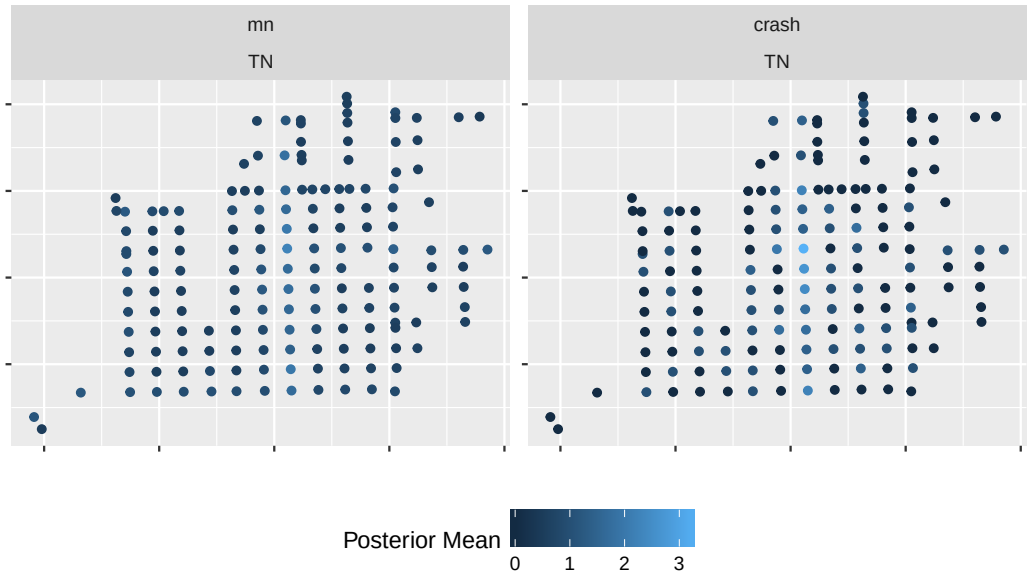

Figure B.6: Tipton: Posterior predictive expectation and actual crashes

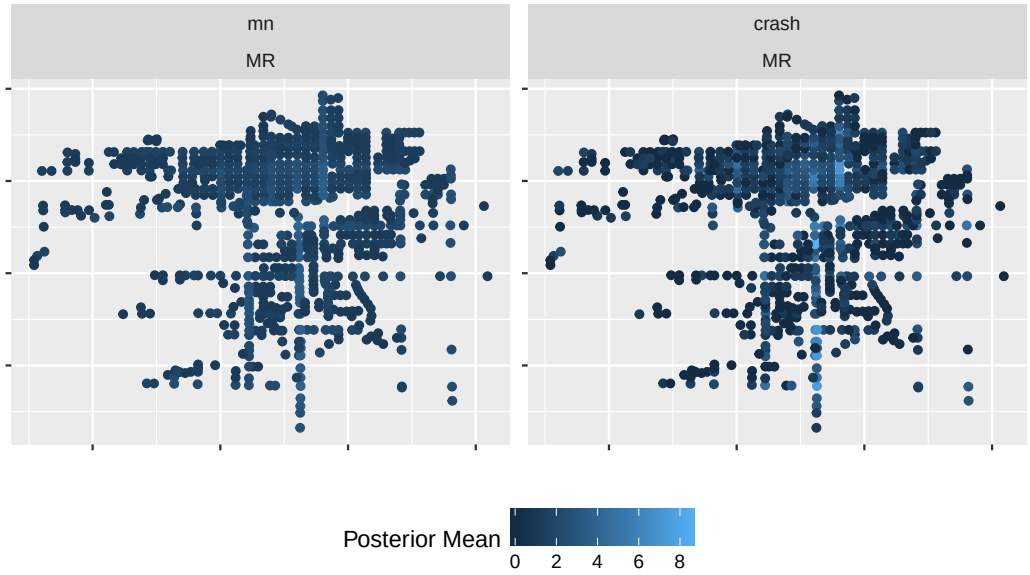

Figure B.7: Marshaltown: Posterior predictive expectation and actual crashes

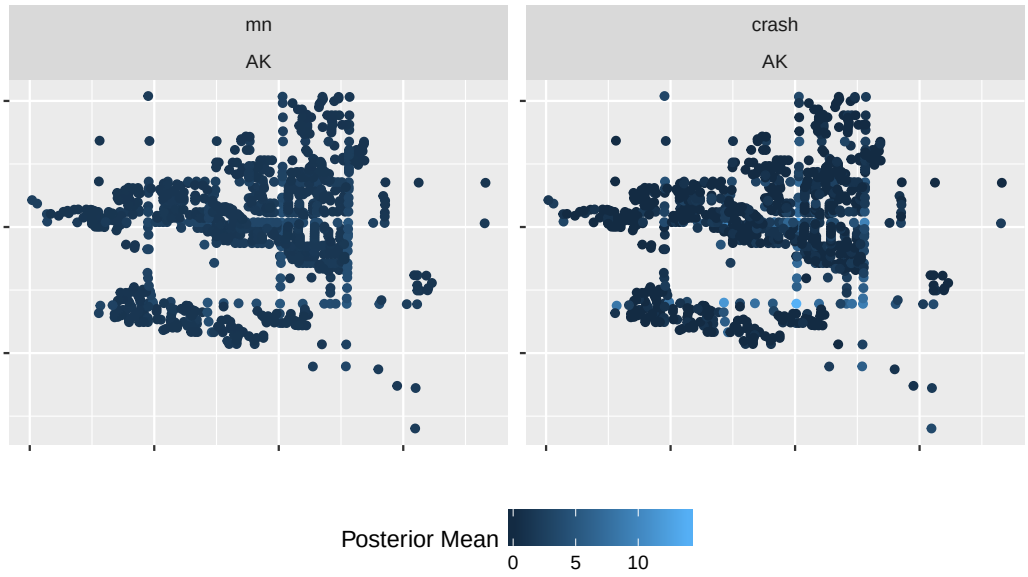

Figure B.8: Ankeny: Posterior predictive expectation and actual crashes