Crispin M. Mutshinda[1] / Andrew J. Irwin[1] / Mikko J. Sillanpää[2]

A Bayesian Framework for Robust Quantitative Trait Locus Mapping and Outlier Detection

[1] Department of Mathematics and Statistics, Dalhousie University, 6316 Coburg Road, Halifax, Nova Scotia B3H 4R2, Canada, E-mail: Crispin.Mutshinda@dal.ca, a.irwin@dal.ca. https://orcid.org/0000-0001-9671-7812.
[2] Department of Mathematical Sciences, University of Oulu, Oulu, Finland, E-mail: Mikko.Sillanpaa@oulu.fi

Abstract:
We introduce a Bayesian framework for simultaneous feature selection and outlier detection in sparse high-dimensional regression models, with a focus on quantitative trait locus (QTL) mapping in experimental crosses. More specifically, we incorporate the robust mean shift outlier handling mechanism into the multiple QTL mapping regression model and apply LASSO regularization concurrently to the genetic effects and the mean-shift terms through the flexible extended Bayesian LASSO (EBL) prior structure, thereby combining QTL mapping and outlier detection into a single sparse model representation problem. The EBL priors on the mean-shift terms prevent outlying phenotypic values from distorting the genotype-phenotype association and allows their detection as cases with outstanding mean shift values following the LASSO shrinkage. Simulation results demonstrate the effectiveness of our new methodology at mapping QTLs in the presence of outlying phenotypic values and simultaneously identifying the potential outliers, while maintaining a comparable performance to the standard EBL on outlier-free data.

Keywords: extended Bayesian LASSO, Hamiltonian Monte Carlo, mean-shift outlier model, QTL mapping, Stan
DOI: 10.1515/ijb-2019-0038
Received: April 9, 2019; **Revised:** January 20, 2020; **Accepted:** February 4, 2020

1 Introduction

In multiple quantitative trait locus (QTL) mapping, the phenotypic trait values $y_1, \ldots, y_n$ of n individuals from the study population are regressed on their genotypes at p markers scored across the genome, to identify the genetic loci associated with variation in the quantitative trait of interest, known as quantitative trait loci, and evaluating their genetic effects. For the purpose of this paper, we restrict attention to experimental crosses derived from two inbred lines such backcross (BC) and double haploid (DH) progenies with one of two only possible genotypes at any locus. The regression model for multiple QTL mapping in BC or DH has the form

$$y_i = \alpha + \sum_{j=1}^{p} x_{i,j}\,\beta_j + \varepsilon_i \tag{1}$$

where α and $x_{i,j}$ denote respectively the population intercept and the genotype of the ith individual at locus j ($j = 1, \ldots, p$) coded herein as 0 for one genotype and 1 for the other, so that β_j is the effect of genotype substitution at locus j from the genotype coded as 0 to the alternative genotype coded as 1. Stochasticity comes in through the random error terms ε_i ($i = 1, \ldots, n$), typically assumed to be independent and normally distributed around zero with common variance σ^2, implying that the phenotypic values are also normally distributed, conditional on the marker genotypes. However, real-world phenotypic data often violate the normality assumption as pointed out by Nascimento et al. [1] among others, due for instance to the presence of outliers. An outlier is defined as an observation whose value deviates so much from other observations in a dataset as to arouse suspicions that it results from a different mechanism than the one underlying the bulk of data [2]. Outliers may arise from various kinds of errors or from the variability inherent in the actual data generating process [3]. While outliers of the first category represent erroneous information that can be rightfully discarded from the data prior to the analysis [4], those of the second category are legitimate data and may be the most interesting cases in various contexts including medical screening, fraud detection, and genetic engineering.

The presence of outlying phenotypic values may unduly impair the robustness of QTL mapping analyses and overly inflate QTL detection error rates [5, 6]. However, QTL mapping models designed to accommodate and automatically detect potential outliers are scarce. In this paper, we introduce such a model, with a focus on

Crispin M. Mutshinda is the corresponding author.

QTL mapping in experimental crosses, and illustrate its implementation with simulated and real data. Before delving into the details of our new outlier-robust QTL mapping model, we start by reviewing the prevailing outlier handling methods.

Existing outlier handling methods fall into two broad categories: diagnostic and robust methods [7]. Diagnostic methods attempt to separate outliers from the bulk of the data, with a view to discarding the outliers and allowing the analysis of the "cleansed" dataset by classical methods such as the ordinary least squares (OLS). In essence, diagnostic methods assume a unique data-generating distribution, deeming as outlier any observation with critically low probability under the hypothesized distribution. This involves the screening of all data points, one at a time, for "outlier-ness" using suitable test quantities such as studentized residuals driven by the leave-one-out scheme [8]. Despite their proven efficiency in single-outlier situations, diagnostic methods suffer from masking and swamping issues in the presence of multiple outliers [9, 10]. Briefly put, we say that an outlier masks a second outlier if the latter emerges as an outlier alone, but not in the presence of the former. Alternatively, we say that an outlier swamps another outlier if the latter arises as an outlier only in the presence of the former. Unlike diagnostic techniques, robust outlier handling methods involve mechanisms for mitigating the impact of potential outliers without the need to remove them from the data prior to the analysis. Robust methods are better suited to handling multiple outliers without suffering from masking and swamping issues affecting diagnostics methods in the presence of multiple outliers. In contrast to diagnostic methods, robust outlier-handling methods assume that the data arose from a mixture of two distributions comprising a "core" distribution supposed to generate the bulk of the data and an alternative distribution held responsible for producing outliers. Mixture models are straightforward to design under the hierarchical Bayesian framework [11] and relatively easy to fit to data by Markov chain Monte Carlo (MCMC) simulation methods [12]. In a regression set-up, an outlier is a data point whose response value deviates markedly from the bulk of response values in the data, whereas a data point with atypical value of the predictor variable is called as a high-leverage point. While both outliers and high-leverage points may unduly affect regression analyses, high-leverage issues are rarely reported for categorical predictors. When dealing with multiple predictors, high leverage points may represent observations with an extreme value for a single predictor or observations with "atypical" combinations of predictor values. This remains a topic for further research.

The two most popular robust outlier-handling methods are the variance-inflation and the mean-shift methods [13], which rest on premises that outliers originate from shifts in the scale (variability) and in the location (mean) of the data-generating process, respectively.

The multiple QTL mapping regression model involving the variance-inflation mechanism has the form

$$y_i = \alpha + \sum_{j=1}^{p} x_{i,j}\,\beta_j + \nu_i\,\varepsilon_i \tag{2}$$

where α, $x_{i,j}$ and β_j are defined as in (1), ε_i is a normally-distributed error with mean zero mean and variance σ^2, and $\nu_i > 0$ $(i = 1, \ldots, n)$ are idiosyncratic variance parameters expected to be one for mainstream instances and larger than one for outliers. Under this model, outlier detection boils down to identifying the instances with variance inflation values v_i significantly larger than 1 [14]. A value of ν_i much less than one (or very close to zero) indicates that y_i is very close to $\alpha + \sum_{j=1}^{p} x_{i,j}\,\beta_j$. Model (2) can be compactly written in matrix form as $\boldsymbol{y} = \alpha\mathbf{1}_n + \boldsymbol{X\beta} + V\varepsilon$, where $\boldsymbol{y} = [y_1, \ldots, y_n]^{\mathrm{T}}$ is the n-vector of phenotypic trait values, $\mathbf{1}_n$ is the n-vector of ones, $\boldsymbol{X}$ is the $n \times p$ matrix of genetic codes, $\boldsymbol{\beta} = \left[\beta_1, \ldots, \beta_p\right]^{\mathrm{T}}$ is the p-vector of genetic effects, $\boldsymbol{V} = diag\,(v_i)$ is the $n \times n$ diagonal matrix with individual-specific variance inflation terms on the main diagonal, and ε is the n-vector of random error terms assumed to be normally distributed with mean zero and common variance σ^2.

The robust mean-shift outlier-handling mechanism can be incorporated in the multiple QTL mapping regression model (1) as

$$y_i = \alpha + u_i + \sum_{j=1}^{p} x_{i,j}\,\beta_j + \varepsilon_i \tag{3}$$

where α, $x_{i,j}$ and β_j and ε_i are defined as in (1), and u_i is the mean-shift term associated with the ith individual $(i = 1, \ldots, n)$. The mean-shift u_i is expected to be zero if y_i is not an outlier and non-zero if y_i is an outlier. Since outliers are exceptions rather than the norm [10], the mean-shift vector $\boldsymbol{u} = [u_1, \ldots, u_n]^{\mathrm{T}}$ is fundamentally sparse as most entries are expected to be zero, in contrast to the variance-inflation vector $\boldsymbol{v} = [v_1, \ldots, v_n]^{\mathrm{T}}$ where most entries are expected to equal 1. In compact matrix notation, eq. (3) becomes $\boldsymbol{y} = \alpha\mathbf{1}_n + \boldsymbol{u} + \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{y}$, $\boldsymbol{\beta}$, $\boldsymbol{X}$, $\mathbf{1}_n$, and $\boldsymbol{\varepsilon}$ are defined as in the matrix form of model (2), and $\boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \sigma^2\boldsymbol{I}_n\right)$. In genome-wide QTL mapping, the number p of candidate markers is typically larger than the sample size n, meaning that the phenotype-to-genotype regression model is generally oversaturated in the sense that it involves more parameters than data points. However, it is expected that most of the markers will have very little or no effect

on the phenotype, by virtue of the documented sparsity of genetic basic of quantitative traits. From a statistical perspective, oversaturated models do not have enough degrees of freedom for parameter estimation by usual techniques such as OLS, the only way out being the bet on the sparsity principle by assuming that most of the effects are null. Since both the p-vector $\boldsymbol{\beta}$ of genetic effects and the n-vector $\boldsymbol{u}$ of mean-shifts are sparse, their concatenation $\tilde{\boldsymbol{\beta}} = [\boldsymbol{\beta} \,|\, \boldsymbol{u}]$ with $\tilde{\beta}_1, \ldots, \tilde{\beta}_p = \beta_1, \ldots, \beta_p$ and $\tilde{\beta}_{p+1}, \ldots, \tilde{\beta}_{p+n} = u_1, \ldots, u_n$ is sparse as well. Letting $\widetilde{\boldsymbol{X}}$ denote the $n \times (p+n)$ matrix obtained by appending the $n \times n$ identity matrix $\boldsymbol{I}_n$ to the design matrix $\boldsymbol{X}$, eq. (3) can be expressed in matrix form as

$$\boldsymbol{y} = \alpha \mathbf{1}_n + \widetilde{\boldsymbol{X}} \tilde{\boldsymbol{\beta}} + \boldsymbol{\varepsilon} \tag{4}$$

This model is inexorably oversaturated since the number $(p+n)$ of regression coefficients, excluding the intercept, exceeds the sample size n. Model (4) re-casts the QTL mapping and outlier detection into a single variable selection problem. Given the inherent sparsity of the extended feature vector $\tilde{\boldsymbol{\beta}}$, sparsity-inducing shrinkage priors, particularly LASSO-type priors, can be used to provide a sparse model representation by shrinking spurious QTL effects and redundant mean-shift terms towards zero to simultaneously map QTLs and identify outlying phenotypic values.

In this paper, we introduce a Bayesian framework for simultaneous variable selection and outlier detection in large-scale and potentially oversaturated regression models, with a focus on quantitative trait locus (QTL) in experimental crosses. The rationale of our new methodology is to insert the robust mean-shift outlier mechanism into the multiple QTL mapping regression model and assign extended Bayesian LASSO (EBL) priors [15–17] on the genetic effects and the mean-shift terms. The EBL priors on the mean-shift terms prevent potential outlying phenotypic values from distorting the genotype-phenotype association, while allowing their detection as cases with outstanding mean-shifts terms following the LASSO shrinkage. The concurrent prescription of EBL priors on the genetic effects and the mean-shift terms allows us to rely on a single decision-rule for QTL identification and outlier detection. We carry out extensive simulations to evaluate our new model, comparing its performance to two alternatives with no outlier handling mechanism namely, the standard EBL with Gaussian errors as proposed by Mutshinda and Sillanpää [15] and the ostensibly robust EBL-t assuming heavy-tailed Student-t rather than Gaussian errors. We fit all three models to the same sets of outlier-contaminated and outlier-free synthetic data replicates and evaluate their performance with regard to the root mean square error and the QTL detection sensitivity. As an application to real-world data, we re-analyze the genetic basis of the time to heading in two-row barley (*Hordeum vulgare* L.) using data from the North American Barley Genome Mapping Project. At the outset, we fit our new model and the EBL to the barley data with actual phenotypic values. We then introduce some outlying phenotypic values and fit the two models to the outlier-contaminated data to evaluate our new model's ability to map QTLs in the presence of outlying phenotypic values and simultaneously identify the outlying cases.

2 Materials & Methods

2.1 Model specification

Our working model is the multiple QTL mapping regression model extended to incorporate mean-shift terms according to eq. (3) *i. e.* $\boldsymbol{y} = \alpha \mathbf{1}_n + \boldsymbol{u} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. In order to achieve robust QTL mapping in the presence of outlying phenotypic values and concurrently identify the potential outliers, we independently assign EBL priors (see below) on the genetic effects β_j $(j = 1, \ldots, p)$ and the mean-shift terms u_i $(i = 1, \ldots, n)$. For conciseness sake, we only describe the EBL prior placed on the mean-shift parameters since the priors assigned on genetic effects have the exact same structure, except for the indexing is on genetic markers rather than on individuals.

For each mean-shift parameter u_i $(i = 1, \ldots, n)$, we assume that $u_i|\sigma_i^2 \sim N\left(0, \sigma_i^2\right)$ and $\sigma_i^2|\lambda_i^2 \sim Exp\left(\lambda_i^2/2\right)$ independently, where σ_i^2 and $\lambda_i > 0$ represent respectively the variance and the regularization parameter specific to the ith individual. The regularization parameter λ_i controls the degree of shrinkage experienced by u_i, with large values of λ_i resulting in stronger shrinkage of u_i towards zero and *vice-versa*. Following the EBL rationale [15–17], each λ_i is factored as $\lambda_i = \delta\eta_i$, where $\delta > 0$ and $\eta_i > 0$ determine respectively the overall sparsity level of the mean-shift vector $\boldsymbol{u} = [u_1, \ldots, u_n]^{\mathrm{T}}$ and the degree of shrinkage specific to u_i. The hyper-parameters δ and η_i are assigned priors and estimated from data alongside the other model parameters. The explicit separation between the overall model sparsity and the distinctive degrees of shrinkage resulting from the factorization $\lambda_i = \delta\eta_i$ of the idiosyncratic regularization parameters under the EBL causes LASSO to adaptively release the shrinkage pressure on important parameters while inflexibly shrinking spurious ones towards zero [15, 16]. This differential shrinkage obviates the tuning of the regularization parameter, which remains a challenging

issue in Bayesian shrinkage analysis [18, 19]. The EBL has been successfully applied with uninformative priors on the hyper-parameters δ and η, including uniform priors [15, 16]. However, when prior information is available, there is no reason for not using it, and the ability to incorporate prior information is one of the attractive features of Bayesian inference.

Since no shrinkage is required on the population intercept α and the error variance σ^2, we use the non-informative priors $\alpha \sim N(0, 100)$ and $\sigma^2 \sim IG(0.1,\ 0.1)$, where $IG(\varphi,\ \omega)$ denotes the inverse-Gamma distribution with parameters φ and ω. The model specification is completed with explicit statements of priors on the hyper-parameters δ and η_i $(i = 1, \ldots, n)$. For the EBL priors on the genetic effects β_j $(j = 1, \ldots, p)$, we denote the counterparts of $\delta > 0$ and $\eta_i > 0$ by $\rho > 0$ and $\kappa_j > 0$, respectively. We assign $Ga(1,\ 1)$ priors independently on δ and ρ and $Unif(0,\ 4)$ on η_i and on κ_j independently for $i = 1, \ldots, n$ and $j = 1, \ldots, p$, where $Ga(\varphi,\ \omega)$ and $Unif(a,\ b)$ denote respectively the Gamma distribution with mean φ/ω and variance φ/ω^2, and the Uniform distribution over the interval $(a,\ b)$, $b > a$. We elaborate on the rationale for choosing a prior of the form for η_i and κ_j further on.

2.2 Model fitting and posterior inferences

We use Markov chain Monte-Carlo methods to simulate from the joint posterior which is not available in closed-form. With the widespread of high-throughput sequencing technologies, the scalability of genetic mapping to large datasets has become critically important. Standard MCMC algorithms such as the Gibbs sampler [20] and the Metropolis-Hastings algorithm [21, 22] may be prohibitively slow in the presence of strongly correlated parameters [[23], p. 567]. Hamiltonian Monte Carlo (HMC) [24] provides a valuable alternative to *status quo* MCMC algorithms, with tremendous promises for faster mixing and better scalability to complex and high-dimensional models owing to its more efficient exploration of the parameter space. HMC considers the target log probability density and its gradient and then generates an extremely efficient Markov transition. While a HMC sampler is difficult to set up, it is easy to implement through the Bayesian probabilistic language Stan [25], with the default No-U-Turn sampler (NUTS) allowing fast exploration of the most important parts of the parameter space, regardless of the covariance structure [26]. In this paper, we rely on HMC simulation *via* Stan.

Under the EBL, we expect the mean-shifts of outlying phenotypic values to undergo less shrinkage towards zero relative to the bulk of data. The overall sparsity of the mean-shift vector $\boldsymbol{u}$ and the degree of shrinkage experienced by the bulk of mean-shift values are controlled by the "global" hyper-parameter δ, whereas the idiosyncratic hyper-parameters η_i distinguish the outliers from the bulk of phenotypic values by being consistently less than 1 for outliers and *vice versa*. On this premise, Mutshinda and Sillanpää [16] developed a fully Bayesian decision rule for variable selection under the EBL, which boils down to the test of whether or not η_i is less than 1. In practice, one may define Bayes factors (BFs [27]) to provide a rule of thumb for hypothesis testing. Suppose that prior distributions $p(\eta_i)$ are independently assigned to the hyper-parameters η_i with some pre-specified "outlier-ness" probability $\Pr(\eta_i < 1) = w$ (the prior probability that y_i is an outlier) which corresponds to prior odds of $w/(1-w)$ to 1 for any phenotypic value being an outlier (H_1) *versus* being consistent with the bulk of data (H_0). Consequently, the Bayes factor, $BF^i_{1,0}$, for outlier-ness of the ith phenotypic value is simply the ratio of the posterior odds to the prior odds for H_1. Choosing a prior of the form $Unif(0, w)$, $w > 1$ simplifies the computation of the prior probabilities $\Pr(\eta_i < 1)$ and $\Pr(\eta_i \geq 1)$ as $\Pr(\eta_i < 1) = 1/w$ and $\Pr(\eta_i \geq 1) = (w-1)/w$, so that

$$BF^i_{1,0} = \frac{(1-w)\Pr(\eta_i < 1|data)}{w\Pr(\eta_i \geq 1|data)} \tag{5}$$

Since w is predetermined, the key input in (4) is $s_i = \Pr(\eta_i \langle 1|\, data)$, the posterior probability that the ith phenotypic value is an outlier, and this probability can be straightforwardly evaluated from MCMC samples as

$$\Pr(\eta_i < 1|data) = \frac{1}{N}\sum_{k=1}^{N} I(\eta_i^k < 1|data) \tag{6}$$

where N, $(\eta_i^k < 1|data)$ and $I(.)$ denote respectively, the number of post burn-in MCMC samples, the kth post burn-in posterior MCMC sample for η_i, and the indicator function taking the value 1 when its argument is true and the value 0 otherwise.

The Bayes factor $BF_{1,0}$ for hypothesis H_1 *versus* H_0 is usually interpreted against the Jeffreys' scale of evidence [28]. On the Jeffreys' scale, as slightly amended by Kass and Raftery [27], $BF_{1,0} < 1$ indicates a negative support for H_1 (support for H_0), $1 \leq BF_{1,0} < 3$ indicates a support for H_1 that is "not worth more than a bare mention", $3 \leq BF_{1,0} < 20$ indicates a positive support for H_1, whereas $20 \leq BF_{1,0} < 150$ and $BF_{1,0} \geq 150$ indicate

respectively a strong and a very strong support for H_1. Based on this scale of evidence, we consider 3 as cut-off Bayes factor for outlier and QTL detection in the simulation study and the real data analysis, noting that, under the prior $\Pr(\eta_i < 1) = 0.25$ assumed here, $BF^i_{1,2} \geq 3$ is equivalent to $s_i = \Pr(\eta_i < 1 \,|\, data) \geq 0.5$. Similarly, we assume that $\Pr(\kappa_j < 1) = 0.25$ *a priori* for QTL presence at any locus j so that $BF^j_{1,0} \geq 3$ corresponds to $r_j = \Pr(\kappa_j < 1 \,|\, data) \geq 0.5$.

3 Report on the simulation study

In this section, we report on a simulation study designed to evaluate our new model by comparing its performance to two alternatives namely, the standard EBL with Gaussian residuals [15] and a seemingly robust version of the EBL, herein the EBL-t, assuming heavy-tailed Student-t rather than Gaussian errors. For computational convenience, we parameterize the Student-t distribution as a scale mixture of normal distributions with scaled-inverse χ^2 mixing variances [29, 30]. More specifically, if $Z|\sigma^2 \sim N(0, \sigma^2)$ and $\sigma^2 \sim IG(v/2, v/2)$, then $Z \sim t_v$, where t_v denotes the Student-t distribution with v degrees of freedom. In the EBL-t model, we pick $v = 5$, which is small enough to guarantee the heavy tailed-ness of the ensuing Student-t distribution. Before delving into the details of the data simulation process and subsequent analyses, we start by describing the outlier-labelling rule and the model performance measures considered here.

3.1 Outlier labeling rule

A commonly used outlier labeling procedure is the "k standard deviation rule" where any observation beyond k standard deviations from the mean value for a specific $k > 0$ is flagged as an outlier. However, this approach is only sensible when the data distribution is approximately normal. The inter-quartile range multiplier (IQRM) approach pioneered by Tukey [30] provides a valuable alternative to the k standard deviation rule. This approach involves finding the inter-quartile range $\text{IQR} = Q_3 - Q_1$ of the data, where Q_1 and Q_3 denote respectively the first and the third quartile of the data distribution. Multiplying the IQR by a tuning parameter g we define the range $[Q_1 - g \times \text{IQR},\ Q_3 + g \times \text{IQR}]$ outside of which observations are flagged as outliers. This method is applicable to data with skewed or non bell-shaped distributions, providing the sample size is not too small. In keeping with Tukey [31], we use $g = 1.5$ for outlier labeling in the simulation study.

3.2 Performance measures for model evaluation

We consider two performance measures for model evaluation namely, the root mean square error and the QTL detection sensitivity, which are briefly described below.

3.2.1 The root mean squared error

The root mean squared error is defined as $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}$, where y_i is the observed value of the response variable for the ith individual ($i = 1, 2, \ldots, n$) and $\widehat{y_i}$ the model fitted counterpart defined by $\widehat{y_i} = \hat{\alpha} + \hat{u}_i + \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ when the model involves mean-shift terms ore $\widehat{y_i} = \hat{\alpha} + \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ otherwise.

3.2.2 QTL detection sensitivity

Sensitivity quantifies the avoidance of false negatives whereas specificity does the same for false positives. Let TP and TN denote respectively the proportions of QTL loci (positives) and the proportion of non-QTL loci (negatives) correctly identified as such, and let also FP and FN denote respectively, the proportion of negatives predicted to be positives and the proportion of positives predicted to be negatives. A model with high QTL detection accuracy is expected to have high TP and TN and low FP and FN, resulting in high QTL detection sensitivity $\text{Sn} = \text{TP}/(\text{TP} + \text{FN})$ and specificity $\text{Sp} = \text{TN}/(\text{TN} + \text{FP})$. Since oversaturation ($p > n$) typically induces low TP and high FN, we consider QTL detection sensitivity as performance measure for QTL detection ability. It is worth mentioning that when mapping QTLs using tight marker maps, inherent high dependencies

between genotypes of nearby markers may cause adjacent markers to QTLs loci to emerge instead of the actual QTL loci [32, 33], and these should not be considered as false positives.

3.3 Data simulation and statistical analyses

We generate synthetic phenotypic data replicates using a dense marker dataset simulated through the WinQTL Cartographer 2.5 program [34], and involving 1000 markers for 100 backcross progeny (10 times as many markers as individuals). The markers span two chromosomes with 500 markers each and only 1 cM gap between consecutive markers, implying a high level of correlation between adjacent markers. We generate the phenotypic values from the regression model $\boldsymbol{y} = \alpha \mathbf{1}_n + \boldsymbol{u} + \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ with $\alpha = 0$, $\boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \sigma^2 \boldsymbol{I}_n\right)$, and $\sigma = 1$, assuming sparse underlying biology involving just six QTLs at loci 25, 202, 451, 633, 815 and 986 with respective QTL effects as β_{25}=−3, $\beta_{202} = 2$, $\beta_{451} = 2$, $\beta_{633} = 3$, $\beta_{815} = -3$, $\beta_{986} = 2$, and β_j set to zero for the rest of loci. To generate outlier-contaminated phenotypic data (scenario 1), we set all entries of the mean-shift vector **u** to zero, except for five individuals, namely individual number 9, 37, 58, 71 and 83 selected to have outlying phenotypic values, and whose mean-shifts are uniformly drawn between 10 and 12. In order to assess potential statistical losses incurred by fitting our new robust model to outlier-free data, we also generate outlier-free phenotypic values (scenario 2) assuming the same model and same QTL effects as in the scenario 1, but with all mean-shifts $u_1, \ldots u_n$ set equal to zero (no outlier). Figure 1 shows a histogram, a boxplot and a normal Q-Q plot of representative phenotypic data replicates under scenarios 1 (bottom) and 2 (top).

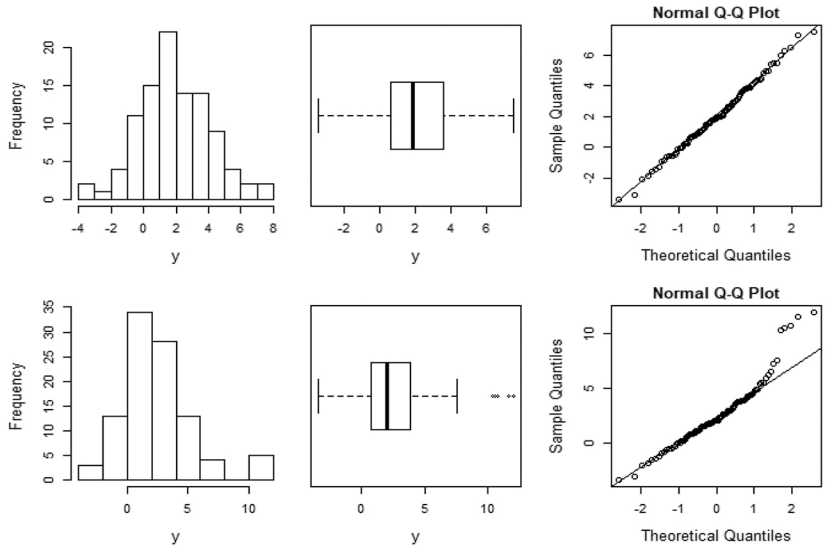

Figure 1: Histogram, boxplot and normal Q-Q plot of a typical outlier-free (top) and a typical outlier-contaminated (bottom) phenotypic data replicate.

The boxplot of the outlier-contaminated phenotypic data (Figure 1, bottom center panel) shows the five simulated outliers clearly standing out from the bulk. We generated 20 synthetic data replicates under each scenario, fit the three models (our new model, the EBL-t and the EBL) to each data replicate, and evaluate the root mean squared error (RMSE) and the QTL detection sensitivity (Sn) for each model over the simulated data replicates.

3.4 Simulation results

We used Hamiltonian Monte Carlo, through Stan (Stan Development Team, 2018), to simulate from the joint posterior distribution of the model parameters. At the outset, we ran 6000 iterations of three Markov chains. The approximate running time was 40 minutes on a personal computer equipped with a 64-bit CORE i5 Intel processor @ 2.50 Hz. After about 1000 iterations, the chains reached the target distribution and mixed well, jumping freely around the parameter space. During the simulation study, we ran a single MCMC for 4000 iterations and discarded the first 2000 iterations as burn-in, thinning the remainder by a factor of 5. The results based on our new model clearly separate QTLs form non-QTL loci on the one hand (Figure 2(a)) and outlying phenotypic values from the bulk on the other hand (Figure 2(c)).

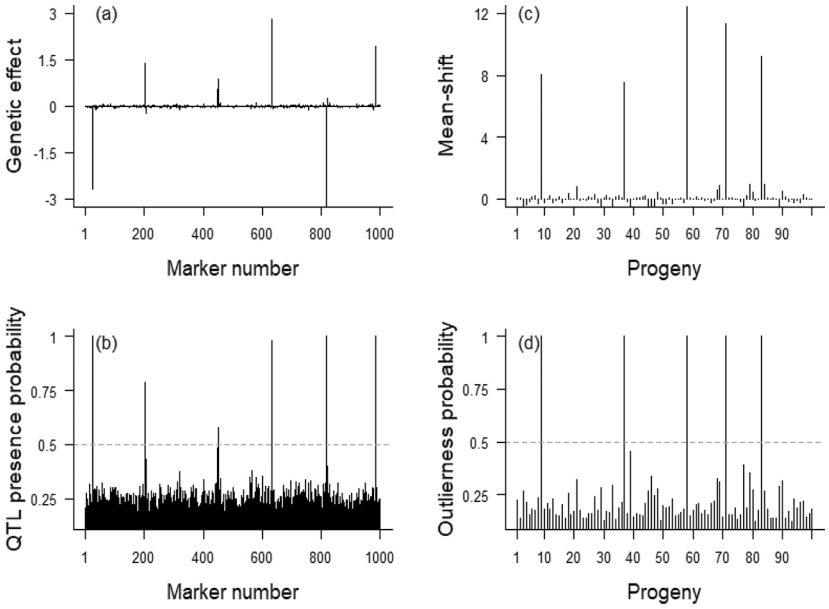

Figure 2: Posterior means of (a) genetic effects, (b) QTL presence probabilities, (c) mean-shift terms, and (d) outlier-ness probabilities averaged over 20 synthetic outlier-contaminated data replicates for our new model. The dashed horizontal lines in the lower panels indicate the posterior probability cut-off value 0.5 for QTL and outlier detection corresponding to a Bayes factor of 3 under our prior assumption.

In addition, the posterior means of QTL presence and outlier-ness probabilities shown in panels 2b and 2d respectively are way beyond the detection threshold for the simulated QTL loci and the outlying phenotypic values. However, the LASSO shrinkage on genetic effects tends to be excessively strong under the two models with no outlier detection mechanism (the EBL-t and the EBL), irrespective of whether or not a locus harbours a QTL for the quantitative trait of interest (Figure 3(a), (c)). As a result, QTL loci tend to go undetected under these two models, particularly when the magnitude of the genetic effect is low.

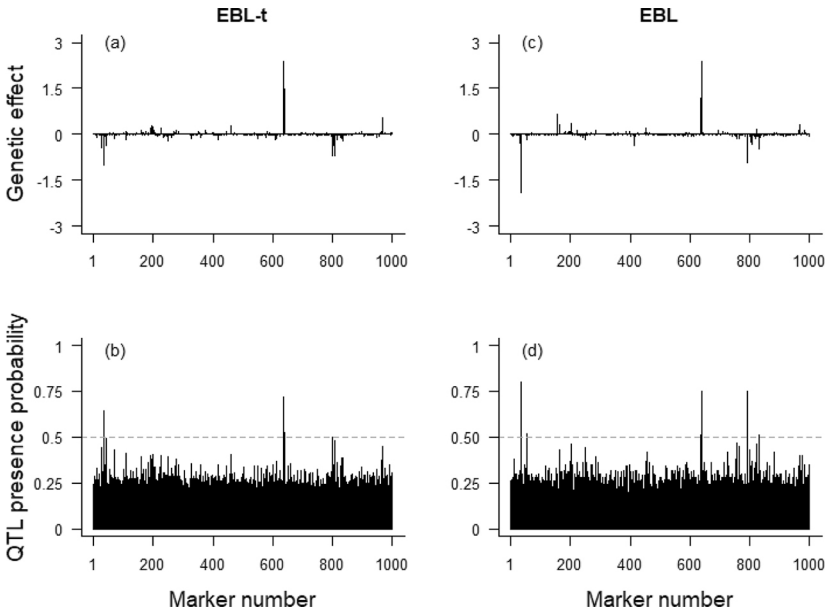

Figure 3: Posterior means of genetic effects for the EBL-t (a) and the EBL (c) and posterior means of QTL presence probabilities for the EBL-t (b) and the EBL (d) averaged over 20 synthetic outlier-contaminated data replicates. The dashed horizontal lines in the lower panels indicate the posterior probability cut-off value 0.5 for QTL detection corresponding to a Bayes factor of 3 under our prior assumption.

Our proposed model outperforms the EBL-t and the EBL on outlier-contaminated data (under scenario 1) with regard to all performance measures under consideration, standing out with lower root mean squared error and higher QTL detection sensitivity (Table 1).

Table 1: Performance measures namely, the root mean squared error (RMSE) and the QTL detection sensitivity (Sn) averaged over 20 simulated outlier-contaminated data replicates for the three models under consideration namely, our new model, the EBL and the EBL-t.

Scenario	Model	RMSE	Sn (%)(%)
Scenario 1	NewModel	0.43	97
Outlier-contaminated	EBLt	4.50	58
phenotypes	EBL	2.63	61

On outlier-free data, our model neatly separates QTLs from non-QTL loci, with the simulated QTL loci standing clearly out with posterior means of genetic effects largely different from zero (Figure 4(a)) and QTL presence probabilities far beyond the detection threshold 0.5 corresponding to a Bayes factor 3 under our prior assumption (Figure 4(b)). Interestingly, all mean-shift terms undergo a stringent shrinkage towards zero on outlier-free data under our new model (Figure 4(c)), preventing any phenotypic value from emerging as an outlier, and confining the posterior outlier-ness probabilities well below the outlier detection threshold (Figure 4(d)).

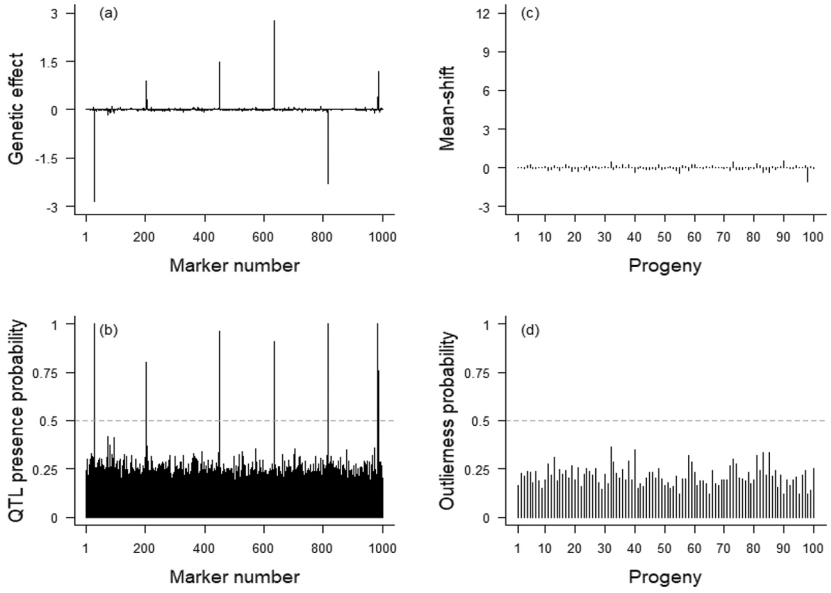

Figure 4: Posterior means of (a) genetic effects, (b) QTL presence probabilities, (c) mean-shift terms, and (d) outlier-ness probabilities averaged over 20 synthetic outlier-free data replicates for our new model. The dashed horizontal lines in the lower panels indicate the posterior probability cut-off value 0.5 for QTL and outlier detection corresponding to a Bayes factor of 3 under the assumed priors.

On outlier-free data replicates, the posterior means of genetic effects are significantly different from zero and the posterior QTL presence probabilities are largely beyond the QTL detection threshold for the six simulated QTLs under both the EBL-t and the EBL (Figure 5).

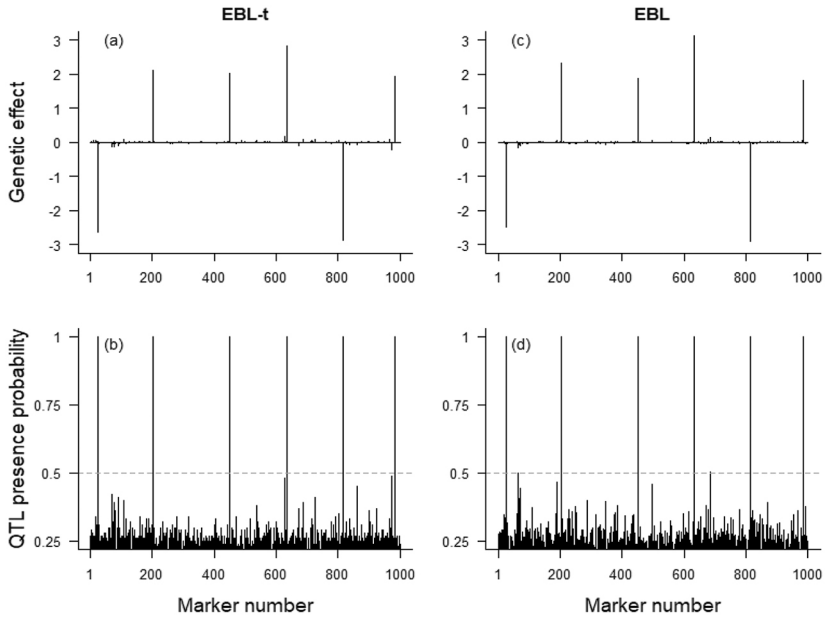

Figure 5: Posterior means of genetic effects for the EBL-t (a) and the EBL (c) and posterior means of QTL presence probabilities for the EBL-t (b) and the EBL (d) averaged over 20 synthetic outlier-free data replicates. The dashed horizontal lines in the lower panels indicate the posterior probability cut-off value 0.5 for QTL detection corresponding to a Bayes factor of 3 under our prior assumption.

On outlier-free data, the posterior means of genetic effects and QTL presence probabilities for the EBL-t and the EBL displayed in Figure 5 are broadly similar to corresponding estimates under our new model (Figure 4(a), (b)). The performance measures are also similar between the three models, with our new model and the EBL achieving nearly 100 % QTL detection sensitivity and low RMSEs (Table 2).

Table 2: Performance measures namely, the root mean squared error (RMSE) and QTL detection sensitivity (Sn) averaged over 20 simulated outlier-free data replicates for the three models under consideration (our new model, the EBL and the EBL-t).

Scenario	Model	RMSPE	Sn (%)
Scenario 2	NewModel	0.44	98
Outlier-free	EBLt	0.47	93
phenotypes	EBL	0.40	98

With the prevalence of high-throughput genotyping and sequencing technologies, the scalability of genetic mapping procedures to large data has become an important issue. We examined the time complexity for our new model relative to the standard EBL. We fitted both models to the same sets of synthetic data replicates with gradually increasing number of markers and monitored the running time of each model. The plot of the average running time against the number of markers displayed in Figure 6 is roughly linear in the number of markers for both models, implying a time complexity of $O(n)$.

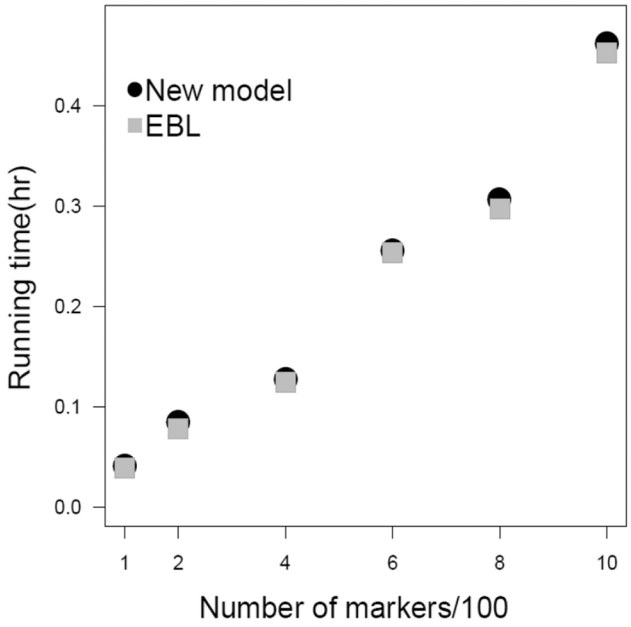

Figure 6: Running time (in hours) against the number of markers (in hundreds) on outlier-contaminated data for our new robust model (black filled circles) and the EBL (grey filled squares).

Our analysis demonstrates the value of Hamiltonian Monte Carlo as a fast and scalable alternative to *status quo* Markov chain Monte Carlo techniques such as the Gibbs sampler [20] and the Metropolis-Hastings algorithm [21, 22].

Linkage analysis (QTL mapping) and association analysis refer to two different ways of mapping QTLs for target phenotype. The former is usually performed either using inbred line crosses or sets of known relatives whereas the latter, also known as "linkage disequilibrium mapping", takes advantage of historic linkage disequilibrium to link phenotypes to genotypes using dense marker maps, with a view to uncovering genetic associations in sets of largely unrelated individuals. Association mapping usually considers high marker density to map QTLs to a fine resolution. However, association mapping suffers from a number of drawbacks including: (1) the dependence of the power of detecting QTLs on allele frequencies, reflected in a lack of power to detect rare alleles. (2) A high sensitivity to population structure, leading to many false positives if population structure is not properly accounted for. (3) Scalability issues due to the high dimensionality of the feature space in contrast to linkage mapping which generally relies on few hundreds to few thousands markers. We tailored our methodology to linkage mapping by focusing on experimental crosses to avoid dealing with the range of issues associated with association mapping. However, the methodology is applicable to association mapping as well. Scalability to ultra-high dimensional feature spaces may be achieved by resorting to sure independence screening (SIS), a two-stage learning approach due to Fan and Lv [35], whereby a large-scale screening is first applied to reduce the model dimensionality from p features to a moderate number $d < n$, where n is the sample size n. SIS involves the ranking of the initial p features in decreasing order of their association measures with the response variable, y, typically $\left|cor\left(\boldsymbol{y}, \boldsymbol{x}_j\right)\right|$ or $\left|\beta_j\right|$ and retaining the top d covariates with the largest marginal association values with $\boldsymbol{y}$. Inference is then conducted on the reduced feature space [35, 36]. Fan and Lv [35] showed that such a ranking of features possesses the sure independence screening property, meaning that it retains the important variables in the model with probability very close to 1.

4 Real data analysis

In this section, we re-analyze the genetic basis of the time (number of days) to heading in two-row barley (*Hordeum vulgare* L.) using real data from the North American Genome Mapping project [37, 38]. The barley data involve the phenotypes (days to heading) of 145 double haploid (DH) progenies, along with their genotypes at 127 markers covering seven chromosomes with a 10.5 cM gap between consecutive markers. To remain in the $p > n$ setting, we only consider a subset of 100 individuals randomly selected without replacement among the 145 progenies in the barley data. The empirical distributions of the restricted and the full phenotypic data are comparable as expected since the former is a simple random sample from the latter. We first analyzed the data with the actual barley phenotypic data using our new model and the EBL. We then introduced a few outlying phenotypic values and analyzed the outlier-contaminated data using the two models. We generated the

outlier-contaminated phenotypic data by replacing the phenotypic values of progeny 7, 48, and 83 by random numbers uniformly drawn on the interval $[Q_3 + 2 \times IQR, Q_3 + 3 \times IQR]$ where Q_3 and IQR denote respectively the third quartile and the inter-quartile range of the phenotypic data under consideration. Figure 7 shows the histograms, boxplots and normal Q-Q plots of the outlier-free (top panels) and outlier-contaminated (bottom panels) phenotypic data.

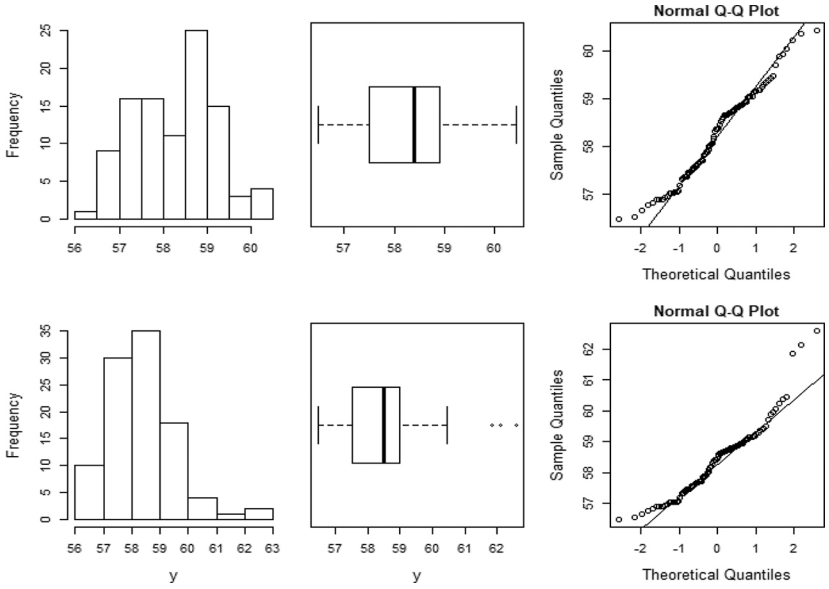

Figure 7: Histograms, boxplots and normal Q-Q plots of the outlier-free (top) and outlier-contaminated (bottom) barley phenotypic values.

We carried out the model fitting to data by MCMC simulation via Hamiltonian Monte Carlo implemented in Stan, under the prior probabilities $\Pr(\kappa_j < 1) = 0.20$ for QTL presence at any locus and $\Pr(\eta_i < 1) = 0.20$ for any of the phenotypic values being an outlier. These prior assumptions correspond to prior odds of 1 to 4 for QTL presence at any locus and for any of the phenotypic values being an outlier. We ran 10,000 iterations of three parallel Markov chains following a burn-in period of 4,000 iterations, and thinned the post burn-in MCMC samples by a factor of 25. The posterior means of the QTL effects and QTL presence probabilities for the new model and the EBL are similar on outlier-free data. The two models identified roughly the same set of loci as QTLs namely, locus 5, 10, 47, 59, 63, 86, 112, 119 and 120 (Figure 8), consistent with previous analyses of the genetic basis of time to heading in barley using the original phenotypic trait values [15, 16, 38].

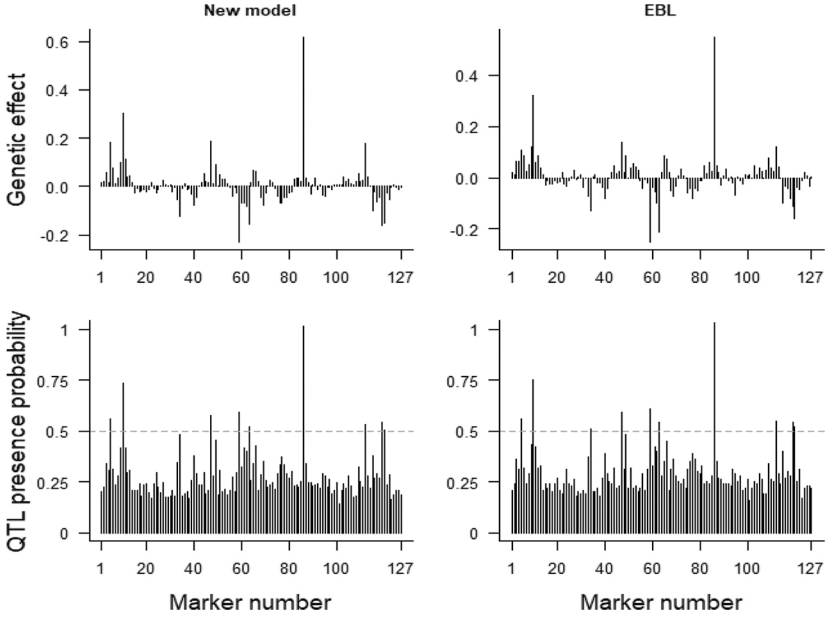

Figure 8: Posterior means of the QTL effects for our new model (top-left) and the EBL (top-right), and posterior means of QTL presence probabilities for our new model (bottom-left) and the EBL (bottom-right) on outlier-free data. The dashed horizontal lines in the lower panels indicate the posterior probability cut-off for declaring QTLs, which corresponds to a BF of 3 under the 1:3 prior odds for QTL presence assumed here.

While our new model identified roughly the same QTLs on outlier-contaminated and on outlier-free data (Figure 9, left panels) more than half of the presumed QTL loci on outlier-free data under the EBL went undetected as outliers were introduced (Figure 9, right panels).

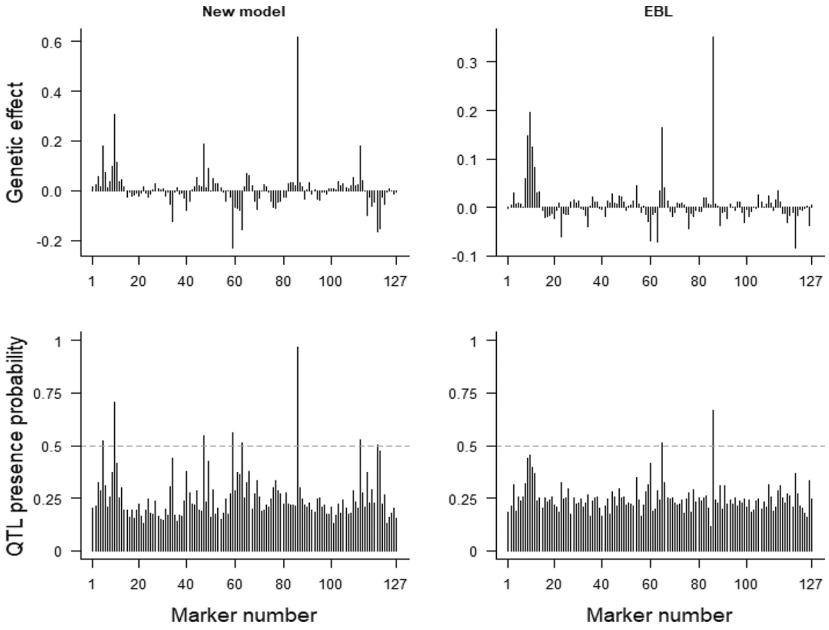

Figure 9: Posterior means of the QTL effects for our new model (top-left) and the EBL (top-right), and posterior means of QTL presence probabilities for our new model (bottom-left) and the EBL (bottom-right) on outlier-contaminated data. The dashed horizontal lines in the lower panels indicate the posterior probability cut-off for declaring QTLs, which corresponds to a BF of 3 under the 1:3 prior odds for QTL presence assumed here.

Turning to the outlier identification aspect of our new model, the stringent shrinkage of all mean-shift terms towards zero on outlier-free data (scenario 1) caused their magnitudes to be virtually zero *a posteriori* (Figure 10, top-left panel). In addition, all posterior outlier-ness probabilities were constrained to much lower values than the outlier detection threshold (Figure 10, bottom-left panel), thereby preventing false outlier detection. Likewise, the mean-shift terms associated with non-outlying cases underwent a stringent shrinkage towards zero on outlier-contaminated data, while the five simulated outliers stood out with posterior mean-shift values way larger than zero (Figure 10, top-right panel) and posterior outlier-ness probability far beyond the detection threshold (Figure 10 bottom-right panel).

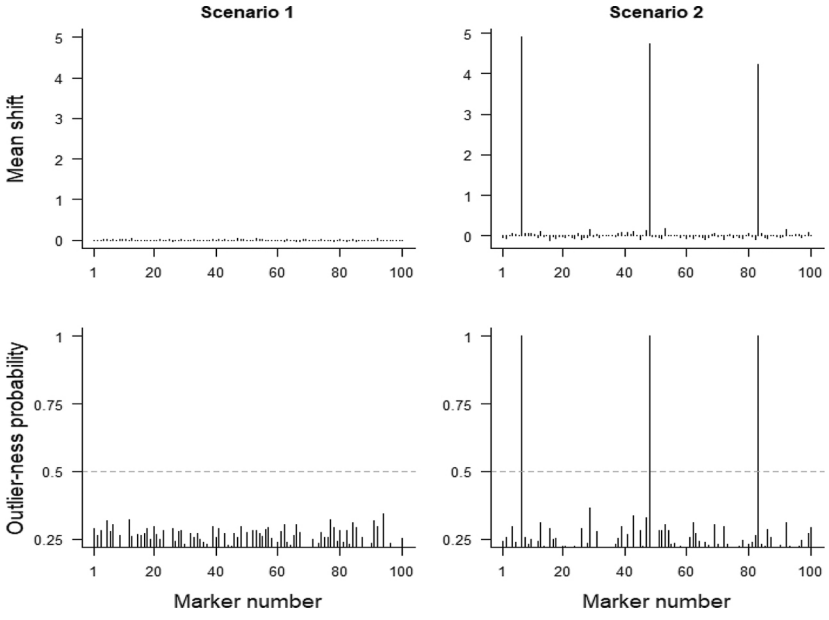

Figure 10: Posterior outlier-ness probabilities for our new model under the two scenarios, with the dashed horizontal lines in the bottom panels indicating the posterior probability cut-off for declaring outliers which, under the 1:3 prior odds for any phenotypic value being an outlier assumed here, corresponds to a BF of 3.

Overall, the barley data analysis corroborates the effectiveness of our new approach for robust QTL mapping with simultaneous outlier detection at a relatively low computational cost.

5 Discussion

In this paper, we introduced a Bayesian framework for concurrent feature selection and outlier detection in sparse high-dimensional regression, with a focus on QTL mapping in experimental crosses. The rationale of our methodology is to integrate the robust mean-shift outlier handling mechanism into the regression model for multiple QTL mapping, and prescribe LASSO shrinkage on the genetic effects and the mean shift terms using the flexible extended Bayesian LASSO (EBL) prior, which provides much flexibility with regard to variable selection. The EBL priors assigned to the mean-shift terms precludes outlying phenotypic values from misrepresenting the genotype-phenotype association, while allowing their detection as cases with outstanding mean shift values following the LASSO shrinkage. The concurrent prescription of EBL priors on the genetic effects and the mean-shift allowed us to rely on a single decision rule for QTL identification and outlier detection, with hypotheses tests for QTL identification and outlier detection following as by-products of the MCMC model fitting process [16].

We carried out extensive simulations to evaluate our model, comparing its performance on synthetic outlier-contaminated and on outlier-free data to the EBL and the ostensibly robust EBL-t assuming heavy-tailed Student-t rather than Gaussian errors. Our new model outperformed the EBL and the EBL-t on outlier-contaminated data, standing out with lower mean squared errors and higher QTL detection sensitivity, and effectively detected the simulated outliers. Interestingly, our new model performed comparably to the EBL and the EBL-t on outlier-free data with the LASSO inflexibly shrinking all mean shift terms towards zero (Figure 4(c) and Figure 10, top-left panel).

We analyzed the genetic basis of the time to heading in two-row barley (*Hordeum vulgare* L.) using data from the North American Genome Mapping project, with the standard EBL serving as benchmark for performance evaluation. The barley genetic data involve 127 markers for 145 double haploid individuals. In order to remain in the ($p > n$) setting, we considered a sample of 100 individuals. We first fitted our new model and the EBL to the data with actual phenotypic trait values. We then introduced three outlying phenotypic values and fitted the two models to the outlier-contaminated data. The results substantiated the ability of our new model to provide robust QTL mapping in the presence of outlying phenotypic values and identify the potential outliers. The QTL mapping results on outlier-contaminated data and outlier-free data were comparable under our new model, and consistent with findings of previous analyses of the barley data under consideration [15, 16, 38], corroborating the robustness of our proposed model to the presence of outliers. While the posterior estimates of QTL effects and QTL presence probabilities were comparable between our new model and the EBL on outlier-free data, most of the QTLs detected under the EBL on outlier-free data (Figure 8) went undetected in the presence of outliers (Figure 9).

Robust regression and outlier detection in a high-dimensional regression set-up are fundamental problems in statistics with several applications across disciplines. Generally, robust regression focuses on estimating regression coefficients in the presence of outliers without necessarily localizing the potential outliers. A documented approach to robust regression is to replace the normality assumption on the error terms by heavy-tailed alternatives, typically the Student-t distribution or the Laplacian or Double Exponential distribution [39]. The use of thick-tailed distributions for robust prediction of complex trait has a long history in the animal breeding literature [40–46]. Gianola et al. [40] present a robust alternative to best linear unbiased prediction for genomic prediction purposes, which involves a linear model with Student-t or Laplace rather than Gaussian errors. However, Lambert-Lacroix and Zwald [47] emphasize that for normally distributed data, models assuming heavy-tailed residual distributions such as the Student-t or the Laplace tend to be less efficient than mean-shifted models with normally distributed errors, which may explain the poor performance of the EBL-t relative to the robust model proposed here.

Variable selection in sparse high-dimensional regression models and outlier detection through mean shifting are both sparse model representation problems. Our new methodology integrates these two problems into a single sparse robust regression model and tackles them concurrently from a Bayesian regularization perspective using the flexible EBL prior and its attendant decision rule for variable selection [16].

The prevalence of next-generation sequencing has led to an explosion in marker number, raising new modeling and computational challenges in linkage studies [48]. With high-density marker data, the genotypes of consecutive markers tend to co-vary widely. However, LASSO selects a single variable from a group of strongly correlated predictors [40]. In practice, it may be desirable to account for all potential predictors particularly in genomic selection based on dense molecular markers, which entails estimating the simultaneous effects of all

genes and combining the estimates to predict the total genomic breeding value [49–51]. An extension of the model proposed here with the L_2 constraint on the vector β of genetic effects may be required according to the elastic net method [52] or the approach of Xu [53] to promote a grouping effect.

It is worth emphasizing that outliers do not necessarily represent flawed data. Atypical instances may be the focus of interest in applications from various domains including genetics, bioinformatics, finance, climate, etc. The model proposed here provides, in combination with Hamiltonian Monte Carlo simulation, an expedient approach to robust analysis of sparse high-dimensional regression models in an effective and scalable manner.

Two words of caution are in order before closing this discussion. (1) The simulated data may not be representative of the large data sets typically encountered by plant and animal breeders. (2) The magnitude of simulated outliers were chosen to be sizably large for illustration purposes. Nevertheless, our analyses provide proof-of-concept for the robust mean-shift EBL model introduced here. It is up to the scientific community to assess its value in real-world situations.

Acknowledgements

CMM and AJI were supported by the Simons Collaboration on Computational Biogeochemical Modeling of Marine Ecosystems/CBIOMES (Grant ID: 549935, AJI).

Code and data availability

The R and Stan code used in the analyses as well as the simulated marker data are available in the online Supplementary Material.

References

[1] Nascimento M, Silva FF, de Resende MD, Cruz CD, Nascimento AC, Viana JM, et al. Regularized quantile regression applied to genome-enabled prediction of quantitative traits. Genet Mol Res. 2017;16:gmr16019538.
[2] Hawkins DM. Identification of outliers. London: Chapman and Hall, 1980.
[3] Anscombe FJ. Rejection of outliers. Technometrics. 1960;2:123–47.
[4] Liu H, Shah S, Jiang W. On-line outlier detection and data cleaning. Comput Chem Eng. 2004;28:1635–47.
[5] Jansen RC, Stam P. High resolution of quantitative traits into multiple loci via interval mapping. Genetics. 1994;136:1447–55.
[6] Feingold E. Regression-based quantitative-trait-locus mapping in the twenty-first century. Am J Hum Genet. 2002;71:217–22.
[7] Barnett V, Lewis T. Outliers in statistical data. 3rd ed. Chichester, UK: John Wiley & Sons, 1994.
[8] Weisberg S. Applied linear regression. 2nd ed. New York, NY: Wiley, 1985.
[9] Hadi AS, Simonoff JS. Procedures for the identification of multiple outlier in linear models. J Am Stat Assoc. 1993;88:1264–72.
[10] She Y, Owen AB. Outlier detection using nonconvex penalized regression. J American Stat Assoc. 2011;106:626–39.
[11] Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis. 2nd ed. London, England: Chapman and Hall, 2013.
[12] Gilks WR, Richardson S, Spiegelhalter DJ, eds. Markov Chain Monte Carlo in practice. London, England: Chapman and Hall, 1996.
[13] Guttman I. Care and handling of univariate or multivariate outliers in detecting spuriousity: a Bayesian approach. Technometrics. 1973;15:723–38.
[14] Box GE, Tiao GC. A Bayesian approach to some outlier problems. Biometrika. 1968;55:119–29.
[15] Mutshinda CM, Sillanpää MJ. Extended Bayesian LASSO for multiple quantitative trait loci mapping and unobserved phenotype prediction. Genetics. 2010;186:1067–75.
[16] Mutshinda CM, Sillanpää MJ. A decision rule for quantitative trait locus detection under the extended Bayesian LASSO model. Genetics. 2012;192:1483–91.
[17] Onogi A, Iwata H. VIGoR: variational Bayesian inference for Genome-Wide regression. J Open Res Software. 2016;4:e11. DOI: 10.5334/jors.80.
[18] Park T, Casella G. The Bayesian LASSO. J Am Stat Assoc. 2008;103:681–6.
[19] Mutshinda CM, Sillanpää MJ. Swift block-updating EM and pseudo-EM procedures for Bayesian shrinkage analysis of quantitative trait loci. Theor Appl Genet. 2012;125:1575–87.
[20] Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell. 1984;6:721–41.
[21] Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E. Equations of state calculations by fast computing machines. J Chem Phys. 1953;21:1087–92.
[22] Hastings WK. Monte Carlo sampling methods using Markov Chains and their applications. Biometrika. 1970;57:97–109.
[23] Gelman A, Hill J. Data analysis using regression and multilevel/hierarchical models. New York: Cambridge University Press; 2007.

[24] Monnahan CC, Thorson JT, Branch TA. Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. Meth Ecol Evol. 2017;8:339–48.
[25] Stan Development Team. Stan modeling language users guide and reference manual. Version 2.18.0, 2018. http://mc-stan.org.
[26] Hoffman MD, Gelman A. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. J Mach Learn Res. 2014;15:1351–81.
[27] Kass RE, Raftery AE. Bayes factors. J Am Stat Assoc. 1995;90:773–95.
[28] Jeffreys H. The theory of probability. 3rd ed. Oxford, UK: Oxford University Press; 1961.
[29] Andrews DF, Mallows CL. Scale mixtures of normal distributions. J R Stat Soc B. 1974;36:99–102.
[30] West M. On scale mixtures of normal distributions. Biometrika. 1987;74:646–8.
[31] Tukey JW. Exploratory data analysis. Cambridge, MA: Addison-Wesley, 1977.
[32] Xu S. An expectation-maximization algorithm for the Lasso estimation of quantitative trait locus effects. Heredity. 2010;105:483–94.
[33] Mutshinda CM, Sillanpää MJ. Bayesian shrinkage analysis of QTLs under shape-adaptive shrinkage priors, and accurate re-estimation of genetic effects. Heredity. 2011;107:405–12.
[34] Wang S, Basten CJ, Zeng Z-B. Windows QTL cartographer 2.5. Raleigh, NC: Department of Statistics, North Carolina State University, 2006.
[35] Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. J R Statl Soc Series B: Stat Method. 2008;70:849–911.
[36] Fan J, Song R. Sure independence screening in generalized linear models with np-dimensionality. The Ann Stat. 2010;38:3567–604.
[37] Tinker NA, Mather DE, Rosnagel BG, Kasha KJ, Kleinhofs A. Regions of the genome that affect agronomic performance in two-row barley. Crop Sci. 1996;36:1053–62.
[38] Knürr T, Läärä E, Sillanpää MJ. Genetic analysis of complex traits via Bayesian variable selection: the utility of a mixture of uniform priors. Genet Res. 2011;93:303–18.
[39] von Rohr P, Hoeschele I. Bayesian QTL mapping using skewed student-t distributions. Genet Sel Evol. 2002;34:1–21.
[40] Gianola D, Cecchinato A, Naya H, Schon C-C. Prediction of complex traits: robust alternatives to best linear unbiased prediction. Front Genet. 2018;9:195.
[41] Strandén I, Gianola D. Attenuating effects of preferential treatment with student-t mixed linear models: a simulation study. Genet Sel Evol. 1998;30:565.
[42] Strandén I, Gianola D. Mixed effects linear models with t-distributions for quantitative genetic analysis: a Bayesian approach. Genet Sel Evol. 1999;31:25–42.
[43] Rosa GJ, Padovani CR, Gianola D. Robust linear mixed models with normal/independent distributions and Bayesian MCMC implementation. Biometrical J. 2003;45:573–90.
[44] Rosa GJ, Gianola D, Padovani CR. Bayesian longitudinal data analysis with mixed models and thick-tailed distributions using MCMC. J Appl Stat. 2004;31:855–73.
[45] Cardoso FF, Rosa GJ, Tempelman RJ. Multiple-breed genetic inference using heavy-tailed structural models for heterogeneous residual variances. J Anim Sci. 2005;83:1766–79.
[46] Varona L, Mekkawy W, Gianola D, Blasco A. A whole-genome analysis using robust asymmetric distributions. Genet Res. 2006;88:143–51.
[47] Lambert-Lacroix S, Zwald L. Robust regression through the Hubers criterion and adaptive LASSO penalty. Electron J Stat. 2011;16:1015–53.
[48] Mutshinda CM, Noykova N, Sillanpää MJ. A hierarchical Bayesian approach to multi-trait clinical quantitative trait locus modeling. Front Genet. 2012;3:97.
[49] Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819–29.
[50] Gianola D, Perez-Enciso M, Toro MA. On marker-assisted prediction of genetic value: beyond the ridge. Genetics. 2003;163:347–65.
[51] Ogutu JO, Torben S-S, Piepho H-P. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. BMC Proc. 2012;6:S10.
[52] Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc B. 2005;67:301–20.
[53] Xu S. Genetic mapping and genomic selection using recombination breakpoint data. Genetics. 2013;195:1103–15.

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/ijb-2019-0038).