Published in final edited form as:
Stat Appl Genet Mol Biol. ; 11(6): . doi:10.1515/1544-6115.1803.

Hierarchical Shrinkage Priors and Model Fitting for High-dimensional Generalized Linear Models

Nengjun Yi[1] and **Shuangge Ma**[2]

[1]Department of Biostatistics, Section on Statistical Genetics, University of Alabama at Birmingham, Birmingham, AL 35294

[2]School of Public Health, Yale University, New Haven, CT 06520

Abstract

Genetic and other scientific studies routinely generate very many predictor variables, which can be naturally grouped, with predictors in the same groups being highly correlated. It is desirable to incorporate the hierarchical structure of the predictor variables into generalized linear models for simultaneous variable selection and coefficient estimation. We propose two prior distributions: hierarchical Cauchy and double-exponential distributions, on coefficients in generalized linear models. The hierarchical priors include both variable-specific and group-specific tuning parameters, thereby not only adopting different shrinkage for different coefficients and different groups but also providing a way to pool the information within groups. We fit generalized linear models with the proposed hierarchical priors by incorporating flexible expectation-maximization (EM) algorithms into the standard iteratively weighted least squares as implemented in the general statistical package R. The methods are illustrated with data from an experiment to identify genetic polymorphisms for survival of mice following infection with *Listeria monocytogenes*. The performance of the proposed procedures is further assessed via simulation studies. The methods are implemented in a freely available R package BhGLM (http://www.ssg.uab.edu/bhglm/).

Keywords

Adaptive Lasso; Bayesian inference; Generalized linear model; Genetic polymorphisms; Grouped variables; Hierarchical model; High-dimensional data; Shrinkage prior

1. Introduction

Scientific experiments routinely generate very many highly correlated predictor variables for complex response outcomes. For example, genetic studies of complex traits (e.g. QTL (quantitative trait loci) mapping and genetic association analysis) usually genotype numerous genetic markers across the entire genome, with closely linked markers being nearly collinear. The main goals of such studies are to identify which predictor variables are associated with responses of interest and/or to predict outcomes of new individuals using all the predictor variables. To control for potential confounding effects, it is desirable to simultaneously fit as many predictors as possible in a model. Due to high-dimensional and correlated structure, however, classical generalized linear models are usually nonidentifiable, and thus a variety of penalization methods and Bayesian hierarchical models have been proposed to solve the problem.

Penalization approach introduces constraints (or penalties) on the coefficients and estimates the parameters by maximizing the penalized objective function (Hastie, Tibshirani and Friedman, 2001). A promising penalization method is the Lasso of Tibshirani (1996), which

uses the L_1-penalty, $\lambda\Sigma_{j=1}^{J}|\beta_j|$, where $\lambda(\geq 0)$ is the tuning parameter controlling the amount of penalty. Bayesian approach places prior distributions on the coeffificients and summarizes posterior inference using MCMC algorithms or by finding posterior modes (Gelman et al., 2003). A commonly used prior distribution is the double-exponential distribution, $p\left(\beta_j\right)=\frac{\lambda}{2}e^{-\lambda|\beta_j|}$, where λ (≥ 0) is the shrinkage parameter controlling the amount of shrinkage (Park and Casella, 2008; Tibshirani, 1996a, b; Yi and Xu, 2008). Under this prior, the posterior modes of the coefficients correspond to the Lasso estimates (Park and Casella, 2008; Tibshirani, 1996a). Another widely used prior distribution is the Student-t distribution, $p(\beta_j) = t_\nu(0,s^2)$, with the degrees of freedom ν and the scale s controlling the amount of shrinkage (Gelman et al., 2008; Yi and Xu, 2008). The family of the Student-t distributions includes various distributions as special cases, for example, Normal $\beta_j \sim N(0, s^2)$ ($\nu = \infty$), Jeffreys' prior $\beta_j \propto 1/|\beta_j|$ ($\nu = s = 0$) and Cauchy ($\nu = 1$). One remarkable feature of the above distributions is that they can be expressed as scale mixtures of normal distributions (Park and Casella, 2008; Yi and Xu, 2008). This hierarchical formulation leads to the development of straightforward MCMC and Expectation-Maximization (EM) algorithms (Figueiredo, 2003; Gelman et al., 2008; Kyung et al., 2010; Park and Casella, 2008; Yi and Banerjee, 2009; Yi and Xu, 2008).

The above approaches have shown excellent performance in many situations, however they have some limitations. First, the Lasso (also Student-t) uses a unique tuning parameter to equally penalize all coefficients, and thus can either include a number of irrelevant variables or over-shrink large coefficients. This is particularly critical for sparse high-dimensional data (as in most genetic studies) where among hundreds or thousands of variables only a few have detectable effects. To address this issue, Zou (Zou, 2006) introduced the adaptive Lasso that uses a weighted L_1-penalty $\Sigma_{j=1}^{J}\lambda_j|\beta_j|$, where $\lambda_j=\lambda/|\widehat{\beta}_j^0|$ with $\widehat{\beta}_j^0$ being some preliminary estimate of β_j such as the least-squares estimate. The intuition of the adaptive lasso is to use different penalty parameters for different coefficients and thus differently shrink coefficients. Although theoretically attractive, the practical performance of the adaptive Lasso heavily depends on the quality of the initial estimates. Secondly, some previous methods cannot appropriately address the hierarchical structures of predictor variables. To accommodate the grouping structure, several group approaches based on composite penalties have been proposed. Both the inner and outer penalties can take multiple forms, including ridge, Lasso, elastic net and others. If the inner penalty is ridge as with group Lasso, then the approaches have an "all in or all out" property for variables within the same groups. If the inner penalty is Lasso-type, then two level selection may be achieved. Despite theoretical effectiveness of such methods, they may suffer a high computational cost.

In this article, we adopt a Bayesian approach and propose hierarchical prior distributions for high-dimensional generalized linear models with grouped predictor variables. We express our hierarchical priors as scale mixtures of normal distributions: $\beta_j \sim N\left(0, \tau_j^2\right)$, $\tau_j^2 \sim \text{Expon}\left(\lambda_j^2/2\right)$ or $\text{Inv}-\chi^2\left(\nu, s_j^2\right)$, and assume that the scale parameters λ_j^2 (or λ_j) or s_j^2 further follow gamma distributions with unknown group-specific hyperparameters. The variable-specific scale parameters are similar in spirit to the variable-specific penalties in the adaptive Lasso of Zou (2006), but they can be easily estimated in our Bayesian framework. Similar Bayesian approaches have been proposed. Griffin and Brown (2010) (Griffin and Brown, 2010) and Leng et al. (2010) (Leng, Minh Ngoc Tran and Nott, 2010) have generalized the Bayesian Lasso by including variable-specific scale parameters λ_j^2 in the exponential prior with gamma mixing distributions. Sun et al. (2010) (Sun, Ibrahim and Zou, 2010) and Armagan et al. (2010) (Armagan, Dunson and Lee, 2010) assume the parameters

λ_j rather than λ_j^2 in the exponential prior to follow gamma distributions and show that this treatment induces simpler posterior distributions facilitating study of properties and implementation. Carvalho et al. (2009, 2010) (Carvalho, Polson and Scott, 2009, 2010) proposed the horseshoe prior that is similar to our prior with $\tau_j^2 \sim \text{Inv} - \chi^2\left(1, s_j^2\right)$. More recently, Lee et al (2012) (Lee et al., 2012) discussed the use of Bayesian sparity priors obtained through hierarchical mixtures of normals for the analysis of genetic association. However, these previous methods either prefix the hyperparameters in gamma distributions or have not considered grouped variables. Our hierarchical priors consist of both the group-specific and variable-specific parameters, and thus not only adopt different shrinkage for different coefficients and different groups but also providing a way to pool the information within groups.

To develop our algorithms for fitting generalized linear models with the proposed hierarchical prior distributions, we first derive the conditional posterior distributions for all parameters. These posterior distributions may allow us to implement MCMC algorithms. In this paper, however, we focus on much faster EM algorithms for finding posterior modes. We incorporate our EM algorithms into the usual iteratively weighted least squares (IWLS) for fitting classical generalized linear models as implemented in the general statistical package R. This strategy allows us to take advantage of the existing algorithm and leads to stable and flexible computational tools.

The rest of the paper is organized as follows. We first introduce hierarchical generalized linear models with grouped variables in Section 2 and then describe the hierarchical prior distributions in Section 3. We derive our EM-IWLS algorithms for fitting the hierarchical GLMs with the proposed priors in Section 4 and describe the implementation in R in Section 7. In Section 6, we discuss the relationship between the proposed models and existing approaches. Section 7 illustrates the methods with data from an experiment to identify genetic polymorphisms for survival of mice following infection with *Listeria monocytogenes*. Section 8 demonstrates the performance of the proposed procedures via simulation studies. Finally, some concluding remarks and potential extensions are discussed in Section 9.

2. Hierarchical Generalized Linear Models with Grouped Predictors

We consider the problem of variable selection and coefficient estimation in generalized linear models with a large number of coefficients or highly correlated predictor variables. The observed values of a continuous or discrete response are denoted by $y = (y_1, \cdots, y_n)$. We assume that the predictor variables can be divided into K groups, G_k, $k = 1, \cdots, K$, and the k-th group G_k contains J_k variables, where $K \geq 1$ and $J_k > 1$. There may be multiple ways of defining the groups. In genetic studies, for example, we can use genomic regions or candidate genes, or the types of the effects (e.g., additive and dominance effects) to construct groups. We also include in the model some variables (e.g., gender indicator, age, etc.) that do not belong to any groups.

A generalized linear model consists of three components: the linear predictor η, the link function h, and the data distribution p (Gelman et al., 2003; McCullagh and Nelder, 1989). The linear predictor for the i-th individual can be expressed as

$$\eta_i = \beta_0 + \sum_{j=1}^{J_0} x_{ij}\beta_j + \sum_{k=1}^{K}\sum_{j\in G_k} z_{ij}\beta_j = X_i\beta \quad (1)$$

where β_0 is the intercept, x_{ij} and z_{ij} represent observed values of ungrouped and grouped variables, respectively, β_j is a coefficient, the notation $j \in G_k$ indicates the group of variable j, X_i contains all variables, and β is a vector of all the coefficients and the intercept. For simplicity, we denote $X_i = (1, x_{i1}, \cdots, x_{iJ})$ and $\beta = (\beta_0, \beta_1, \cdots, \beta_J)'$, where $J = \sum_{g=0}^{K} J_g$ is the total number of variables.

The mean of the response variable is related to the linear predictor via a link function h:

$$E(y_i|X_i) = h^{-1}(X_i\beta) \quad (2)$$

The data distribution is expressed as

$$p(y|X\beta, \phi) = \prod_{i=1}^{n} p(y_i|X_i\beta, \phi) \quad (3)$$

where φ is a dispersion parameter, and the distribution $p(y_i|X_i,\beta,\varphi)$ can take various forms, including Normal, Gamma, Binomial, and Poisson distributions. Some GLMs, for example the Poisson and the Binomial models, do not require a dispersion parameter; that is, φ is fixed at 1.

Generalized linear models with many coefficients or highly correlated variables can be nonidentifiable classically. An approach to overcoming the problem is to use Bayesian inference. We use a hierarchical framework to construct priors for coefficients. At the first level, we assume an independent normal distribution with mean 0 and variable-specific variance τ_j^2 for each coefficient β_j:

$$\beta_j|\tau_j^2 \sim N\left(0, \tau_j^2\right) \quad \text{for} \quad j=1, \cdots, J \quad (4)$$

The variance parameters τ_j^2 directly control the amount of shrinkage in the coefficient estimates; if $\tau_j^2=0$, the coefficient β_j is shrunk to zero, and if $\tau_j^2=\infty$, there is no shrinkage. Although these variances are not the parameters of interest, they are useful intermediate quantities to estimate for easy computation of the model.

We treat the variances τ_j^2 as unknowns and will further assign prior distributions to them as discussed in the next sections. Enough data is available to estimate the intercept β_0 and the dispersion parameter φ. Thus, we can use any reasonable non-informative prior distributions for these two parameters; for example, $p\left(\beta_0|\tau_0^2\right) = N\left(0, \tau_0^2\right)$ with τ_0^2 set to a large value, and $p(\log\varphi) \propto 1$.

3. Hierarchical Adaptive Shrinkage Priors for Variance Parameters

The prior distributions for the variance parameters play a crucial role on variable selection and coefficient estimation. We consider two types of priors; the first is the half-Cauchy distribution for τ_j, and the second is the exponential prior distribution for τ_j^2.

3.1. Half-Cauchy prior distribution for τ_j

A half-Cauchy distribution can be expressed as

$$\tau_j|\alpha_{k[j]} \sim \text{half–Cauchy}\left(\alpha_{k[j]}\right) \propto \left(\tau_j^2+\alpha_{k[j]}^2\right)^{-1}, \quad \text{for} \quad j=1,\cdots,J; \quad k=1,\cdots,K \quad (5)$$

which has a peak at zero and a scale parameter $\alpha_{k[j]}$, where the subscript $k[j]$ indexes the group k that the j-th predictor belongs to. The scale parameter controls the amount of shrinkage in the variance estimates; small scales force most of the variances close to zero (Gelman, 2006). For grouped variables, we treat the scale parameters $\alpha_{k[j]}$ as random variables and assign a noninformative prior distribution on the logarithmic scale, $p(\log \alpha_{k[j]}) \propto 1$. The common scale $\alpha_{k[j]}$ induces a common distribution for variables within a group. For ungrouped variables, we cannot estimate the scale and thus preset $\alpha_{k[j]}$ to a known value (say $\alpha_j = 1$).

There is no direct way to fit the model with the above half-Cauchy prior. Therefore, we use the hierarchical formulation of the half-Cauchy distribution. The half-Cauchy variable can be expressed as the product of the absolute value of a normal random variable with variance $\alpha_{k[j]}^2$ and the square root of an inverse-χ^2 variable with degree-of-freedom 1 and scale 1, i.e., $\tau_j = |s_j|\eta_j$, where $s_j \sim N\left(0, \alpha_{k[j]}^2\right)$ and $\eta_j^2 \sim \text{Inv}-\chi^2(1,1)$. For computational simplicity, we deal with τ_j^2 rather than τ_j and express the prior distribution of τ_j^2 hierarchically:

$$\tau_j^2|s_j \sim \text{Inv}-\chi^2\left(\nu, s_j^2\right), \quad s_j^2|b_{k[j]} \sim \text{Gamma}\left(a, b_{k[j]}\right), \quad p(\log b_k) \propto 1 \quad (6)$$

where $\nu = 1$, $a = 0.5$, and $b_{k[j]}=0.5\alpha_{k[j]}^{-2}$. We describe the computational algorithm for arbitrary values of ν and a in the next section. The half-Cauchy prior distribution corresponds to $\nu = 1$ and $a = 0.5$, and is free of user-chosen hyperparameters.

3.2. Exponential prior distribution for τ_j^2

The second prior distribution assumes that the variances τ_j^2 follow exponential or equivalently gamma distributions with variable-specific hyperparameters s_j:

$$\tau_j^2|s_j \sim \text{Expon}\left(s_j^2/2\right) = \text{Gamma}\left(1, s_j^2/2\right) \quad (7)$$

The hyperparameter s_j controls the amount of shrinkage in the variance estimate; a large value of s_j forces the variance τ_j^2 closer to zero. We treat the hyperparameters s_j as random variables with the Gamma hyper-prior distributions:

$$s_j|b_{k[j]} \sim \text{Gamma}\left(a, b_{k[j]}\right) \quad (8)$$

where the subscript $k[j]$ indexes the group k that the j-th predictor belongs to, and a and $b_{k[j]}$ are two hyperparameters. As shown later, placing the gamma prior on s_j rather than s_j^2 leads to a simpler posterior distribution of s_j that is independent of the variance τ_j^2 and thus facilitates computation.

As default, we set $a = 0.5$. Theoretically, we need not worry so much about how to select a, because the shrinkage can be determined by $b_{k[j]}$ and thus the hyperparameter a has less effect on inference. For ungrouped variables, we preset $b_{k[j]}$ to a known value (say $b_{k[j]} = 0.5$). For grouped variables, we treat the scale parameters $b_{k[j]}$ as unknown parameters and

assign a noninformative prior distribution on the logarithmic scale of $b_{k[j]}$, i.e., $p(\log b_{k[j]}) \propto 1$.

The above prior distributions include group-specific parameters b_k and variable-specific parameters s_j. The group-specific parameters provide a way to pool the information among variables within a group and also to induce different shrinkage for different groups, while the variable-specific parameters allow different shrinkage for different variables. Indeed, as we will see in later numerical experiments, the estimates of b_k for groups which include no important variables will be much different from those with important variables, and the estimates of s_j for zero β_j will be different from those for nonzero β_j.

Conditional on the scale parameters s_j, the hierarchical priors (6) and (7) induce the Student-t distributions $\beta_j \sim t_\nu\left(0, s_j^2\right)$ and the double-exponential distributions $\beta_j \sim DE(0, s_j)$ on the coefficients, respectively. Hereafter, we refer these two prior distributions as to *hierarchical t* and *hierarchical double-exponential* distributions, respectively. For the hierarchical t prior, we use the hierarchical Cauchy distribution as a default choice (i.e., setting $\nu = 1$).

4. EM-IWLS Algorithm for Model Fitting

We fit the above hierarchical generalized linear models by estimating the marginal posterior modes of the parameters (β, φ). We modify the usual iterative weighted least squares (IWLS) for fitting classical GLMs and incorporate an EM algorithm into the modified IWLS procedure. The EM-IWLS algorithm increases the marginal posterior density of the parameters (β, φ) at each step and thus converges to a local mode. Our EM algorithm treats the unknown variances τ_j^2 and the hyperparameters s_j and $b_{k[j]}$ as missing data and estimates the parameters (β, φ) by averaging over these missing values. At each step of the iteration, we replace the terms involving the parameters (β, φ) and the missing values $(\tau_j^2, s_j, b_{k[j]})$ by their conditional expectations, and then update the parameters (β, φ) by maximizing the expected value of the joint log-posterior density,

$$\log \quad p\left(\beta, \phi, \tau^2, s, b|y\right) \propto \sum_{i=1}^{n} \log \quad p\left(y_i|X_i\beta, \phi\right) - \frac{1}{2}\sum_{j=0}^{J}\frac{\beta_j^2}{\tau_j^2} + \text{terms that do not depend on} \quad (\beta, \phi) \quad (9)$$

4.1. Conditional posterior distributions and conditional expectations

For the E-step of the algorithm, we take the expectation of the above joint log-posterior density with respect to the conditional posterior distributions of the variances and the hyperparameters. For the hierarchical t prior distribution, the conditional posterior distributions are

$$\tau_j^{-2}|\beta_j, s_j^2 \sim \frac{\chi_{1+\nu}^2}{\nu s_j^2 + \beta_j^2} \quad (10)$$

$$s_j^2|\tau_j^2, b_{k[j]} \sim \text{Gamma}\left(\nu/2 + a, \quad \tau_j^{-2}\nu/2 + b_{k[j]}\right) \quad (11)$$

$$b_k|\left\{s_j^2; j \in G_k\right\} \sim \text{Gamma}\left(aJ_k, \sum_{j \in G_k}^{J_k} s_j^2\right) \quad (12)$$

Therefore, we have the conditional expectations

$$E\left(\tau_j^{-2}|\beta_j, s_j^2\right)=\frac{1+\nu}{\nu s_j^2+\beta_j^2} \quad (13)$$

$$E\left(s_j^2|\tau_j^2, b_{k[j]}\right)=(\nu/2+a)\,/\left(\tau_j^{-2}\nu/2+b_{k[j]}\right) \quad (14)$$

$$E\left(b_k|\left\{s_j^2; j\in G_k\right\}\right)=aJ_k/\sum_{j\in G_k}^{J_k} s_j^2 \quad (15)$$

For the hierarchical double-exponential prior distribution, the conditional posterior distributions are

$$\tau_j^{-2}|\beta_j, s_j \sim \text{Inv}-\text{Gauss}\left(s_j/|\beta_j|, s_j^2\right) \quad (16)$$

$$s_j|\beta_j, b_{k[j]} \sim \text{Gamma}\left(1+a, |\beta_j|+b_{k[j]}\right) \quad (17)$$

$$b_k|\left\{s_j; j\in G_k\right\} \sim \text{Gamma}\left(aJ_k, \sum_{j\in G_k}^{J_k} s_j\right) \quad (18)$$

Therefore, we have the conditional expectations

$$E\left(\tau_j^{-2}|\beta_j, s_j\right)=s_j/|\beta_j| \quad (19)$$

$$E\left(s_j|\beta_j, b_{k[j]}\right)=(1+a)\,/\left(|\beta_j|+b_{k[j]}\right) \quad (20)$$

$$E\left(b_k|\left\{s_j; j\in G_k\right\}\right)=aJ_k/\sum_{j\in G_k}^{J_k} s_j \quad (21)$$

It is worth noting that under the hierarchical double-exponential prior the posterior $p(s_j|\beta_j, b_{k[j]})$ is independent of the variance τ_j^2, which may speed up convergence.

4.2. Estimating (β,φ) conditional on the prior variances τ_j^2

Since only the terms $\beta_j^2/\tau_j^2 \quad (j=1,\cdots,J)$ include both the parameters and the missing values, only the conditional expectations $E\left(\tau_j^{-2}|\beta_j, s_j\right)$ directly affect the M-step. Given the prior variances τ_j^2, $j = 0, \cdots, J$, we thus estimate (β,φ) by fitting the generalized linear model with a normal prior distribution for the coefficients the generalized linear model $y_i \sim p(y_i|X_i,\beta,\varphi)$ with the normal priors $\beta_j|\tau_j^2 \sim N\left(0, \tau_j^2\right)$ (Gelman et al., 2008; Yi and Banerjee, 2009; Yi, Kaklamani and Pasche, 2011). Following the usual iteratively weighted least squares (IWLS) algorithm for fitting generalized linear models (as implemented in the glm function

in R), we approximate the generalized linear model likelihood $p(y_i \mid X_i,\beta,\varphi)$ by the weighted normal likelihood

$$p(y_i|X_i\beta,\phi) \approx N\left(z_i|X_i\beta, w_i^{-1}\phi\right) \quad (22)$$

where the 'normal response' z_i and 'weight' w_i are called the pseudo-response and pseudo-weight, respectively. The pseudo-response and pseudo-weight are calculated by

$$z_i=\widehat{\eta}_i - \frac{L'(y_i|\widehat{\eta}_i)}{L''(y_i|\widehat{\eta}_i)}, \quad w_i=-L''(y_i|\widehat{\eta}_i) \quad (23)$$

where $\widehat{\eta}_i=X_i\widehat{\beta}$, $L(y_i|\widehat{\eta}_i)=\log \quad p\left(y_i|X_i\widehat{\beta},\phi=1\right)$, $L'(y_i \mid \eta_i) = dL(y_i \mid \eta_i)/d\eta_i$, $L''(y_i \mid \eta_i) = d^2L(y_i \mid \eta_i)/d\eta_i^2$, and $\widehat{\beta}$ is the current estimate of β.

The prior $\beta_j|\tau_j^2 \sim N\left(0,\tau_j^2\right)$ can be incorporated into the weighted normal likelihood as an 'additional data point' 0 (the prior mean) with corresponding 'explanatory variables' equal to 0 except x_j which equals 1 and a 'residual variance' τ_j^2 (Gelman et al., 2003; Gelman et al., 2008). Therefore, we can update β by running the augmented weighted normal linear regression

$$z_* \sim N(X_*\beta, \quad \phi\Sigma_*) \quad (24)$$

where $z_*=\begin{pmatrix} z \\ 0 \end{pmatrix}_{(n+J+1)\times 1}$ is the vector of all z_i and all $(J+1)$ prior means 0, $X_*=\begin{pmatrix} X \\ I_{J+1} \end{pmatrix}_{(n+J+1)\times(J+1)}$ is constructed by the design matrix X of the regression $z_i \sim N\left(X_i\beta, w_i^{-1}\phi\right)$ and the identity matrix $I_{(J+1)}$, and $\Sigma_*=\text{diag}\left(w_1^{-1},\cdots,w_n^{-1},\tau_0^2/\phi,\cdots,\tau_J^2/\phi\right)$ is the diagonal matrix of all pseudo-weights and prior variances. With the augmented X_*, this regression is identified and thus the resulting estimate $\widehat{\beta}$ is well defined and has finite variance, even if the original data are high-dimensional and have collinearity or separation that would result in nonidentifiability of the classical maximum likelihood estimate (Gelman et al., 2008). Therefore, we obtain the estimate of β, $\widehat{\beta}=\left(X_*'\Sigma_*^{-1}X_*\right)^{-1}X_*'\Sigma_*^{-1}z_*$, and its variance $\text{Var}\left(\widehat{\beta}\right)=\left(X_*'\Sigma_*^{-1}X_*\right)^{-1}\widehat{\phi}$. If a dispersion parameter, φ, is present, we can update φ at each step of the iteration by

$$\widehat{\phi}=\frac{1}{n}\left(z_*-X_*\widehat{\beta}\right)^T\Sigma_*^{-1}\left(z_*-X_*\widehat{\beta}\right) \quad (25)$$

4.3. EM-IWLS algorithm and inference

In summary, the EM-IWLS algorithm can be described as follows:

1. Start with a crude parameter estimate.
2. For $t = 1, 2, \cdots$:

   E-step: Calculate the conditional expectations

   1. $\tau_j^{-2(t)}=E\left(\tau_j^{-2}|\beta_j^{(t-1)}, s_j^{(t-1)}\right)$

2. $s_j^{2(t)} = E\left(s_j^2 | \tau_j^{(t)}, b_{k[j]}^{(t-1)}\right)$ for the hierarchical t prior, or $s_j^{(t)} = E\left(s_j | \beta_j^{(t-1)}, b_{k[j]}^{(t-1)}\right)$ for the hierarchical double-exponential prior
3. $b_k^{(t)} = E\left(b_k | \left\{s_j^{(t)}; j \in G_k\right\}\right)$

M-step:

1. Based on the current value of β, calculate the pseudo-data $z_i^{(t)}$ and the pseudo-weights $w_i^{(t)}$.
2. Update β by running the augmented weighted normal linear regression.
3. If φ is present, update φ.

We choose the starting values of the parameters as follows. An initial estimate to the linear predictor η is found by the standard method as implemented in the R function glm (This standard method is not affected by the number of variables and can be safely used in high-dimensional settings). With the initial linear predictor, we can obtain initial values for the pseudo-response and the pseudo-weight, z_i and w_i, from Equation (22). We set the starting value of φ (if present) to 1, and the variances τ_j^2 to 1 for all the predictors except for the intercept for which τ_0^2 is prefixed at a large value (say 10^{10}). This initial value of τ_j^2 corresponds to first setting initial values for the hyperparameters b_k to be 0.5 or 0.125 for the hierarchical t or double-exponential priors, respectively, and then taking the prior means for the scale parameters s_j and the variances τ_j^2. These hyperparameter values lead to a weakly informative prior on the coefficients, and could be reasonable. With the initial values of the pseudo-response and the pseudo-weight z_i and w_i and the variances τ_j^2, the estimate of β is obtained by the augmented weighted normal linear regression (23). From these starting values, the EM-IWLS algorithm can converge rapidly.

We apply the criterion in the R function glm to assess convergence, i.e., $|d^{(t)} - d^{(t-1)}|/(0.1 + |d^{(t)}|) < \varepsilon$, where $d^{(t)} = -2 \log p(y | X\beta^{(t)}, \varphi^{(t)})$ is the estimate of deviance at the t^{th} iteration), and ε is a small value (say 10^{-5}). At convergence of the algorithm, we obtain the latest estimates $(\widehat{\beta}, \widehat{\phi})$ and the covariance matrix $\text{Var}(\widehat{\beta}) = \left(X_*' \widehat{\Sigma}_*^{-1} X_*\right)^{-1} \widehat{\phi}$. As in the classical framework, the p-values for testing the hypotheses H_0: $\beta_j = 0$ can be calculated using the statistics $\widehat{\beta}_j / \sqrt{\text{Var}(\widehat{\beta}_j)}$, which approximately follows a standard normal distribution or a Student-t distribution with n degrees of freedom, if the dispersion φ is fixed in the model or estimated from the data, respectively.

4. Relationship with Existing Methods

The Bayesian GLMs with the hierarchical t and double-exponential priors includes various previous methods in the literature as special cases. The EM-IWLS algorithm described above can be easily adapted to fit these existing models.

1. The hierarchical t prior with $s_j = \infty$ and the double-exponential prior with $s_j = 0$ correspond to a flat distribution. Placing flat priors on all β_j corresponds to classical models, which are usually non-identifiable for high-dimensional data. Our framework has the flexibility of setting flat priors to some predictors (e.g., relevant covariates) that perform no shrinkage;

2. At $\nu_j = \infty$, the t prior is equivalent to a normal distribution $\beta_j \sim N\left(0, s_j^2\right)$, which leads to a ridge regression when setting a common scale $s_j \equiv s$;
3. For the hierarchical t prior, setting $\nu_j = s_j = 0$ corresponds to placing Jeffreys' prior on each variance, $p\left(\tau_j^2\right) \propto 1/\tau_j$, which is equivalent to a flat prior on log τ_j^2, leading to improper priors $p\left(\beta_j\right) \propto |\beta_j|^{-1}$. For the hierarchical double-exponential prior, setting $a = b_k = 0$ in the Gamma hyper-prior distribution (8) leads to the Normal-Jeffrey's prior on β_j (Armagan et al., 2010);
4. Setting a common scale parameter to all variables, $s_j \equiv s$, leads to the Bayesian lasso or t models discussed in Park and Casella (2008) and Yi and Xu (2008);
5. Ignoring the group structure but using variable-specific scale parameters s_j leads to the Bayesian adaptive lasso described in Leng et al. (2010).

6. Implementation

We have created an R function bglm for setting up and fitting the Bayesian hierarchical generalized linear models. As described above, the Bayesian hierarchical generalized linear models include various models as special cases, and thus the R function bglm can be used not only for general data analysis but also for high-dimensional data analysis using various prior distributions. Our computational strategy is based on extending the well-developed IWLS algorithm for fitting classical GLMs to our Bayesian hierarchical GLMs. The IWLS algorithm is executed in the glm function in R (http://www.r-project.org/). The bglm function implements the EM-IWLS algorithm by inserting the E-step for updating the missing values (i.e., the variances τ_j^2 and the hyperparameters s_j and $b_{k[j]}$) and the steps for calculating the augmented data and the dispersion parameter into the IWLS procedure in the glm function, and includes all the glm arguments and also some new arguments for the hierarchical modeling. We have incorporated the bglm function into the freely available R package BhGLM (http://www.ssg.uab.edu/bhglm/).

7. Applications

We illustrate the methods by analyzing the mouse data of (Boyartchuk et al., 2001). This dataset consisted of 116 female mice from an intercross (F_2) between the BALB/cByJ and C57BL/6ByJ strains. Each mouse was infected with *Listeria monocytogenes*. Approximately 30% of the mice recovered from the infection and survived to the end of the experiment (264 hours). We denoted the survival status for the i-th animal by y_i (= 0 or 1 if the i-th animal was dead or alive, respectively). The mice were genotyped at 133 genetic markers spanning 20 chromosomes, including two at the X chromosome. The numbers of markers on an autosome range from 4 to 13. The goal of the study was to identify markers that are significantly associated with the survival status and to estimate the genetic effects of these markers. The single-marker analysis has been previously applied to this data set, identifying significant QTL on chromosomes 5 and 13 (Boyartchuk et al., 2001). As shown below, our hierarchical model analyses detected additional significant QTL.

For each autosomal marker which consists of three genotypes, we constructed two main-effect variables, the additive and the dominance, using the Cockerham genetic model; the additive predictor is defined as $x_a = -1 \times \Pr(aa) + 0 \times \Pr(Aa) + 1 \times \Pr(AA)$ and the dominance predictor as $x_d = -0.5 \times [\Pr(aa) + \Pr(AA)] + 0.5 \times \Pr(Aa)$, where Pr(aa), Pr(AA) and Pr(Aa) are probabilities of homozygotes aa, AA and heterozygote Aa, respectively. For observed genotypes, one of these probabilities equals 1. The resulting 262 main-effect variables were clustered into 38 groups based on their located chromosomes and effect types

(i.e., additive or dominance). Each marker at the X chromosome consists of two genotypes. Therefore, we defined a binary variable for each of these markers and treated these variables as ungrouped. The genotype data contains ~ 11% missing values. We calculated the genotypic probabilities of missing marker genotypes conditioning on the observed marker data, and then used these conditional probabilities to construct additive and dominance predictors (Yi and Banerjee, 2009).

We used logistic models with the proposed prior distributions to simultaneously fit all the variables. Figure 1 displays the coefficient estimates, standard errors, and *p*-values for all the variables. The two analyses obtained fairly similar results, identifying four or five effects significantly associated with the survival status and shrinking all other effects to zero. Although the two models detected different additive predictors on Chromosome 13 (i.e., c13.18.9a or c13.26.2a), these two variables were strongly correlated ($r^2 = 0.87$). The model with the hierarchical double-exponential prior detected an additional dominance predictor located on Chromosome 15.

Figure 2 shows the estimates of hyper-parameters s_j and b_k for the grouped variables. For the hierarchical Cauchy prior, larger s_j and smaller b_k would induce weaker shrinkage on the corresponding coefficients. As can be seen in Figure 2, the estimates of s_j for the detected variables (i.e., c5.25.5a, c6.18.2a and c13.18.9a) were larger than those of other variables, and the estimates of b_5, b_6 and b_{13} were much smaller than other b_k's. For the hierarchical double-exponential priors, smaller s_j and larger b_k are expected to yield weaker shrinkage on the corresponding coefficients. Figure 2 clearly shows that the detected variables had much smaller estimates of s_j, and the groups including significant effects had much larger estimates of b_k. Therefore, allowing variable-specific parameters s_j and group-specific parameters b_k enable us to achieve hierarchically adaptive shrinkage on the coefficients.

For comparison purposes, we analyzed the data using the prior distributions with fixed values of hyper-parameters. Figure 3 displays the results from the logistic models with the prior distributions $\tau_j^2 \sim \text{Inv}-\chi^2\left(\nu, s^2\right)$ and $\tau_j^2 \sim \text{Expon}\left(s^2/2\right)$ with several fixed values of s^2. These prior distributions lead to the existing models described earlier. All these analyses were clearly unsatisfactory, failing to detect the strong signals that were found previously. Figure 4 shows the analyses using the prior distributions with fixed values of b_k but unknown variable-specific parameters s_j^2. Since hyper-parameters that are deeper in the hierarchy have less effect on inference (Leng et al., 2010), some of these analyses could yield results similar to the previous findings. However, the choice of the parameters b_k largely affect the results.

8. Simulation Studies

We used simulations to validate the proposed models and algorithm and to study the properties of the method. We compared the proposed method with several alternative models. Our simulation studies used the real genotype data in the above mouse survival study of *Listeria monocytogenes*, and generated a binary response y_i for each of 116 mice from the binomial distribution Bin(1, $\text{logit}^{-1}(X_i\beta^{\text{true}})$) conditional on the assumed 'true' coefficients β^{true}, where X_i was constructed as in the above real data analysis. The 'true' coefficients β^{true} were set based on the fitted logistic model with the hierarchical Cauchy prior (see the left panel of Figure 1), equaling the estimated values for the intercept and the detected predictors, c5.25.5a, c6.18.2a and c13.18.9a, and 0 for the others. For each situation, 1000 replicated datasets were simulated. We calculated the frequency of each effect estimated as significant at the threshold level of 0.05 over 1000 replicates. These frequencies corresponded to the empirical power for the simulated non-zero effects and the

type I error rate for other coefficients, respectively. We also examined the accuracy of estimated coefficients by calculating the mean and 95% interval estimates.

For each simulated dataset, we simultaneously fitted all the 264 main-effect variables using logistic linear models with the proposed prior distributions. As in the real data analysis, the variables were clustered into 38 groups. As shown in the left panel of Figure 5, the three non-zero effects were detected with higher power than all other effects. Given the small sample size and the relatively large number of predictors, these powers are reasonable. The model with the hierarchical double-exponential prior generated higher power for two of the three simulated effects. The type I error rates for effects on chromosomes without simulated non-zero effects were close to zero. For chromosomes 5, 6 and 13, however, there were several non-simulated effects detected with non-zero type I error rates. This may be expected because these variables were highly correlated with the non-zero variables. The right panel of Figure 5 shows the assumed values, the estimated means and the 95% intervals of all coefficients for the analysis with the hierarchical double-exponential prior. The estimates of all effects were accurate; the estimated means overlapped the simulated values.

We then analyzed the simulated datasets using logistic regressions with the prior distributions $\tau_j^2 \sim \text{Inv}-\chi^2\left(\nu, s^2\right)$ and $\tau_j^2 \sim \text{Expon}\left(s^2/2\right)$ with several fixed values of s_j. As shown in Figure 6, these analyses generated lower power than the proposed hierarchical prior distributions for all the simulated effects. With an inappropriate choice of the hyper-parameter, these alternative priors had no power to detect the simulated effects.

9. Discussion

We have proposed two prior distributions, hierarchical Cauchy and double exponential distributions, for simultaneous variable selection and coefficient estimation in high-dimensional generalized linear models. The hierarchical priors include both variable-specific and group-specific tuning parameters. Numerical results showed that our methods can impose different shrinkage for different coefficients and different groups, and hence allow reliable estimates of parameters and increase the power for detection of important variables. Although both the two proposed priors perform well, we found that with the double exponential distribution the power for detection of important variables is usually higher, and the algorithm converges more rapidly. Therefore, we recommend the hierarchical double exponential model as default in high-dimensional data analysis. The proposed algorithm extends the standard procedure for fitting classical generalized linear models in the general statistical package R to our Bayesian models, leading to the development of stable and flexible software. Although a fully Bayesian computation that explores the posterior distribution of parameters provides more information, our mode-finding algorithm quickly produces all results as in routine statistical analysis.

The key to Bayesian hierarchical modeling is to express shrinkage prior distributions as scale mixtures of normals with unknown variable-specific variances τ_j^2 (Kyung et al., 2010; Park and Casella, 2008; Yi and Xu, 2008). We have used this hierarchical formulation to obtain our adaptive shrinkage priors and to develop our algorithms. Kyung et al. (2010) and Leng et al. (2010) have showed that various penalized likelihood methods, including the elastic net (Zou and Hastie, 2005), the group Lasso (Yuan and Lin, 2006), and the composite absolute penalty (Zhao, Rocha and Yu, 2009), can be expressed as Bayesian hierarchical models by assigning certain priors on these variances. Our hierarchical priors can be incorporated into various Lasso methods, leading to new Bayesian hierarchical models.

Our computational algorithms also take advantage of the hierarchical formulation of the prior distributions. Given the variances τ_j^2, the normal priors on the coefficients can be included in the model as additional 'data points', and thus the coefficients can be estimated using the standard iterative weighted least squares, regardless of the specific prior distributions on the variances. The conditional expectations of the variances and other hyperparameters are independent of response data, and thus the same updating scheme can be used to update the variances and hyperparameters regardless of the response distribution. Therefore, our approach can be straightforwardly applied to a broad class of models. An alternative approach does not require the introduction of these variances (Hans, 2009; Sun et al., 2010). However, it has the disadvantage that the step for updating the coefficients is complicated. As a result, the approach is less extendable.

We describe our algorithm by simultaneously estimating all coefficients β. This can be referred to as the *all-at-once* algorithm. This method can be very fast when the number of variables is not very large (say $J < 1000$) and has the advantage of accommodating the correlations among all the variables. However, it can be slow or even cannot be implemented when the number of variables is large (say $J > 2000$) due to memory storage and convergence problems. We can extend the algorithm to update all coefficients at a group given all the others, referred to as the *group-at-time* algorithm. At each of the iteration, the group-at-time algorithm proceeds by cycling through all the groups of parameters and treats the linear predictor of all other groups as an *offset* in the model. This method updates coefficients in a conditional manner, significantly reducing the number of parameters in each M-step, and thus can deal with large number of variables.

Acknowledgments

We are grateful for Dr. Jun Liu for his suggestions and comments. This work was supported in part by the research grants: NIH 5R01GM069430-07 and R01CA142774.

References

Armagan A, Dunson D, Lee J. Bayesian generalized double Pareto shrinkage. Biometrika. 2010

Boyartchuk VL, Broman KW, Mosher RE, D'Orazio SE, Starnbach MN, Dietrich WF. Multigenic control of Listeria monocytogenes usceptibility in mice. Nat Genet. 2001; 27:259–260. [PubMed: 11242105]

Carvalho C, Polson N, Scott J. Handling sparsity via the horseshoe. JMLR: W&CP 5. 2009

Carvalho C, Polson N, Scott J. The horseshoe estimator for sparse signals. Biometrika. 2010; 97:465–480.

Figueiredo MAT. Adaptive Sparseness for Supervised Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2003; 25:1150–1159.

Gelman A. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis. 2006; 1:515–533.

Gelman, A.; Carlin, J.; Stern, H.; Rubin, D. Bayesian data analysis. Chapman and Hall; London: 2003.

Gelman A, Jakulin A, Pittau MG, Su YS. A weakly informative default prior distribution for logistic and other regression models. Annals of Applied Statistics. 2008; 2:1360–1383.

Griffin, JE.; Brown, PJ. Technical report. IMSAS, University of Kent; 2010. Bayesian adaptive lassos with non-convex penalization.

Hans C. Bayesian lasso regression. Biometrika. 2009; 96:835–845.

Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning; Data Mining, Inference and Prediction. Springer-Verlag; NewYork: 2001.

Kyung M, Gill J, Ghosh M, Casella G. Penalized Regression, Standard Errors, and Bayesian Lassos. Bayesian Analysis. 2010; 5:369–412.

Lee A, Caron F, Doucet A, Holmes C. Bayesian Sparsity-Path-Analysis of Genetic Association Signal using Generalized t Priors. Statistical Applications in Genetics and Molecular Biology. 2012; 11:1544–6115.

Leng, C.; Minh Ngoc, Tran; Nott, D. Bayesian Adaptive Lasso. 2010. Arxiv preprint arXiv:1009.2300

McCullagh, P.; Nelder, JA. Generalized linear models. Chapman and Hall; London: 1989.

Park T, Casella G. The Bayesian Lasso. Journal of the American Statistical Association. 2008; 103:681–686.

Sun W, Ibrahim J, Zou F. Genomewide multiple-loci mapping in experimental crosses by iterative adaptive penalized regression. Genetics. 2010; 185:349–359. [PubMed: 20157003]

Tibshirani R. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society. Series B. 1996a; 58:267–288.

Tibshirani R. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society. Series B. 1996b:267–288.

Yi N, Banerjee S. Hierarchical generalized linear models for multiple quantitative trait locus mapping. Genetics. 2009; 181:1101–1113. [PubMed: 19139143]

Yi N, Kaklamani VG, Pasche B. Bayesian analysis of genetic interactions in case-control studies, with application to adiponectin genes and colorectal cancer risk. Ann Hum Genet. 2011; 75:90–104. [PubMed: 20846215]

Yi N, Xu S. Bayesian LASSO for quantitative trait loci mapping. Genetics. 2008; 179:1045–1055. [PubMed: 18505874]

Yuan M, Lin Y. Model Selection and Estimation in Regression with Grouped Variables. Journal of the Royal Statistical Society, Series B. 2006; 68:49–67.

Zhao P, Rocha G, Yu B. The composite absolute penalties family for grouped and hierarchical variable selection. Ann. Statist. 2009; 37:3468–3497.

Zou H. The Adaptive Lasso and Its Oracle Properties. Journal of the American Statistical Association. 2006; 101:1418–1429.

Zou H, Hastie T. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B. 2005; 67:301–320.

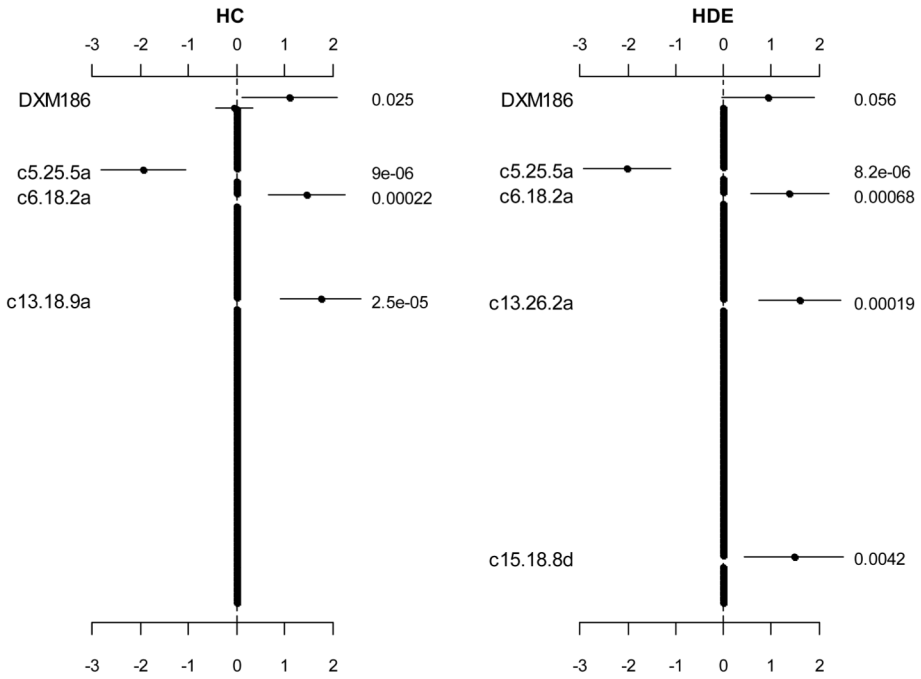

Figure 1.

Jointly fitting two effects of markers (DXM186 and DXM64) on the X chromosome and 262 main effects of 131 markers on 19 autosomes using the hierarchical logistic models with the two prior distributions, hierarchical Cauchy (HC) and hierarchical double-exponential (HDE). The points, short lines and numbers at the right side represent estimates of effects, ± 2 standard errors, and *p*-values, respectively. Only effects with *p*-value < 0.06 are labeled. The notation, e.g., c5.25.5a, indicates the additive predictor of the marker located at 25.5 cM on chromosome 5.

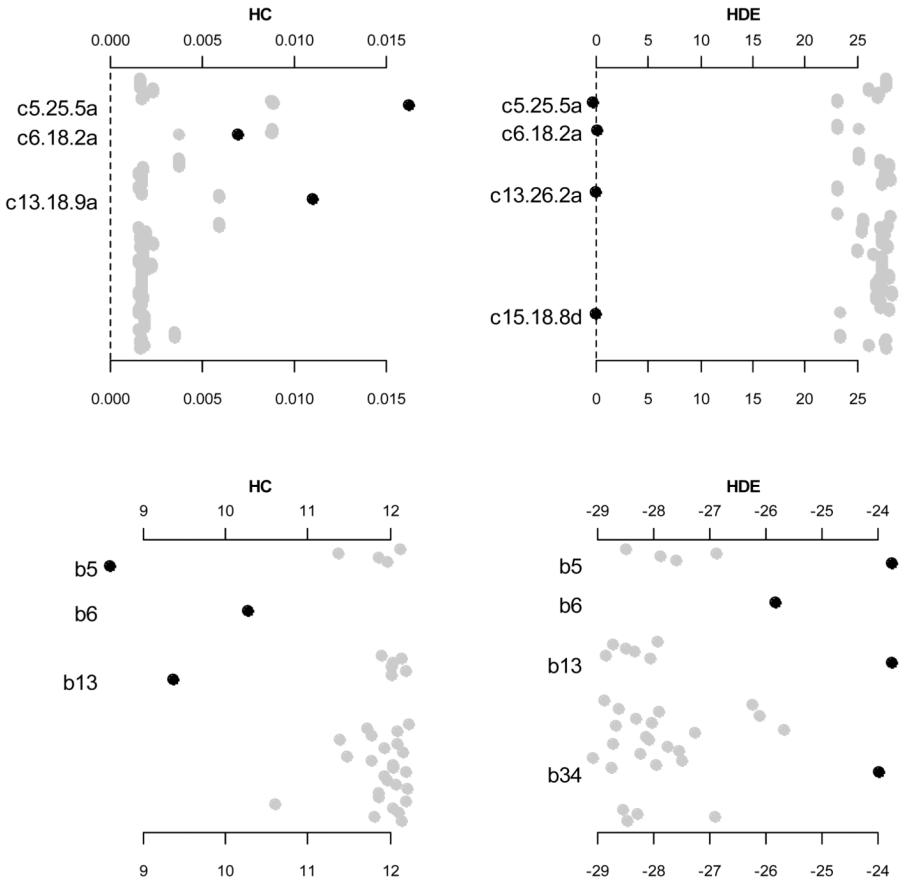

Figure 2.

Estimates of hyper-parameters of grouped variables for the two prior distributions, hierarchical Cauchy (HC) and hierarchical double-exponential (HDE). The top panels are the estimates of s_j and $\log(s_j)$, and the bottom panels are the estimates of $\log(b_k)$, for hierarchical Cauchy and hierarchical double-exponential priors, respectively. Only terms corresponding to effects with p-value < 0.05 are labeled and blacked. The notation, e.g., c5.25.5a, indicates the additive predictor of the marker located at 25.5 cM on chromosome 5.

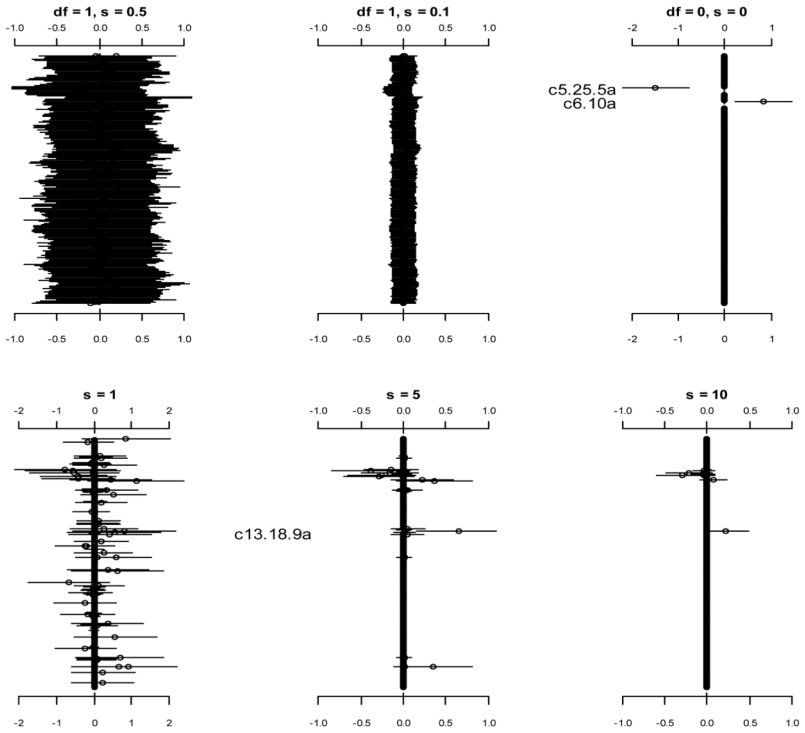

Figure 3.

Analyses using the prior distributions $\tau_j^2 \sim \text{Inv} - \chi^2\left(\nu, s^2\right)$ (the top panels) and $\tau_j^2 \sim \text{Expon}\left(s^2/2\right)$ (the bottom panels) with fixed values of s^2. The points and short lines represent estimates of effects and ± 2 standard errors, respectively. Only effects with p-value < 0.05 are labeled.

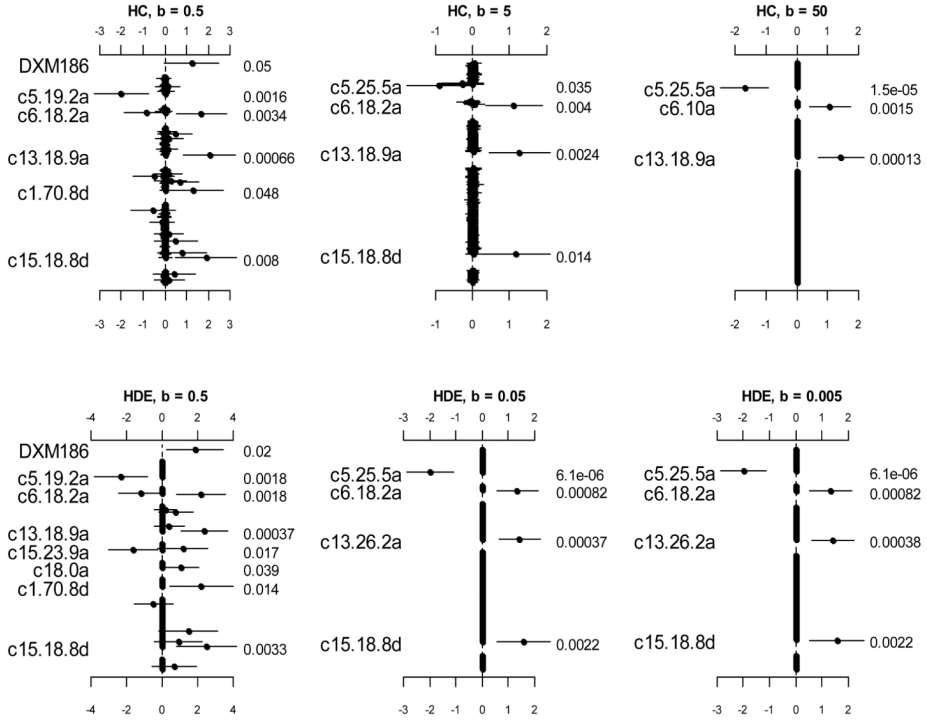

Figure 4.

Analyses using the hierarchical Cauchy (HC) and hierarchical double-exponential (HDE) with fixed values group-specific parameters b_k but unknown variable-specific parameters s_j^2. The points, short lines and numbers at the right side represent estimates of effects, ± 2 standard errors, and p-values, respectively. Only effects with p-value < 0.05 are labeled.

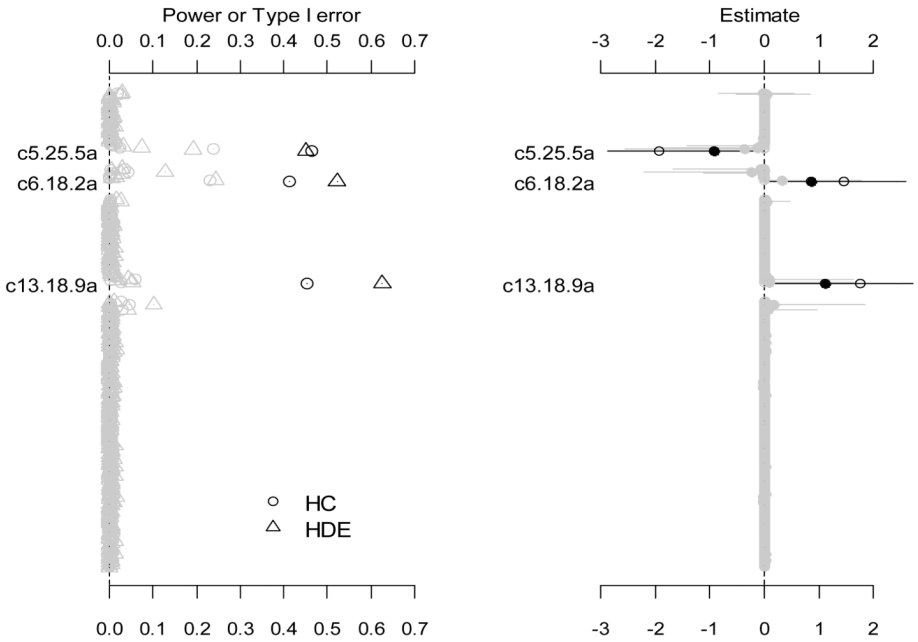

Figure 5.
The left panel shows the frequency of each effect estimated with p-value smaller than 0.05 over 1000 replicates with the two prior distributions, hierarchical Cauchy (HC) and hierarchical double exponential (HDE). The right panel shows the assumed values (circles), the estimated means (points) and the 95% intervals (short lines). Only effects with non-zero simulated value are labeled and blacked. The notation, e.g., c5.25.5a, indicates the additive predictor of the marker located at 25.5 cM on chromosome 5.

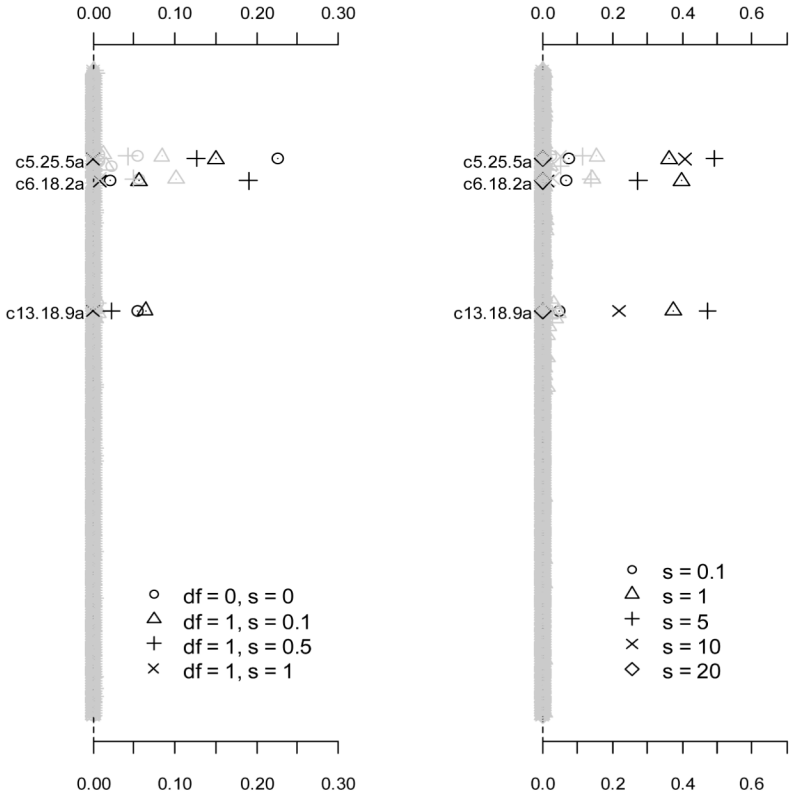

Figure 6.

Frequency of each effect estimated with p-value smaller than 0.05 over 1000 replicates using the prior distributions $\tau_j^2 \sim \text{Inv} - \chi^2\left(\nu, s_j^2\right)$ (the left panel) and $\tau_j^2 \sim \text{Expon}\left(s_j^2/2\right)$ (the right panel) with fixed values of s_j. Only effects with non-zero simulated value are labeled and blacked. The notation, e.g., c5.25.5a, indicates the additive predictor of the marker located at 25.5 cM on chromosome 5.