

Bayesian semi-parametric multiple shrinkage

Richard F. MacLehose¹, David B. Dunson¹

¹Biostatistics Branch, National Institute of Environmental Health Sciences,
National Institute of Health

Abstract

High dimensional and highly correlated data leading to non- or weakly-identified effects are commonplace. Maximum likelihood will typically fail in such situations and a variety of shrinkage methods have been proposed. Standard techniques, such as ridge regression or the lasso, shrink estimates toward zero, with some approaches allowing coefficients to be selected out of the model by achieving a value of zero. When substantive information is available, estimates can be shrunk to non-null values; however, such information may not be available. We propose a Bayesian semi-parametric approach that allows shrinkage to multiple locations. Coefficients are given a mixture of heavy tailed double exponential priors, with location and scale parameters assigned Dirichlet process hyperpriors to allow groups of coefficients to be shrunk toward the same, possibly non-zero, mean. Our approach favors sparse, but flexible structure, by shrinking towards a small number of random locations. The methods are illustrated using a study of genetic polymorphisms and multiple myeloma.

Key Words: Dirichlet process, Hierarchical model, Lasso, MCMC, Mixture model, Nonparametric; Regularization, Shrinkage prior

1. Introduction

Researchers in many disciplines routinely collect data for high-dimensional, and potentially highly-correlated, predictors. In studies relating genetic polymorphisms to phenotypic outcomes, advances in genotyping technologies have made large dimensional data commonplace, with the number of predictors typically exceeding the number of observations. Inexpensive microarray chips with the ability to genotype even larger numbers of single nucleotide polymorphisms (SNPs) will make this problem much more severe in the near future (Thomas et al., 2005). Often, the effects of the predictors will not be estimable without the incorporation of prior information. The focus of this article is on the use of shrinkage to address this problem.

It has long been known that shrinkage, or regularization, can improve estimation performance, reducing mean square error while introducing bias (Hoerl and Kennard, 1970b). This is true even in low dimensions, though the impact is particularly apparent in higher dimensional models. Shrinkage estimators typically have a Bayesian interpretation, with different estimators corresponding to different priors. Ridge regression (Hoerl and Kennard, 1970a,b) is obtained using independent normal priors centered at zero, with the degree of shrinkage controlled by the prior variance. Replacing the normal prior with a double exponential (Laplace) distribution centered at zero results in the lasso procedure of Tibshirani (1996). The double exponential prior concentrates more of its mass near zero, but also has heavier tails. This favors a sparse structure, with many of the coefficients having values close to zero and few with large values. In addition, maximum a posteriori (MAP) estimates under the Laplace prior can take zero values, allowing variable selection, though posterior means or medians do not exhibit this property (Park and Casella, 2005).

There is a rich recent literature on shrinkage methods for high dimensional predictors. Griffin and Brown (2006) extend the lasso by expressing the double exponential

distribution as a scale mixture of normals, with hyperpriors used to allow data adaptive prior choice. Relevance vector machines (Tipping, 2001) are an extension of ridge regression in which hyperprior variances specific to each coefficient are estimated using type II maximum likelihood to promote sparsity. Gelman et al. (2006) proposed independent heavy tailed Cauchy priors as a default shrinkage procedure. Model selection and shrinkage have also been achieved by assuming a mixture model for the prior distribution; typically a normal prior and a point mass at zero, such as in Geweke (1996). The mixture prior allows a coefficient to have positive probability of being zero (and dropping out of the model) or else being shrunk toward the normal prior.

The common theme of these methods is shrinkage towards a single prior mean, which is most commonly chosen as zero. Individual coefficients could be shrunk toward their own individual prior means, when sufficient prior knowledge is available. Alternatively, coefficients could be assumed exchangeable within pre-specified groups, allowing coefficients in each group to be shrunk toward different means as in Witte et al. (1994) and Greenland (1992, 1994). However, these approaches assume significant prior knowledge, which will be lacking in many instances. George (1986a,b,c) proposes a minimax multiple shrinkage estimator that is a mixture of James-Stein estimators and allows coefficients to shrink towards a fixed number of locations. Previous research by Dunson et al. (2007), MacLehose et al. (2007) and Dahl and Newton (2007) attempts to cluster coefficients for different predictors after standardization in order to borrow strength. Do et al. (2005) used Dirichlet mixture models and Loennstedt and Britton (2005) used parametric hierarchical models to allow shrinkage to two fixed means in gene expression studies.

In this paper, we propose a fundamentally different approach, which uses a Bayesian semiparametric hierarchical model to allow shrinkage of coefficients toward

multiple prior means, with the locations of these means unknown. By placing a Dirichlet process (DP) prior (Ferguson, 1974) on the unknown mean and scale parameters, we induce clustering into a small number of groups with the degree of shrinkage varying across groups. In order to develop an efficient approach for posterior computation, which can be feasibly implemented even for very large numbers of predictors, we rely on a retrospective MCMC algorithm related to that proposed recently by Papaspiliopoulos and Roberts (2007).

In section 2 we introduce the proposed hierarchical structure. In section 3 we outline the MCMC algorithm. Section 4 presents simulated data results, while Section 5 implements the model in a commonly-used diabetes dataset. Section 6 applies the approach to a dataset with $p \gg n$: an analysis of SNP data on early versus late onset multiple myeloma. Section 7 contains a discussion.

2. Semi-Parametric Multiple Shrinkage Priors

2.1 Model and Prior Formulation

Suppose we collect data (y_i, \mathbf{x}_i) , $i = 1 \dots n$, where \mathbf{x}_i is a $p \times 1$ vector of predictors (with p possibly much larger than n) and y_i is a binary outcome. A standard approach is to estimate the coefficients $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ in a regression model (we focus on the most common model in epidemiologic studies, a logistic regression for a binary outcome; however, extensions to other generalized linear models are straightforward):

$$\text{Logit}\{\Pr(y_i = 1 | \mathbf{x}_i)\} = \gamma + \mathbf{x}_i' \boldsymbol{\beta}. \quad (1)$$

For large p , maximum likelihood estimates will tend to have high variance and may not be unique. However, to regularize expression 1, we could incorporate a penalty by using a lasso prior $\pi(\boldsymbol{\beta}) = \prod_{j=1}^p \text{DE}(\beta_j | 0, \tau)$. Here, DE denotes a double exponential distribution with location parameter 0 and scale parameter τ . This model is easily implemented in a Bayesian setting by recognizing that the double exponential

distribution is equivalent to a scale mixture of normals as shown in West (1987),

$$DE(\beta_j | 0, \tau) = \int_0^\infty N(\beta_j | 0, \lambda_j) \text{Exp}(\lambda_j | 2/\tau) d\lambda_j, \quad (2)$$

where $\text{Exp}(\cdot | 2/\tau)$ denotes an exponential distribution with mean $2/\tau$. If a data augmentation step is implemented for the binary outcome (Albert and Chib, 1993; O'Brien and Dunson, 2004), the conditional distributions implied by expression 2 are conjugate and a MCMC sampling algorithm is straightforward to implement, similar to that in Park and Casella (2005).

The prior distribution in expression 2 induces shrinkage toward the prior mean of zero; however, in many situations shrinkage toward non-null values will be beneficial. We extend the lasso model specification by introducing a mixture prior with separate prior location and scale parameters for each coefficient: $\pi(\boldsymbol{\beta}) = \prod_{j=1}^p DE(\beta_j | \mu_j, \tau_j)$. Because it tends to be unclear how best to choose the hyperparameters subjectively, we allow the data to play more of a role in their choice while favoring sparsity through the use of a carefully-tailored hyperprior. This is accomplished through a Dirichlet process (DP) prior, which is non-parametric and allows clustering of parameters to help reduce dimensionality. Our proposed prior structure can then be summarized as follows:

$$\begin{aligned} \beta_j &\sim N(\beta_j | \mu_j, \lambda_j) \\ \lambda_j &\sim \text{Exp}(\lambda_j | 2/\tau_j) \\ (\mu_j, \tau_j) &\sim \pi \delta(\mu_j | 0) \times \text{Gamma}(\tau_j | a_0, b_0) + (1 - \pi) D \\ \pi &\sim \text{Beta}(\pi | 1, \alpha) \\ D &\sim \text{DP}(\alpha D_0) \\ D_0 &\equiv N(\mu_j | c, d) \times \text{Gamma}(\tau_j | a_1, b_1), \end{aligned} \quad (3)$$

where $\delta(\mu_j | 0)$ indicates that the random variable μ_j has a degenerate distribution

with all its mass at 0. Thus, with probability π a coefficient is shrunk toward zero as in standard lasso estimation. With probability $1 - \pi$ the coefficient is shrunk toward non-zero mean, μ_j . The amount of shrinkage a coefficient exhibits toward its prior mean is determined by τ_j , with larger values resulting in greater shrinkage. The Gamma distribution is parameterized as $\text{Gamma}(\tau | a, b) = 1/[b^a \Gamma(a)] \tau^{a-1} \exp[-\tau/b]$ and has mean of $a \times b$. We can specify a_0 and b_0 to give support to larger values of τ_j in order to allow strong shrinkage to 0 and a_1 and b_1 to give support to smaller values of τ_j to allow less shrinkage toward non-zero prior locations.

Because the DP prior implies that D is almost surely discrete, the prior will automatically group the p coefficient-specific hyperparameter values, $\{\mu_j, \tau_j\}_{j=1}^p$, into p^* clusters, $\{\mu_j^*, \tau_j^*\}_{j=1}^{p^*}$, with $p^* \leq p$. One of these clusters will most likely correspond to $\mu_j = 0$, while the other clusters will not have zero means and will vary in the precision parameter and hence the degree of shrinkage. The prior on the number of clusters is controlled by α , with smaller values favoring fewer clusters. However, the data are strongly informative about the number of clusters and the cluster-specific hyperparameters, so we obtain a procedure that adaptively shrinks coefficients toward non-zero locations to an extent suggested by the available data. The clustering property of the DP prior in expression 3 can be seen more clearly when expressed in equivalent stick breaking (Sethuraman, 1994) form:

$$\begin{aligned}
\beta_j &\sim \text{N}(\beta_j | \mu_{k_j}^*, \lambda_j) \\
\lambda_j &\sim \text{Exp}(\lambda_j | 2/\tau_{k_j}^*) \\
k_j &\sim \sum_{t=1}^{\infty} \pi_t \delta(k_j | t) \\
(\mu_t^*, \tau_t^*) &\sim \begin{cases} \delta(\mu_t^* | 0) \times \text{Gamma}(\tau_t^* | a_0, b_0) & \text{if } t=1 \\ \text{N}(\mu_t^* | c, d) \times \text{Gamma}(\tau_t^* | a_1, b_1) & \text{if } t > 1 \end{cases} \quad (4)
\end{aligned}$$

An infinite number of (μ_t^*, τ_t^*) are drawn from their prior distribution and the variable k_j indexes which of these bins the j^{th} coefficient's prior parameters fall into. Coefficients whose prior parameters fall into the first bin (occurring with prior probability π_1) have a standard lasso prior that shrinks toward zero. The prior probability of falling into the t^{th} , with $t > 1$, bin which has nonzero mean is π_t . The random variable π_t is constructed through a stick-breaking process: $\pi_t = V_t \prod_{h < t} (1 - V_h)$ and $V_t \sim \text{Beta}(V_t | 1, \alpha)$. Figure 1 shows a random draw from the prior distribution for one predictor-specific coefficient.

Both the prior locations and scales of the coefficients are clustered. Thus, if a number of coefficients fall into the same bin, the extent toward which they will be shrunk toward the prior mean of that bin, will depend, in part, on how similar the coefficients are to that prior mean. Note that although an infinite number of locations and scales can be drawn, in practice this type of prior encourages sparsity (for small α) since the number of clusters increases more slowly than the number of coefficients and only the first few locations have probability that is noticeably different from zero. By choosing a relatively large value of d we can give support to a wide range of possible prior means, allowing a great deal of flexibility in the model. Thus, unlike previous methods that shrink coefficients to fixed locations, the method we propose allows shrinkage toward an unknown number of prior means, allowing the locations of those prior means to be flexible but encouraging the total number of locations to be very small relative to the number of predictors.

2.2 Default Prior Specification

Prior specification is an important aspect of any Bayesian model. In many cases, substantive information may be used to inform hyperparameter values. However, we recommend standard priors that can be implemented to result in a relatively default regression technique (cf. Gelman et al. (2006)) that can be run in a wide variety of

situations, including those with little prior information. Clustering coefficients for predictors having different scales is unappealing, so we suggest standardizing predictors. However, in most biomedical applications with very many predictors these predictors will tend to be collected on the same scale. For example, indicators of genotypes at different loci intrinsically have the same scale.

To specify a_0 and b_0 , the hyperpriors for Double Exponential prior with mean fixed at zero, we assume that any β_j falling within some ϵ of zero will be viewed as having no meaningful biologic effect. That is, we treat the double exponential prior distribution with mean fixed at zero as a null cluster and coefficients assigned to that cluster should have values sufficiently close to zero to be treated as a null result. With this in mind, we choose a_0 and b_0 such that $\int_{-\epsilon}^{\epsilon} DE(\beta_j | 0, \tau_j) = z$ where z is the prior probability that a coefficient chosen randomly from the null bin has a null effect. For instance, if $z = 0.95$ and $\epsilon = 0.1$, then values of $a_0 = b_0 = 30$ guarantee that 95% of coefficients drawn from this bin will have an effect that is viewed as indistinguishable from the null. Values of a_1 and b_1 need to be specified for the priors on the scale parameter for the non-null bins. We recommend choosing smaller values for these hyperparameters that are large enough to encourage shrinkage but not so large as to overwhelm the data and arbitrarily force a huge number of bins to be generated. We specify default priors for the Gamma distribution as $a_1 = b_1 = 6.5$, so the DE prior has prior credible intervals of unit width. This is a robust choice allowing the data to inform about the amount of shrinkage. We set $\alpha = 1$, a common default in DP models.

We suggest setting the parameters c and d , the prior mean and variance for the location parameter portion of the base distribution for the DP, so that the μ_p have support over a wide range of values. Setting $c = 0$, we choose $d = 4$ such that we assign 95% probability to a very wide range of reasonable prior effects. In some

instances, particularly in epidemiology, prior knowledge may indicate that an effect larger than some value U is very unlikely. In this instance we recommend setting $d = (U/1.96)^2$. However, unless prior knowledge suggests otherwise, we recommend against setting d too large as it may impede convergence by proposing prior locations at unlikely locations.

2.3 Multiple Testing

In many applied settings, a decision will need to be made about the effect of some variable. In genetic settings, researchers will typically need to decide whether certain SNPs require further study. We outline a general procedure for flagging predictors using the output of the MCMC algorithm. Müller et al. (2004) consider a variety of loss functions and demonstrate the optimal approach is to flag an effect when its posterior probability exceeds some threshold. Let $H_j = 0$ indicate that the null hypothesis is true for the j^{th} predictor and $H_j = 1$, otherwise. We estimate $\pi_j = Pr(H_j = 1|\text{data})$ from our model by assuming that $|\beta_j| < \epsilon$ is null, for scientific purposes. Then, $\pi_j = \sum_{g=1}^G I(|\beta_j^g| > \epsilon)/G$ where G is the number of MCMC iterations, β_j^g is the g^{th} MCMC iterate for the j^{th} coefficient and I is an indicator function. We can estimate the posterior expected false discovery rate (FDR) for a threshold, c , as less than or equal to $\sum_j (1 - \pi_j) I(\pi_j > c) / \sum_j I(\pi_j > c)$. Typically, the threshold will be chosen as the smallest value such that the FDR is less than or equal to some meaningful value such as 20% or 50%. An alternative to this approach is to simply list the predictors ordered by their posterior probabilities of $H_j = 1$. In genomics applications, this allows investigators to identify promising genes for further study, which is a common goal in analyzing high-throughput data.

3: Posterior Computation

Our approach initially augments the data using the O'Brien and Dunson (2004) al-

gorithm by assuming the outcome $y_i = 1$ occurs when a latent variable, $g_i > 0$. We assume $g_i = \mathbf{X}'_i \boldsymbol{\beta} + \epsilon_i / \phi_i$, where $\epsilon_i \sim N(0, \sigma^2)$ and $\phi_i \sim \text{gamma}(\nu/2, 2/\nu)$. This scale mixture of normals with $\sigma^2 = \pi^2(\nu - 2)/3\nu$ and $\nu = 7.3$ is a near exact representation of the logistic distribution. The data augmentation approach allows this algorithm to be easily modified for other regression models, such as probit or linear. We propose a Metropolis within Gibbs sampling algorithm that proceeds through the following steps:

- 1a. Augment the data with $\mathbf{g} = (g_1 \dots g_n)'$ sampled from $f(g_i | y_i, \boldsymbol{\beta}, \phi_i)$

$$\begin{aligned} f(g_i | \mathbf{y}_i = 1, \boldsymbol{\beta}, \phi_i) &= N^+(g_i | \mathbf{x}'_i \boldsymbol{\beta}, \sigma^2 / \phi_i) \\ f(g_i | \mathbf{y}_i = 0, \boldsymbol{\beta}, \phi_i) &= N^-(g_i | \mathbf{x}'_i \boldsymbol{\beta}, \sigma^2 / \phi_i) \end{aligned} \quad (5)$$

where N^+ is a normal distribution truncated to the left of 0 and N^- is truncated to the right of 0.

- 1b. Update ϕ_i by sampling from $f(\phi_i | \boldsymbol{\beta}, g_i)$.

$$f(\phi_i | \boldsymbol{\beta}, g_i) = \text{Gamma}\left(\phi_i \mid \frac{\nu + 1}{2}, \frac{2}{\nu + (g_i - \mathbf{x}_i \boldsymbol{\beta})^2 / \sigma^2}\right) \quad (6)$$

2. Use the current estimates of $\boldsymbol{\mu} = (\mu_{k_1}^* \dots \mu_{k_p}^*)'$ to update the regression coefficients. Assume the intercept, γ has prior distribution $N(\gamma | \gamma_0, \lambda_0)$ with γ_0 and λ_0 fixed hyperpriors. Let $\boldsymbol{\mu}_0 = (\gamma_0, \boldsymbol{\mu}^{*'})'$, $\boldsymbol{\beta}_0 = (\gamma, \boldsymbol{\beta}')'$ and \mathbf{X} be the $n \times (p + 1)$ design matrix with first column equal to 1. Then we update by sampling from

$$f(\boldsymbol{\beta}_0 | \Gamma, \Lambda, \boldsymbol{\mu}_0, \mathbf{g}) = N(\boldsymbol{\beta}_0 | E_{\boldsymbol{\beta}}, V_{\boldsymbol{\beta}}) \quad (7)$$

where $V_{\boldsymbol{\beta}} = (\mathbf{X}'\Gamma^{-1}\mathbf{X} + \Lambda^{-1})^{-1}$ and $E_{\boldsymbol{\beta}} = V_{\boldsymbol{\beta}}(\mathbf{X}'\Gamma^{-1}\mathbf{g} + \Lambda^{-1}\boldsymbol{\mu}_0)$. The matrix Λ is a $(p + 1) \times (p + 1)$ diagonal matrix with j^{th} element λ_{j-1} , and Γ is an $n \times n$ diagonal matrix with i^{th} element σ^2 / ϕ_i .

3. Update the mixing parameter, λ_j .

$$\begin{aligned}
 f(\lambda_j | \beta_j, \mu_{k_j}^*, \tau_{k_j}^*) &\propto (\lambda_j)^{-1/2} \exp \left[-\frac{1}{2} \left(\frac{(\beta_j - \mu_{k_j}^*)^2}{\lambda_j} + \lambda_j \tau_{k_j}^* \right) \right] \\
 &\propto \frac{1}{\text{iG}(\lambda_j | \sqrt{\frac{\tau_{k_j}^*}{(\beta_j - \mu_{k_j}^*)^2}}, \tau_{k_j}^*)}
 \end{aligned} \tag{8}$$

where iG is the inverse Gaussian distribution, $\text{iG}(y | a, b) = \sqrt{\frac{b}{2\pi}} y^{-3/2} \exp[-\frac{b}{2y}(\frac{y}{a} - 1)^2]$ with mean a and variance a^3/b . Samples can be drawn from an inverse Gaussian distribution as in Chhikara and Folks (1989).

4. We then update the prior location and scale parameters using a modified version of the retrospective stick breaking algorithm proposed by Papaspiliopoulos and Roberts (2007). The stick breaking form of the DP in expression 4 cannot be updated directly through a MCMC sampler because of the infinite dimension classification variable. Instead, the retrospective algorithm recognizes that the p prior location and scale parameters will fall in the first p^* bins (with $p^* \leq p$ and typically $p^* \ll p$) thus only the probability of falling into those bins needs to be updated. The variable p^* is allowed to grow or shrink as needed in the algorithm.

4a. Sample $(\mu_1^*, \tau_1^*) \dots (\mu_{p^*}^*, \tau_{p^*}^*)$ from their conditional posterior distribution. Set $p^* = \max(\mathbf{K})$ where $\mathbf{K} = (k_1 \dots k_p)$ and defines which of the p^* bins each of the p priors locations and scales are currently assigned. For instance, $k_1 = 2$ would indicate that the prior location and scale parameters for the first coefficient are (μ_2^*, τ_2^*) . We define the number of predictors that fall in the t^{th} bin as m_t . Begin by updating μ_t . Since the prior for μ_1 has unit probability mass at 0, the posterior distribution is $f(\mu_1^* | \tau_1^*, \boldsymbol{\beta}) = \delta(\mu_1 | 0)$. The conditional posterior for the scale parameter in the first

bin is given by:

$$\begin{aligned}
f(\tau_1^* | \boldsymbol{\lambda}, \mathbf{K}) &\propto \text{Gamma}(\tau_1^* | a_0, b_0) \prod_{j:k_j=1} \text{Exp}(\lambda_j | \frac{2}{\tau_1^*}) \\
&\propto \text{Gamma}\left(\tau_1^* | m_1 + a_0, \frac{1}{\sum_{j:k_j=1} \lambda_j/2 + 1/b_0}\right)
\end{aligned} \tag{9}$$

For subsequent bins, $t > 1$, the conditional posteriors are given by:

$$\begin{aligned}
f(\mu_t^* | \tau_t^*, \boldsymbol{\beta}, \mathbf{K}) &\propto \text{N}(\mu_t^* | c, d) \prod_{j:k_j=t} \text{N}(\beta_j | \mu_t^*, \lambda_j) \\
&\propto \text{N}(\mu_t^* | \widehat{E}_{\mu_t}, \widehat{V}_{\mu_t})
\end{aligned} \tag{10}$$

$$f(\tau_t^* | \boldsymbol{\lambda}, \mathbf{K}) \propto \text{Gamma}\left(\tau_t^* | m_t + a_1, \frac{1}{\sum_{j:k_j=t} \lambda_j/2 + 1/b_1}\right) \tag{11}$$

where $\widehat{V}_{\mu_t} = (1/d + \sum_{j:k_j=t} 1/\lambda_j)^{-1}$ and $\widehat{E}_{\mu_t} = \widehat{V}_{\mu_t} (c/d + \sum_{j:k_j=t} \beta_j/\lambda_j)$. Potentially, no coefficients will belong to one of the bins and in this case, sampling (μ_t^*, τ_t^*) amounts to sampling from the prior distribution, $\text{N}(\mu_t^* | c, d) \times \text{Gamma}(\tau_t^* | a_1, b_1)$.

4b. Sample V_t from $f(V_t | \mathbf{K}, \alpha)$ and generate π_t using the stick breaking formula above.

$$\begin{aligned}
f(V_t | \mathbf{K}, \alpha) &\sim \text{Beta}(m_t + 1, p - \sum_{l=1}^t m_l + \alpha) \\
\pi_t &= \prod_{h=1}^{t-1} (1 - \pi_h) V_t
\end{aligned} \tag{12}$$

4c. Update the vector of coefficient configurations, \mathbf{K} , using a Metropolis step. To generate a proposal configuration, we calculate the posterior probability of the j^{th} predictor falling in the l^{th} bin as:

$$q_j(l) \propto \begin{cases} \pi_l \text{N}(\beta_j | \mu_l^*, \lambda_j) \text{Exp}(\lambda_j | \frac{2}{\tau_l^*}) & \text{for } l \leq \max(\mathbf{K}) \\ \pi_l M_j(\mathbf{K}) & \text{for } l > \max(\mathbf{K}) \end{cases} \tag{13}$$

Following the recommendation of Papaspiliopoulos and Roberts (2007) we specify $M_j(p^*) = \max(\text{N}(\beta_j | \mu_l^*, \lambda_j) \text{Exp}(\lambda_j | \frac{2}{\tau_l^*}), l \leq \max(\mathbf{K}))$ and the normalizing constant for q_j is $n_j(\mathbf{K}) = \sum_{l=1}^{\max(\mathbf{K})} q_j(l) + M_j(\mathbf{K})(1 - \sum_{l=1}^{\max(\mathbf{K})} \pi_l)$.

To determine the proposal configuration for the j^{th} prior, we sample $U_j \sim \text{Uniform}(0, 1)$ and propose to move into bin b , where the probability of falling into bin k is $q_j(l)$. If $U_j > \sum_{l=1}^{max(\mathbf{K})} q_j(l)$, let $p^* = p^* + 1$ and draw new values of $(\mu_{p^*}^*, \tau_{p^*}^*)$ from their prior distributions until $U_j \leq \sum_{l=1}^{max(K)} q_j(l)$.

We now have a proposed configuration $\mathbf{K}' = (k_1, \dots, k_{j-1}, l, k_{j+1}, \dots, k_p)$ for moving the j^{th} coefficient to bin l . We accept the move from configuration K to configuration K' with probability:

$$\min \left(1, \frac{n(\mathbf{K})}{n(\mathbf{K}')} \frac{M_j(\mathbf{K}')}{N(\beta_j | \mu_{k_j}^*, \lambda_j) \text{Exp}(\lambda_j | \frac{2}{\tau_{k_j}^*})} \right) \quad \begin{array}{l} 1 \quad \text{if } l \leq \max(\mathbf{K}) \text{ and } \max(\mathbf{K}) = \max(\mathbf{K}') \\ \text{if } l \leq \max(\mathbf{K}) \text{ and } \max(\mathbf{K}') < \max(\mathbf{K}) \end{array}$$

$$\min \left(1, \frac{n(\mathbf{K})}{n(\mathbf{K}')} \frac{N(\beta_t | \mu_l^*, \lambda_j) \text{Exp}(\lambda_j | \frac{2}{\tau_l^*})}{M_j(\mathbf{K})} \right) \quad \text{if } l > \max(\mathbf{K})$$

We iterate through these steps until convergence is achieved, excluding an initial number of iterates as a burn-in period. Common algorithms for updating in DP model such as those by Escobar and West (1995) slow down considerably as the number of predictors increases. The retrospective sampling algorithm, in our experience, is more robust converging quickly even with many thousands of predictors.

4. Simulations

Allowing shrinkage of model parameters to multiple locations is intuitively appealing; however, we examine the performance of this approach, relative to the Bayesian lasso, in some simple simulations. First, we simulated data in which effects were easily estimated : 400 observations and 20 parameters, with 10 of the parameters having true effect of 2 and the remaining 10 having a true effect of 0. A second set of simulations was performed for a data sets with $p > n$: 100 observations and 200 parameters, 10 of which have true effect of 2 while the remaining have true effect 0. 50 simulated datasets were generated and we examined the performance of methods

using mean squared error (MSE), bias, false positive and true positive rates.

We implemented the multiple shrinkage model using the default specification of $\alpha = 1$, $a_0 = b_0 = 30$, $a_1 = b_1 = 6.5$, $c = 0$, and $d = 4$ for the simulations. Each MCMC algorithm was run for 100,000 iterations with the first 10,000 discarded as burn-in. Output of the algorithms was examined for the first few simulations to determine convergence. The algorithms ran quickly in Matlab v7.5 on a Dell desktop with a 2.99 GHz Xeon chip and 3Gb RAM, taking approximately 5 minutes when $p = 20$ and 45 minutes when $p = 200$.

For comparison we implemented a standard Bayesian lasso through a Gibbs sampler similar to that in Park and Casella (2005). In particular, we assume $\beta_j \sim N(0, \lambda_j^2)$ with $\lambda \sim \text{Exp}(2/\tau)$ and $\tau \sim \text{Gamma}(a, b)$. Hyperparameters a and b in the Bayesian lasso are set equal to a_1 and b_1 in expression 4.

Results from the first set of simulations indicate that the multiple shrinkage prior offers improvement over the standard Bayesian lasso. The MSE estimated over all simulated datasets are smaller for the multiple shrinkage prior. The reduction in MSE is largely a result of decreased bias. The multiple shrinkage prior model tends to identify the correct coefficient clustering. For instance, prior location and scale parameters are often grouped into one cluster for the first 10 coefficients and a second cluster for the last 10 coefficients. Each of the β coefficients is shrunk towards a cluster specific prior mean. The first 10 coefficients are shrunk toward a prior mean that is close to 2, resulting in dramatically decreased bias for $\beta_1 \dots \beta_{10}$ in the multiple shrinkage model (MSE=0.03) compared to the standard lasso model (MSE=1.08). Additionally, those coefficients ($\beta_{11} \dots \beta_{20}$) whose prior means are clustered into the null bin and assigned a prior mean of zero are shrunk more strongly toward that mean in the multiple shrinkage prior model (MSE=0.01) than in the Bayesian lasso model (MSE=0.04), as a result of our prior specification that $a_0 = b_0 > a_1 = b_1$.

Results from the second set of simulations illustrate the advantages of the multiple shrinkage priors in large dimensions. The estimated MSE of the 10 coefficients with an effect of 2 is much lower in the multiple shrinkage model than in Bayesian lasso (MSE=1.5 and 3.2, respectively). The remaining 190 coefficients are estimated with slightly higher MSE in the multiple shrinkage prior model than the Bayesian lasso (MSE=0.08 vs .01, respectively). The improved performance of the multiple shrinkage model is a result of including prior locations away from 0 and near the truth, resulting in shrinkage toward that value and decreased bias. This also results in slightly poorer performance in estimating the coefficients with null effects since they will, occasionally, be shrunk toward non-null values. We suspect that this trade-off in performance is one that investigators will gladly accept. For example, in our simulated datasets, we can choose to flag significant results using the Bayesian FDR discussed in section 2 and control the FDR at 20%. Averaging over the simulations, the Bayesian lasso flags 9% of true positives and 0.1% of true negative predictors as significant while the multiple shrinkage prior flags 46% of true positives and 1% of true negatives as significant. The multiple shrinkage prior results in much greater power and a slightly increase type-I error rate. Similar results are obtained with other values of ϵ .

5. Multiple Shrinkage Priors and Diabetes

We implement our method in the widely used Pima Indian diabetes dataset. Complete information on the outcome, diabetes, and predictors was available for 532 patients. Predictors included in the model were gravidity, plasma glucose concentration, diastolic blood pressure, skin fold thickness, body mass index (BMI), diabetes pedigree function and age. The Pima Indian data is typically split into training and testing datasets of size 200 and 332, respectively. We include variables for all 2-way

interactions and standardize all predictors to have mean 0 and variance 1.

We use the prior distribution on β_p as in expression 4 and use the default specification that we recommend in section 2. The semi-parametric multiple shrinkage model was implemented as in section 3 for 20,000 iterations. The chain converged rapidly and showed little autocorrelation. We excluded the initial 5,000 iterations as a burn in. For comparison, we implemented a Bayesian version of the lasso by assuming $\beta_j \sim N(0, \lambda_j^2)$ with $\lambda_j \sim \text{Exp}(2/\tau_j)$ and $\tau_j \sim \text{Gamma}(a, b)$. A Gibbs sampling algorithm for this model can be found in Park and Casella (2005). We ran this model twice, with hyperpriors $a = a_0, b = b_0$ and once with $a = a_1, b = b_1$, where a_1 and b_1 are the hyperprior terms from the multiple shrinkage prior model.

Posterior median and 90% credible intervals from the three models are shown in Figure 2. The Bayesian lasso model with a very concentrated prior distribution ($a = a_0, b = b_0$) shrinks all coefficients strongly toward zero while the model with a less concentrated prior ($a = a_1, b = b_1$) provided less shrinkage. The multiple shrinkage prior model retained the ability to shrink some estimates strongly toward zero while allowing other estimates to be shrunk toward non-zero locations. For instance, prior location and scale parameters for the effect of the 12th predictor is clustered with the prior for the 16th predictor 51% of the time. These coefficients were shrunk toward their group specific hyperprior mean rather than 0, as in the standard Bayesian lasso resulting in larger effects for these parameters.

The Pima Indian data has been widely used for purposes of prediction. We generated posterior predictive distributions for the outcome, \tilde{y}_i , of the new observations in the testing dataset by integrating over model parameters using the output of the MCMC algorithm. When $\Pr(\tilde{y}_i = 1) \geq 0.5$ we predict the outcome to be 1, and 0 otherwise. The multiple shrinkage prior approach was compared to the two Bayesian lassos and a support vector machine (SVM), a maximum margin classifier described in

Bishop (2006). Interestingly, the SVM had the worst predictive error in this example, with 34% of outcomes in the testing dataset misclassified. The Bayesian lasso with $a = a_0, b = b_0$ offered improvement over the SVM, misclassifying 23% of the outcomes, classifying far too many observations as not being diabetic. The Bayesian lasso with $a = a_1, b = b_1$ provided less shrinkage of effects and improved prediction somewhat, misclassifying 22% of outcomes. The multiple shrinkage prior offered further improvement, with a misclassification rate of 21%. Interestingly, the multiple shrinkage prior approach was better at flagging true diabetics than the other approaches.

6. SNPs and multiple myeloma

Multiple myeloma is a hemotologic cancer (the second most common, after non-Hodkins lymphoma) of the plasma cells with an average incidence of 5.5/100,000 person-years (Ries et al., 2007). Survival rates for individuals afflicted with myeloma are poor with 5 year survival at roughly 30% (Ries et al., 2007). Relatively few environmental causes of myeloma have been established and increasingly research has focused on genetic causes of the disease. Myeloma among individuals less than 40 years of age is relatively uncommon and there has been speculation that these individuals have a genetic predisposition for the disease. We analyze SNP data from 80 individuals diagnosed with multiple myeloma to determine whether any polymorphisms are related to early age (before 40 years) at onset. The collection of these data and a previous analysis of them using support vector machines has been described in Waddell et al. (2005).

We limit our analysis to 135 SNPs without any missing values. Indicator variables are created for each SNP with the heterozygous genotype treated as the referent category, leading to 271 coefficients to estimate (270 SNP coefficients and 1 intercept). We specified the default hyperpriors and ran our MCMC sampling algorithm

for 600,000 iterations, keeping every 10th sample and discarding the initial 10,000 iterations as a burn-in. In genetic applications, interest often focuses less on effect estimation than on significance testing. As previously, we assume that $|\beta_j| < \epsilon$ can be treated as null in substantive terms. We estimate $\pi_j = \sum_g I(|\beta_j^g| > \epsilon)/G$ as the posterior probability of the j^{th} genotype having an effect. Coefficients whose prior location and scale fall into the null bin frequently will typically have a very small posterior probability of having an effect, as shown in Figure 3. The threshold implied by choosing to guarantee a $\text{FDR} \leq 50\%$ is 0.46 and 72 genotype effects with posterior probability above that threshold are flagged as significant. If a $\text{FDR} \leq 30\%$ is specified, a threshold of 0.65 is chosen and 2 genotypes are flagged as significant. The two genotypes (896198 and 912651) flagged under this criterion are both located on the first chromosome. Previous research has linked chromosomal abnormalities in the 1p arm with poorer prognosis following diagnosis of multiple myeloma (Wu et al., 2007) and both of these SNPs fall in the 1p arm. Both SNPs fall in intergenic regions and do not fall on known genes; however, they may be in high linkage disequilibrium with a SNP in a gene that is related to multiple myelomas. SNP 896198 is found near a number of amylase coding and regulating genes (AMY1A, AMY1B and AMY1C, among others) and some research has suggested hyperamylasaemia among multiple myeloma patients (Hata et al., 2006).

We analyze these same data using two versions of the standard Bayesian lasso. As before, the first Bayesian lasso allows for greater shrinkage ($a = a_0, b = b_0$) while the second allows less shrinkage ($a = a_1, b = b_1$). MCMC sampling algorithms for each model are run for 20,000 iterations with the initial 5,000 discarded as a burn in. The Bayesian lasso with $a = a_0, b = b_0$ provided such strong shrinkage toward zero that it did not result in any predictors being flagged as warranting further investigation, under $\text{FDR} \leq 30\%$ or $\text{FDR} \leq 50\%$. The Bayesian lasso with $a = a_1, b = b_1$ flagged

one SNP(912651) as warranting further investigation under $FDR \leq 30\%$. However, under $FDR \leq 50\%$ it flagged all SNPs as significant. By allowing a mixture of prior distributions, the multiple shrinkage prior provides greater flexibility in both the amount of shrinkage and the value toward which coefficients are shrunk thus allowing for more reasonable estimation.

7. Discussion

This article has proposed a novel semi-parametric Bayesian multiple shrinkage prior that extends previous shrinkage methods by allowing coefficients to shrink to multiple, unknown locations. Previous methods either allow shrinkage only toward zero, which can greatly decrease MSE or provide a small number of non-zero fixed locations toward which estimates are shrunk. By allowing shrinkage to an unknown number of unknown locations other than zero, we demonstrate that MSE can often be reduced even further.

We have implemented this model in two drastically different settings: a low dimensional example with $p < n$, and a high dimensional example with $p \gg n$. Our simulated data sets under these two scenarios indicate generally improved performance of the multiple shrinkage model. In the first example, large improvements are seen in terms of MSE (largely as a result of decreasing bias). In the large dimensional dataset, MSE is decreased substantially for coefficients having a non-null effect and is increased slightly for coefficients with null effects. The multiple shrinkage prior model should be of great use to substantive researchers in many settings.

Acknowledgements This research was supported by the Intramural Research Program of the NIH, and NIEHS.

References

- JH Albert and S Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88:669–679, 1993.
- CM Bishop. *Pattern Recognition and Machine Learning*, chapter 7: Sparse Kernel Machines, pages 325–357. Springer, 2006.
- RS Chhikara and L Folks. *The Inverse Gaussian Distribution*. Marcel Dekker, Inc., 1989.
- DB Dahl and MA Newton. Multiple hypothesis testing by clustering treatment effects. *Journal of the American Statistical Association*, 102:517–526, 2007.
- KA Do, P Müller, and F Tang. A Bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society, Series C: Applied Statistics*, 54(3):627–644, 2005.
- DB Dunson, AH Herring, and SM Mulherin-Engel. Bayesian selection and clustering of polymorphisms in functionally related genes. *JASA*, to appear, 2007.
- MD Escobar and M West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1995.
- TS Ferguson. Prior distributions on spaces of probability measures. *The Annals of Statistics*, 2:615–29, 1974.
- A Gelman, A Jakulin, MG Pittau, and YS Su. A default prior distribution for logistic and other regression models. Technical report, Columbia University, 2006.
- EI George. Minimax multiple shrinkage estimation. *The Annals of Statistics*, 14:188–205, 1986a.

- EI George. A formal Bayes multiple shrinkage estimator. *Communications in Statistics: Theory and Methods*, 15:2099–2114, 1986b.
- EI George. Combining minimax shrinkage estimators. *Journal of the American Statistical Association*, 81:437–445, 1986c.
- J Geweke. Variable selection and model comparison in regression. In JM Bernardo, JO Berger, AP Dawid, and AFM Smith, editors, *Bayesian Statistics 5*, pages 609–620. Oxford Press, 1996.
- S. Greenland. A semi-bayes approach to the analysis of correlated multiple associations, with an application to an occupational cancer-mortality study. *Stat Med*, 11(2):219–230, Jan 1992.
- S. Greenland. Hierarchical regression for epidemiologic analyses of multiple exposures. *Environ Health Perspect*, 102 Suppl 8:33–39, Nov 1994.
- JE Griffin and PJ Brown. Alternative prior distributions for variable selection with very many more variables than observations. Technical report, University of Warwick, 2006.
- H Hata, H Matsuzaki, K Tanaka, H Nomura, T Kagimaoto, M Takeya, N Yamane, and K Takatsuki. Ectopic production of salivary-type amylase by a iga-lambda-type multiple myeloma. *Cancer*, 62:1511–1515, 2006.
- AE Hoerl and R Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970a.
- AE Hoerl and RW Kennard. Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12:69–82, 1970b.

- I Loennstedt and T Britton. Hierarchical bayes models for cdna microarray gene expression. *Biostatistics*, 6(2):279–291, Apr 2005.
- RF MacLehose, DB Dunson, AH Herring, and JA Hoppin. Bayesian methods for highly correlated exposure data. *Epidemiology*, 18(2):199–207, 2007.
- P Müller, G Parmigiani, C Robert, and J Rousseau. Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association*, 99:990–1001, 2004.
- SM O’Brien and DB Dunson. Bayesian multivariate logistic regression. *Biometrics*, 60(3):739–46, 2004.
- O Papaspiliopoulos and G.O. Roberts. Retrospective markov chain monte carlo methods for dirichlet process hierarchical models. *Biometrika-in press*, 2007.
- T Park and G Casella. The bayesian lasso. Technical report, University of Florida, 2005.
- LAG Ries, D Melbert, M Kraphco, A Mariotto, BA Miller, EJ Feuer, L Clegg, MJ Horner, N Howlader, MP Eisner, M Reichman, and BK Edwards, editors. *SEER Cancer Statistics Review, 1975-2004*. National Cancer Institute, Bethesda, MD, 2007.
- J. Sethuraman. A constructive definition of the dirichlet process prior. *Statistica Sinica*, 2:639–650, 1994.
- DC Thomas, RW Haile, and D Duggan. Recent developments in genomwide association scans: a workshop summary and review. *Am J Hum Genet*, 77(3):337–45, 2005.

- R Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58:267–288, 1996.
- ME Tipping. Sparse bayesian learning and the relevance vector machine. *The Journal of Machine Learning Research*, 1:211–244, 2001.
- M Waddell, D Page, and J Shaughnessy. Predicting cancer susceptibility from single-nucleotide polymorphism data: a case study in multiple myeloma. In *BIOKDD '05: Proceedings of the 5th international workshop on Bioinformatics*, pages 21–28, New York, NY, USA, 2005. ACM Press. ISBN 1-59593-213-5.
- M West. On scale mixtures of normal distributions. *Biometrika*, 74:646–648, 1987.
- J. S. Witte, S. Greenland, R. W. Haile, and C. L. Bird. Hierarchical regression analysis applied to a study of multiple dietary exposures and breast cancer. *Epidemiology*, 5(6):612–621, Nov 1994.
- KL Wu, B Beverloo, HM Lokhorst, CM Segeren, B van der Holt, MM Steijaert, PH Westveer, PJ Poddighe, GE Verhoef, and P Sonneveld. Abnormalities of chromosome 1p/q are highly associated with chromosome 13/13q deletions and are an adverse prognostic factor for the outcome of high-dose chemotherapy in patients with multiple myeloma. *Br J Haematol*, 136:615–23, 2007.

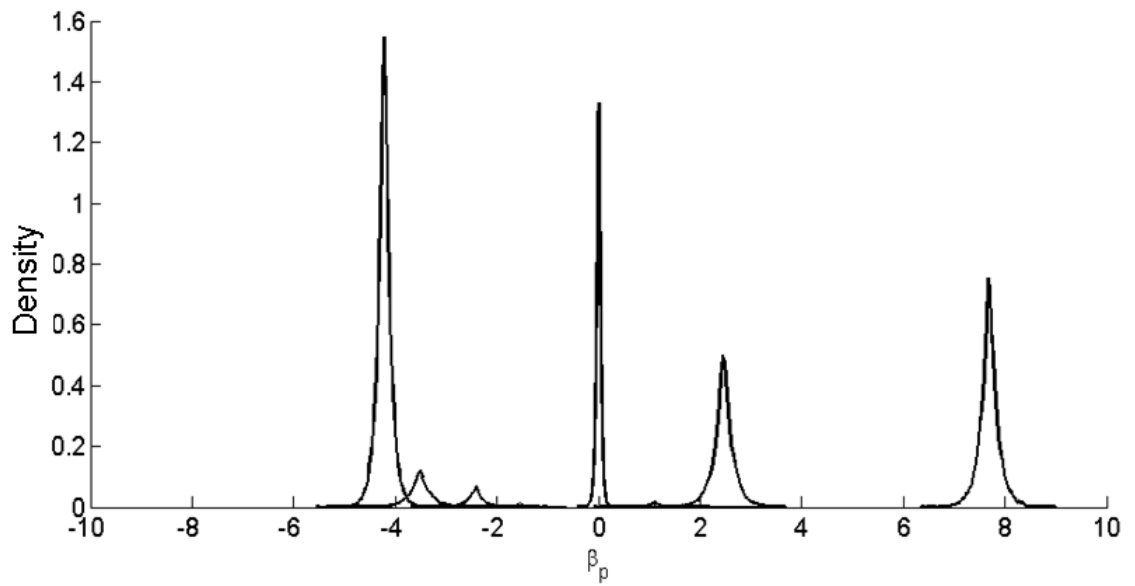


Figure 1: Multiple shrinkage prior distribution with $\alpha = 1$, $c = 0$ and $d = 10$.

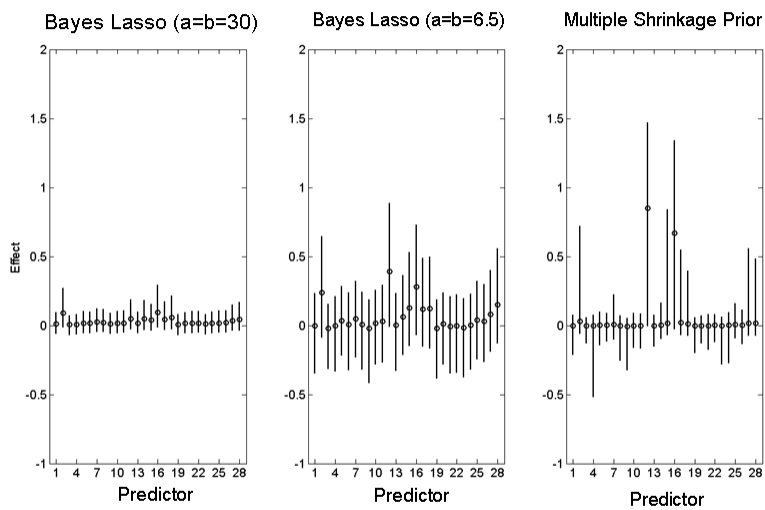


Figure 2: Posterior median and 90% credible intervals for analysis of Pima Indian Diabetes data with Bayes lasso with $a = b = 30$, Bayes lasso with $a = b = 6.5$ and multiple shrinkage prior.

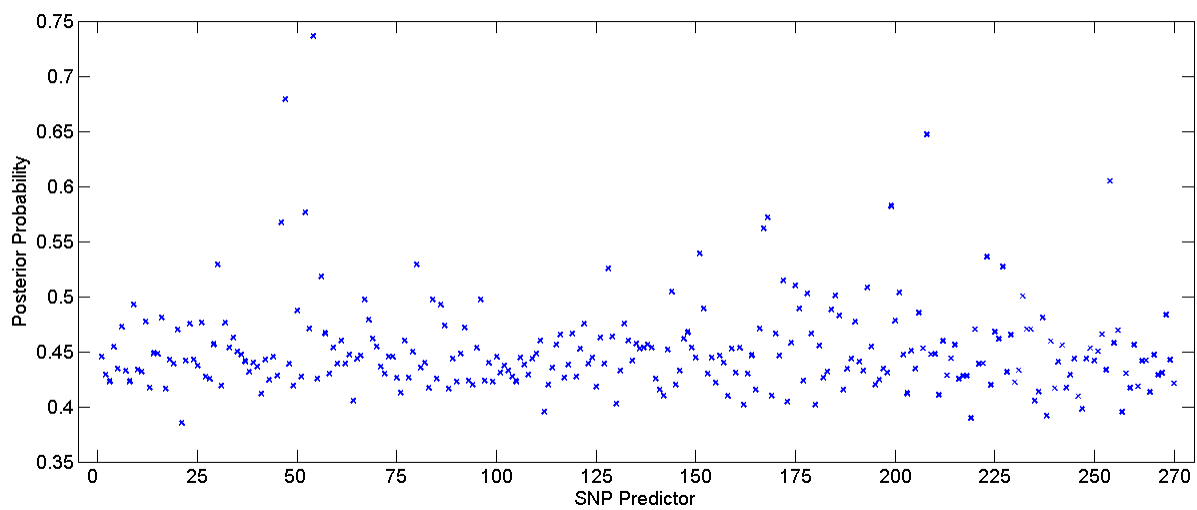


Figure 3: Posterior probability of genotype effects on early onset multiple myeloma.