Published in final edited form as:

Biometrics. 2010 June ; 66(2): 455–462. doi:10.1111/j.1541-0420.2009.01275.x.

Bayesian Semiparametric Multiple Shrinkage

Richard F. MacLehose^{1,2} and David B. Dunson³

Richard F. MacLehose: mac10029@umn.edu

¹Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A

²Division of Epidemiology and Community Health, University of Minnesota, Minneapolis, Minnesota 55454, U.S.A

³Department of Statistical Science, Duke University, Durham, North Carolina 27708, U.S.A

Summary

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 nonnull values; however, such information may not be available. We propose a Bayesian semiparametric 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 nonzero, mean. Our approach favors sparse, but flexible, structure by shrinking toward a small number of random locations. The methods are illustrated using a study of genetic polymorphisms and Parkinson's disease.

Keywords

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, Haile, and Duggan, 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.

© 2009, The International Biometric Society

Correspondence to: Richard F. MacLehose, mac10029@umn.edu.

Supplementary Materials

Web Tables and Figures referenced in Section 4 and 5 are available under the Paper Information link at the Biometrics website <http://www.biometrics.tibs.org>

It has long been known that shrinkage, or regularization, can improve estimation performance, reducing mean square error while introducing bias (Hoerl and Kennard, 1970a). 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, 1970b) 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. 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 toward 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 prespecified groups, allowing coefficients in each group to be shrunk toward group-specific means as Greenland (1992, 1994). However, these approaches assume significant prior knowledge, which will be lacking in many instances. George (1986a, 1986b, 1986c) proposes a minimax multiple shrinkage estimator that is a mixture of James–Stein estimators and allows coefficients to shrink toward a fixed number of locations. Previous research by Dunson, Herring, and Mulherin-Engel (2008), MacLehose et al. (2007), and Dahl and Newton (2007) attempts to cluster coefficients for different predictors after standardization in order to borrow strength. Do, Müller, and Tang (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 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 Markov chain Monte Carlo (MCMC) algorithm related to that proposed recently by Papaspiliopoulos and Roberts (2008).

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

$p \approx n$: an analysis of SNP data in Parkinson's disease patients. Section 7 contains a discussion.

2. Semiparametric 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 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 $f(\boldsymbol{\beta}) = \prod_{j=1}^p \text{DE}(\beta_j|0, \tau)$. 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 Andrews and Mallows (1974),

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

where $\text{Exp}(\cdot|2/\tau)$ denotes an exponential distribution with mean $2/\tau$.

The prior distribution in expression 2 induces shrinkage toward the prior mean of zero; however, when some effects are nonnull, shrinkage toward these nonnull locations may be beneficial. We extend the lasso model specification by introducing a mixture prior with

separate prior location and scale parameters for each coefficient: $f(\boldsymbol{\beta}) = \prod_{j=1}^p \text{DE}(\beta_j|\mu_j, \tau_j)$. 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 of these hyperparameters through the use of a carefully tailored hyperprior. This is accomplished through a nonparametric DP prior, which 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 nonzero 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 of β_j to 0, while

specifying a_1 and b_1 to give support to smaller values of τ_j to allow less shrinkage toward nonzero 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 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. A priori the number of clusters will be proportional to $\alpha \log p$, thus smaller values of α favor a smaller number of clusters. However, the data also inform the number of clusters and the cluster-specific hyperparameters, so we obtain a procedure that adaptively shrinks coefficients toward nonzero 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:

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

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 that 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 $V_t \sim \text{Beta}(V_t | 1, \alpha)$. As is apparent in Figure 1, the prior distribution for each regression coefficient is simply a multimodal extension of the double exponential prior.

Both the prior locations and scales of the coefficients are clustered. Thus, if a number of coefficients fall into the same bin, the extent to which they will be shrunk toward the prior mean of that bin will partly depend 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 prior 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 hyperparameter choices that can be implemented as plug-in values 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 nonnull bins. We recommend choosing smaller values for these hyper-parameters 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_t 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 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.

3. Posterior Computation

Our approach initially augments the data using the O'Brien and Dunson (2004) algorithm by assuming the outcome $y_i = 1$ occurs when a latent variable, $g_i > 0$. We assume $g_i = \gamma + \mathbf{x}_i' \boldsymbol{\beta} + \epsilon_i / \phi_i$, 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.

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

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

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 $f(\phi_i | \boldsymbol{\beta}, g_i)$.

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

(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 $V_\beta = (\mathbf{X}'\Gamma^{-1}\mathbf{X} + \Lambda^{-1})^{-1}$ and $E_\beta = V_\beta(\mathbf{X}'\Gamma^{-1}\mathbf{g} + \Lambda^{-1}\boldsymbol{\mu}_0)$. 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 . Park and Casella (2005) show that $1/\lambda_j$ has

inverse Gaussian (iG) distribution, with parameters $a = \sqrt{\frac{\tau_{k_j}^*}{(\beta_j - \mu_{k_j}^*)^2}}$ and $b = \tau_{k_j}^*$.

The iG distribution is defined as $iG(y|a, b) = \sqrt{\frac{b}{2\pi}} y^{-3/2} \exp[-\frac{b}{2y}(\frac{y}{a} - 1)^2]$ with mean a and variance a^3/b and samples can be drawn from an iG 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 (2008). 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 prior's 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 $((\mu_2^*, \tau_2^*))$. 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 $f(\mu_1^*|\tau_1^*, \beta) = \delta(\mu_1^*|0)$. The conditional posterior for the scale parameter in the first bin is given by

$$f(\tau_1^*|\boldsymbol{\lambda}, \mathbf{k}) \propto \text{Gamma}(\tau_1^*|a_0, b_0) \prod_{j:k_j=1} \text{Exp}\left(\lambda_j \middle| \frac{2}{\tau_1^*}\right) \propto \text{Gamma}\left(\tau_1^*|m_1+a_0, \frac{1}{\sum_{j:k_j=1} \lambda_j/2+1/b_0}\right). \quad (8)$$

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

$$f(\mu_t^*|\tau_t^*, \beta, \mathbf{k}) \propto N(\mu_t^*|c, d) \prod_{j:k_j=t} N(\beta_j|\mu_t^*, \lambda_j) \propto N(\mu_t^*|\widehat{E}_{\mu_t}, \widehat{V}_{\mu_t}) \quad (9)$$

$$f(\tau_t^*|\boldsymbol{\lambda}, \mathbf{k}) \propto \text{Gamma}\left(\tau_t^*|m_t+a_1, \frac{1}{\sum_{j:k_j=t} \lambda_j/2+1/b_1}\right), \quad (10)$$

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 or more of the bins and in this case, sampling $((\mu_t^*, \tau_t^*))$ amounts to sampling from the prior distribution, $N(\mu_t^*|c, d) \times \text{Gamma}(\tau_t^*|a_1, b_1)$.

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

$$f(V_t|\mathbf{k}, \alpha) \sim \text{Beta}\left(m_t+1, p - \sum_{l=1}^t m_l + \alpha\right) \pi_t = \prod_{h=1}^{t-1} (1 - \pi_h) V_t. \quad (11)$$

- (4c) (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

$$q_j(l) \propto \begin{cases} \pi_l N(\beta_j|\mu_l^*, \lambda_j) \text{Exp}\left(\lambda_j \left| \frac{2}{\tau_l^*} \right| \right) & \text{for } l \leq \max(\mathbf{k}) \\ \pi_l M_j(\mathbf{k}) & \text{for } l > \max(\mathbf{k}). \end{cases} \quad (12)$$

Following the recommendation of Papaspiliopoulos and Roberts (2008) we specify

$$M_j(\mathbf{k}) = \max(N(\beta_j|\mu_l^*, \lambda_j) \text{Exp}\left(\lambda_j \left| \frac{2}{\tau_l^*} \right| \right), l \leq \max(\mathbf{k})) \text{ and the normalizing constant for } q_j$$

$$N_j(\mathbf{k}) = \sum_{l=1}^{\max(\mathbf{k})} q_j(l) + M_j(\mathbf{k}) (1 - \sum_{l=1}^{\max(\mathbf{k})} \pi_l).$$

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 l is $q_j(l)$. If

$$U_j > \sum_{l=1}^{\max(\mathbf{k})} q_j(l), \text{ let } p^* = p^* + 1 \text{ and draw new values of } (\mu_{p^*}^*, \tau_{p^*}^*) \text{ from their prior distributions until } U_j \leq \sum_{l=1}^{\max(K)} q_j(l).$$

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 $\mathbf{k}$ to configuration $\mathbf{k}'$ with probability

$$1 \text{ if } l \leq \max(\mathbf{k}) \text{ and } \max(\mathbf{k}) = \max(\mathbf{k}')$$

$$\min \left(1, \frac{n(\mathbf{k})}{n(\mathbf{k}')} \frac{M_j(\mathbf{k}')}{N(\beta_j|\mu_{k_j}^*, \lambda_j) \text{exp}\left(\lambda_j \left| \frac{2}{\tau_{k_j}^*} \right| \right)} \right) \text{ if } l \leq \max(\mathbf{k}) \text{ and } \max(\mathbf{k}') < \max(\mathbf{k})$$

$$\min \left(1, \frac{n(\mathbf{k})}{n(\mathbf{k}')} \frac{N(\beta_j|\mu_l^*, \lambda_j) \text{exp}\left(\lambda_j \left| \frac{2}{\tau_l^*} \right| \right)}{M_j(\mathbf{k})} \right) \text{ if } l > \max(\mathbf{k}).$$

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 predictors.

4. Simulations

Allowing shrinkage of model parameters to multiple locations is intuitively appealing; however, we examine the performance of this approach. We are interested in the performance of these estimators as the sample size and number of predictors change, so we simulate datasets under three combinations of sample size and predictors: (1) $N = 400$, $P = 20$; (2) $N = 300$, $P = 50$; and (3) $N = 100$, $P = 100$. To see how the estimators perform as the distribution of the true effects change, we generate the first 10 regression coefficients under three scenarios: (A) $\beta_j = 1$; (B) $\beta_j \sim \mathcal{N}(1, 0.1)$; and (C) $\beta_j \sim \mathcal{N}(1, 1)$. The remaining coefficients in each simulation are equal to zero. We present nine simulations in the text and additional results are available in an online supplement.

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. We implemented a standard Bayesian lasso through a Gibbs sampler similar to that in Park and Casella (2005). In particular, we assume $\beta_j \sim \mathcal{N}(0, \lambda_j)$ with $\lambda_j \sim \text{Exp}(2/\tau)$ 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. For an additional comparison to a regression model with a heavy-tailed prior, we placed a Cauchy prior on the regression coefficients. We implement the Cauchy prior through a scale mixture of Normals: $\beta_j \sim \mathcal{N}(0, 1/\lambda)$ with $\lambda \sim \text{Gamma}(1/2, 2/1)$.

Results of the simulations are presented in Table 1 and some general conclusions can be drawn from these results. Results from the first set of simulations indicate that the multiple shrinkage prior offers improvement over the standard Bayesian lasso and Cauchy prior. The MSE estimated over all simulated datasets are smaller for the multiple shrinkage prior. The reduction in mean square error (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 toward a cluster-specific prior mean. The first 10 coefficients are shrunk toward a prior mean that is close to 1, resulting in decreased bias for $\beta_1, \dots, \beta_{10}$ in the multiple shrinkage model (MSE = 0.04) compared to the standard lasso model (MSE = 0.09) or Cauchy prior (MSE = 0.06). Additionally, those co-efficients ($\beta_{11}, \dots, \beta_{20}$) whose prior means are clustered into the null bin and assigned a prior mean of zero are shrunk strongly toward that mean in the multiple shrinkage prior model (MSE = 0.03) as well as in the Bayesian lasso model (MSE = 0.03), but less so with the Cauchy prior (MSE = 0.06). If we assume that the nonnull coefficients are drawn from a normal distribution, the MSE of the multiple shrink-age model predictably increases; however, it is still lower than the lasso prior. It is worth noting that drawing the 10 nonnull coefficients from a $\mathcal{N}(1, 1)$ represents a worst-case scenario for the multiple shrinkage model resulting in coefficients that can rarely be clustered together.

Results from the second set of simulations illustrate that as the sample size decreases and number of predictors increase, the performance of all models predictably decreases. Larger MSE is seen in all three models. The multiple shrinkage model seems to have equivalent or marginally better performance in estimating nonnull effects than the Cauchy model, but is still far superior to the standard lasso. The multiple shrinkage model retains excellent performance at estimating null effects, however. Finally, when $N = P$ the multiple shrinkage model results suggest that the multiple shrinkage prior performs well in situations where $N = P$, allowing clusters to be adequately identified.

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 are typically split into training and testing datasets of size 200 and 332, respectively. We include variables for all two-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 semiparametric multiple shrinkage model was implemented as in Section 3 for 50,000 iterations. The chain converged rapidly and showed little autocorrelation. We excluded the initial 5000 iterations as a burn-in. For comparison, we implemented regression with Cauchy priors and a Bayesian version of the lasso by assuming $\beta_j \sim N(0, \lambda_j^2)$ with $\lambda_j^2 \sim \text{Exp}(2/\tau)$ and $\tau \sim \text{Gamma}(a, b)$. 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 four 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 heavy-tailed Cauchy prior accomplishes less shrinkage and less precise effect estimates than the other models. The multiple shrinkage prior model retained the ability to shrink some estimates strongly toward zero while allowing other estimates to be shrunk toward nonzero locations. For instance, prior location and scale parameters for the effect of the 12th predictor is clustered with the prior for the 16th predictor 74% 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.

The Pima Indian data have been widely used for purposes of prediction. We generated posterior predictive distributions for the outcome, $\tilde{y}_p$, of the new observations in the testing dataset by integrating over model parameters using the output of the MCMC algorithm. The multiple shrinkage prior approach was compared to the two Bayesian lassos and a Cauchy prior. The Bayesian lasso with $a = a_0$, $b = b_0$ had a misclassification rate of 23%, 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 slight further improvement, with a misclassification rate of 21%, and the Cauchy prior had slightly better prediction with a 20% misclassification rate. We present sensitivity analyses for this example and trace plots for convergence in online supplementary materials.

6. SNPs and Parkinson's Disease

Parkinson's disease is a common neurologic disorder characterized by tremors, rigidity, and slowness of movement and affects approximately 1 of every 100 people over the age of 60 in the United States. Relatively few environmental risk factors for Parkinson's disease have been found. Exposure to pesticides has been consistently reported to increase risk of Parkinson's disease (Kamel and Hoppin, 2004). The scarcity of established environmental risk factors for Parkinson's disease has led many researchers to focus on genetic causes of the disease. We analyze SNP data from 540 individuals, 270 of whom were diagnosed with Parkinson's and 270 neurologically normal controls, to determine whether any

polymorphisms are related to development of Parkinson's disease. The collection of these data and a previous analysis of them has been described in Fung et al. (2006).

We limit our analysis to 270 SNPs on chromosome 11 without any missing values. Indicator variables are created for each SNP with the most frequent homozygous genotype treated as the referent category, leading to 541 coefficients to estimate (540 SNP coefficients and 1 intercept term). We specified our multiple shrinkage model using the default hyperpriors and ran our MCMC sampling algorithm for 50,000 iterations, keeping every 5th sample and discarding the initial 5,000 iterations as a burn-in. The regression was also implemented using Bayesian lasso priors and Cauchy priors. Examination of trace plots indicated that the chain converged rapidly and exhibited good mixing. In genetic applications, interest often focuses less on effect estimation than on significance testing. As previously, we assume that

$|\beta_j| \in \epsilon$ 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.

Figure 3 shows the resulting posterior probabilities of an effect, $\Pr(|\beta_j| \in \epsilon)$, from each of the four models. The Bayesian lasso models seemed to either provide too little shrinkage (in the case where hyperpriors were set to $a = b = 6.5$) or too much shrinkage (when hyperpriors were set to $a = b = 30$) to provide useful results: for the former, all SNPs retain a relatively high probability of having an effect while for the latter, all SNPs seem to have an extremely small probability of an effect. The Cauchy model, because of its heavy tails, does not provide enough shrinkage toward the null and the resulting imprecise posterior distributions result in high posterior probabilities of effect. 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. It is interesting to note that in this example, where posterior probabilities are of greater interest than effect estimation, the multiple shrinkage model serves to isolate effects more clearly than the either of the Bayesian lasso's or the Cauchy prior. The vast majority of effects have posterior probability of approximately 0.05. Very few SNPs have large posterior probability, with only two coefficients having probability greater than 50% and only 1 greater than 60%. The two genotypes that clearly stand out as being most likely to have an effect on Parkinson's disease in this analysis are rs10501570 and rs12274110. The former occurs in the DLG2 gene, which is a member of the membrane-associated guanylate kinase family, and the latter occurs in a nearby noncoding region of the genome. We are not aware of other genome scans of Parkinson's patients that have found an effect in this region.

7. Discussion

This article has proposed a novel semiparametric 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 nonzero 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 a number of different settings. We note that the multiple shrinkage method works best when $N > P$ and the clusters can be easily identified. However, even when clusters are not easily identified or $N \sim P$, the multiple shrinkage method still performs well. The multiple shrinkage model has better MSE than the standard lasso model since it is able to shrink toward multiple prior means. It has superior

performance to the heavy-tailed Cauchy distribution when $N > P$; however, this performance decreases as the ratio of $N : P$ decreases. In estimating null effects, the multiple shrinkage model, by encouraging substantial shrinkage toward zero tends to perform better than the Cauchy prior. The multiple shrinkage prior model should be of great use to substantive researchers in many settings.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

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

References

- Andrews D, Mallows C. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Series B*. 1974; 36:99–102.
- Chhikara, R.; Folks, L. *The Inverse Gaussian Distribution*. New York: Marcel Dekker, Inc.; 1989.
- Dahl D, Newton M. Multiple hypothesis testing by clustering treatment effects. *Journal of the American Statistical Association*. 2007; 102:517–526.
- Do K, Müller P, Tang F. A Bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society, Series C: Applied Statistics*. 2005; 54:627–644.
- Dunson D, Herring A, Mulherin-Engel S. Bayesian selection and clustering of polymorphisms in functionally related genes. *Journal of the American Statistical Association*. 2008; 103:534–546.
- Escobar M, West M. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*. 1995; 90:577–588.
- Ferguson T. Prior distributions on spaces of probability measures. *The Annals of Statistics*. 1974; 2:615–629.
- Fung H, Scholz S, Matarin M, Simn-Snchez J, Hernandez D, Britton J, Gibbs J. Genome-wide genotyping in Parkinson's disease and neurologically normal controls: First stage analysis and public release of data. *Lancet Neurology*. 2006; 5:911–916. [PubMed: 17052657]
- Gelman, A.; Jakulin, A.; Pittau, M.; Su, Y. Technical report. Columbia University; 2006. A default prior distribution for logistic and other regression models.
- George E. Combining minimax shrinkage estimators. *Journal of the American Statistical Association*. 1986a; 81:437–445.
- George E. A formal Bayes multiple shrinkage estimator. *Communications in Statistics: Theory and Methods*. 1986b; 15:2099–2114.
- George E. Minimax multiple shrinkage estimation. *The Annals of Statistics*. 1986c; 14:188–205.
- Geweke, J. Variable selection and model comparison in regression. In: Bernardo, J.; Berger, J.; Dawid, A.; Smith, A., editors. *Bayesian Statistics*. Vol. 5. Oxford: Oxford Press; 1996. p. 609–620.
- Greenland S. A semi-Bayes approach to the analysis of correlated multiple associations, with an application to an occupational cancer-mortality study. *Statistics in Medicine*. 1992; 11:219–230. [PubMed: 1579760]
- Greenland S. Hierarchical regression for epidemiologic analyses of multiple exposures. *Environmental Health Perspectives*. 1994; 102(Suppl 8):33–39. [PubMed: 7851328]
- Griffin, J.; Brown, P. Technical report. University of Warwick; 2006. Alternative prior distributions for variable selection with very many more variables than observations.
- Hoerl A, Kennard R. Ridge regression: Applications to nonorthogonal problems. *Technometrics*. 1970a; 12:69–82.
- Hoerl A, Kennard R. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*. 1970b; 12:55–67.

- Kamel F, Hoppin J. Association of pesticide exposure with neurologic dysfunction and disease. *Environmental Health Perspectives*. 2004; 112:950–958. [PubMed: 15198914]
- Loennstedt I, Britton T. Hierarchical Bayes models for CDNA microarray gene expression. *Biostatistics*. 2005; 6:279–291. [PubMed: 15772106]
- MacLehose R, Dunson D, Herring A, Hoppin J. Bayesian methods for highly correlated exposure data. *Epidemiology*. 2007; 18:199–207. [PubMed: 17272963]
- O'Brien S, Dunson D. Bayesian multivariate logistic regression. *Biometrics*. 2004; 60:739–746. [PubMed: 15339297]
- Papaspiliopoulos O, Roberts G. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*. 2008; 95:169–186.
- Park, T.; Casella, G. Technical report. University of Florida; 2005. The Bayesian lasso.
- Sethuraman J. A constructive definition of the Dirichlet process prior. *Statistica Sinica*. 1994; 2:639–650.
- Thomas D, Haile R, Duggan D. Recent developments in genome-wide association scans: A workshop summary and review. *American Journal of Human Genetics*. 2005; 77:337–345. [PubMed: 16080110]
- Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*. 1996; 58:267–288.

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 development of Parkinson's disease.

Table 1

Simulation study examining the mean squared error of the multiple shrinkage prior, Lasso prior, and Cauchy prior under various sample size and distributional assumptions

Coefficients		Prior		
		Multiple shrinkage	Lasso	Cauchy
Nonnull	Null	MSE (Nonnull coefficients/null coefficients)		
N= 400, P= 20				
$\beta_1, \dots, \beta_{10} = 1$	$\beta_{11}, \dots, \beta_{20} = 0$	0.04 / 0.03	0.09 / 0.03	0.06 / 0.06
$\beta_1, \dots, \beta_{10} \sim \mathcal{N}(1, 0.25^2)$	$\beta_{11}, \dots, \beta_{20} = 0$	0.07 / 0.02	0.10 / 0.03	0.07 / 0.07
$\beta_1, \dots, \beta_{10} \sim \mathcal{N}(1, 1)$	$\beta_{11}, \dots, \beta_{20} = 0$	0.09 / 0.02	0.10 / 0.03	0.09 / 0.08
N=300, P=50				
$\beta_1, \dots, \beta_{10} = 1$	$\beta_{11}, \dots, \beta_{50} = 0$	0.11 / 0.01	0.18 / 0.02	0.11 / 0.05
$\beta_1, \dots, \beta_{10} \sim \mathcal{N}(1, 0.25^2)$	$\beta_{11}, \dots, \beta_{50} = 0$	0.14 / 0.01	0.18 / 0.02	0.13 / 0.07
$\beta_1, \dots, \beta_{10} \sim \mathcal{N}(1, 1)$	$\beta_{11}, \dots, \beta_{50} = 0$	0.18 / 0.01	0.21 / 0.02	0.15 / 0.07
N=100, P=100				
$\beta_1, \dots, \beta_{10} = 1$	$\beta_{11}, \dots, \beta_{100} = 0$	0.48 / 0.06	0.53 / 0.02	0.54 / 0.36
$\beta_1, \dots, \beta_{10} \sim \mathcal{N}(1, 0.25^2)$	$\beta_{11}, \dots, \beta_{100} = 0$	0.55 / 0.10	0.56 / 0.02	0.41 / 0.25
$\beta_1, \dots, \beta_{10} \sim \mathcal{N}(1, 1)$	$\beta_{11}, \dots, \beta_{100} = 0$	0.66 / 0.06	1.01 / 0.02	1.76 / 0.83