Bayesian Methods for Highly Correlated Exposure Data

Richard F. MacLehose[1,2], David B. Dunson[2], Amy H. Herring[3,4], Jane A. Hoppin[5]

[1]Department of Epidemiology, University of North Carolina at Chapel Hill

[2]Biostatistics Branch, National Institute of Environmental Health Sciences, National Institutes of Health

[3]Department of Biostatistics, University of North Carolina at Chapel Hill

[4]Carolina Population Center, University of North Carolina at Chapel Hill

[5]Epidemiology Branch, National Institute of Environmental Health Sciences, National Institutes of Health

Abstract

Studies that include individuals with multiple highly correlated exposures are common in epidemiology. Because standard maximum likelihood techniques often fail to converge in such instances, hierarchical regression methods have seen increasing use. Bayesian hierarchical regression places prior distributions on exposure-specific regression coefficients to stabilize estimation and incorporate prior knowledge, if available. A common parametric approach in epidemiology is to treat the prior mean and variance as fixed constants. An alternative parametric approach is to place distributions on the prior mean and variance to allow the data to help inform their values. As a more flexible semi-parametric option, one can place an unknown distribution on the coefficients that simultaneously clusters exposures into groups using a Dirichlet process prior. We also present a semi-parametric model with a variable-selection prior to allow clustering of coefficients at zero. We compare these four hierarchical regression methods and demonstrate their application in an example estimating the association of herbicides on retinal degeneration among wives of pesticide applicators.

Highly correlated exposures are ubiquitous in epidemiologic research, and may arise due to an association between the measured exposures and one or more latent factors. For example, pesticide exposures for farm workers may be highly correlated because individuals apply multiple pesticides in a year, with choice of pesticide influenced by type of crop and pest.[1,2] To depict this correlated exposure problem, let $x_1, \ldots, x_k$ denote the levels of k different exposure variables that are highly correlated due to an unmeasured variable or variables, and let y denote the outcome. Researchers will generally be interested in estimating effect measures $\beta_1, \ldots, \beta_k$ for exposures $x_1, \ldots, x_k$. Hence, a common strategy is to fit the logistic regression model:

$$\text{logit}\{\Pr(y_i = 1 \,|\, x_{i1}, \ldots, x_{ik})\} \;\; = \;\; \alpha_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}. \qquad (1)$$

Unfortunately, maximum likelihood estimation of the model in expression (1) can fail to converge when predictors are highly correlated, and estimated coefficients may be unstable even when convergence is achieved.[3]

This problem has led many epidemiologists to fit logistic regression models incorporating one exposure variable at a time. However, the other exposure variables may be confounders and, if so, must be included in order to assess the causal effect of any specific exposure.[4] Another commonly-used strategy is to collapse the specific exposure information into summaries, such as a sum across chemicals in a class or an ever/never indicator. Unfortunately, this strategy results in a loss of information, does not allow inferences on effects of specific exposures, and can be sensitive to the chosen summary measure.

The problems associated with performing maximum likelihood estimation on correlated data have helped motivate increased use of hierarchical models.[5] Ordinary regression models treat the outcome as a random variable, dependent on parameters. For example, in expression (1), y_i is a random variable that depends on the parameters α_0 and $\beta_1 \ldots \beta_k$. Hierarchical regression extends ordinary regression by also treating parameters as ran-

dom variables depending on further coefficients through a prior distribution. Estimates obtained through hierarchical regression are shrinkage estimates in the sense that they are moved away from the asymptotically unbiased maximum likelihood estimate (MLE) and toward the center of the prior distribution. Shrinkage estimators are advantageous in two ways: they often have smaller frequentist mean squared error (MSE) and they represent incorporation of prior knowledge in the Bayesian sense.[6] Such hierarchical models help circumvent problems associated with MLE. Namely, hierarchical models can estimate effects with lower MSE, even in the presence of high correlation.[3,7]

We discuss 4 Bayesian hierarchical models: 2 parametric models (P1 and P2) and 2 semi-parametric models (SP1 and SP2). These 4 models differ in how their prior distribution is specified. The most common Bayesian hierarchical model found in epidemiologic research is the semi-Bayes model,[5,8–13] which we refer to as model P1 (i.e., the 1^{st} parametric model). A typical prior distribution for β_j (where j indexes the k coefficients in expression [1]) is $N(\mu, \phi^2)$, where μ characterizes the investigator's prior knowledge about the true value of β_j and ϕ^2 is the uncertainty regarding that value. Values for μ and ϕ^2 are chosen based on substantive knowledge. The amount that the estimated effects are shrunk away from the MLEs and toward the prior mean is determined by the prior variance, ϕ^2. A large prior variance indicates greater uncertainty about the effect size and causes less shrinkage.

Consider a model such as expression (1) in which 20 coefficients are estimated and each has a $N(0, \phi^2)$ prior. Prior knowledge may exist about the variability of the estimates, but the data also contain information about that variability, with a simplistic estimate being the variance of the 20 MLEs about the prior mean. Model P1 incorporates prior knowledge by treating ϕ^2 as known, but it ignores information regarding the variability of the coefficients that is contained in the observed data. Thus, model P1 has fixed

shrinkage regardless of the support for the prior distribution provided by the data.

Consider, instead, a model that treats these prior parameters as random variables in turn having their own prior distributions (model P2). Unlike model P1, which has fixed shrinkage (because ϕ^2 is constant), model P2 estimates ϕ^2 by combining the observed data with prior knowledge about ϕ^2. This allows the amount of shrinkage to vary depending on how well the data support the prior distribution. If the data lend some support to the prior distribution, model P2 can provide greater shrinkage than model P1. If the data lend little support to the prior distribution, model P2 will result in less shrinkage. In the discussion so far, all coefficients have been shrunk toward a common mean; however, it is straightforward to allow coefficients to be grouped into classes with each set of coefficients shrunk toward separate class-specific means.

Models P1 and P2 have potential disadvantages. A normally distributed prior is commonly assumed for historical reasons and computational convenience; however, results may be sensitive to this assumption. Second, for these methods to shrink estimates towards multiple prior means, the coefficients must be specified into classes (e.g., if the coefficients are the effects of different pesticides, they could be classified as fungicides or fumigants to allow coefficients in those classes to be shrunk toward different means). However, it may be impossible to specify which effects should be grouped into which classes, or even how many classes there should be. A method that allows the data to guide the clustering of coefficients into classes would be preferable. To accomplish this, we place a Dirichlet process prior (DPP) on the distribution of the coefficients.[14–16] A DPP allows researchers to specify their prior knowledge as being "similar" to a known parametric distribution (such as the normal), while remaining flexible enough to allow for substantial deviations from that distribution. Additionally, the DPP attempts to cluster coefficients into groups based on effect size. Coefficients are clustered together probabilistically (soft clustering) rather than

with certainty (hard clustering) and this feature of DPPs can offer dramatic improvements in effect estimation. We will refer to this semi-parametric model with DPP priors as model SP1.

In epidemiologic studies some exposures will typically have virtually no effect, in which case they cannot confound the effect of any other exposure and we might prefer to exclude them from the model. Variable-selection techniques in the epidemiologic literature are limited, generally relying on backward or forward selection strategies that increase the type I error rate.[17–19] However, there has been an increasing focus on variable-selection methods (implemented through variable-selection priors) in the statistics literature based on the advent of microarray technology.[20,21]

To account for the opportunity that some $\beta_j = 0$ we propose a mixture prior that allows an unknown subset of the predictors to have zero coefficients.[22,23] A coefficient is implicitly removed from the model when $\beta_j = 0$, a probability we estimate by combining our prior knowledge of a null effect with the observed data. When using a DPP for the coefficients, the exposures are clustered into groups. By using this mixture prior, we also allow a cluster of exposures that has coefficients equal to zero. Adopting this prior distribution in the DPP to perform simultaneous variable selection and clustering has been shown to have excellent properties.[24] We refer to the semi-parametric model with clustering of coefficients at zero as model SP2.

Parametric models

Both parametric models (P1 and P2) have been discussed in much greater detail elsewhere.[5,6,11,12,25,26] Here, we illustrate some of their properties in the simple setting of an ordinary linear regression model in which centered covariates $x_{i1} \ldots x_{ik}$ are regressed on an outcome y_i. For ease of presentation, we assume the linear model has a known error term,

σ^2, and that the covariates are orthogonal (i.e., they are not correlated); however, the results are generalizable to non-orthogonal situations.

As mentioned above, model P1 incorporates information on β_j through a prior distribution. A typical specification for this model is:

$$\begin{aligned} [y_i|\beta_j] &\sim N\Big(\sum_{j=1}^{k} \beta_j x_{ij}, \sigma^2\Big) \\ [\beta_j] &\sim N\Big(\eta_j, \phi_j^2\Big) \end{aligned} \tag{2}$$

where the prior mean, η_j, incorporates prior evidence regarding the size of the effect for the j^{th} coefficient and the x_{ij} may be standardized so they are all on the same scale. Prior scientific knowledge may indicate that all coefficients have the same prior distribution, that some coefficients have one prior distribution while others have a different prior distribution, or that each coefficient has its own prior distribution. For example, if $\beta_1 \ldots \beta_k$ are the effects of pesticides on retinal degeneration, one could assume that the effects of all pesticides are the same (eg, they all belong to the same class and have a common prior distribution), that the effect varies over different functional groups of pesticides (eg, they could be grouped into classes such as fungicide or fumigant, with each class having a different prior distribution), or that each pesticide has a different prior distribution.[2] Indicator variables, z_{lj}, denoting a pesticide class can be introduced into the prior distribution by allowing $\eta_j = \sum_{l=1}^{p} \theta_l z_{lj}$. However, these classes need not be mutually exclusive and more complicated prior specifications can be included where biologically relevant. The prior variance ϕ_j^2 represents the uncertainty that $\beta_j = \eta_j$. The prior variance could be specified from a meta-analysis or could be calculated by choosing a range within which the researcher believes 95% of effect estimates on this topic would lie. Solving the standard confidence interval formula for the variance term allows the researcher to specify the prior variance. The lack of a prior distribution on η_j or ϕ_j^2 is the distinguishing feature of model

P1.

The posterior distribution (i.e., the distribution that results when the prior distribution for β_j is updated with the observed data) for β_j is given by:

$$[\beta_j|Data] \sim N\Big(\frac{\eta_j/\phi_j^2 + \sum x_{ij}y_i/\sigma^2}{1/\phi_j^2 + \sum x_{ij}^2/\sigma^2}, \frac{1}{1/\phi_j^2 + \sum x_{ij}^2/\sigma^2}\Big) \tag{3}$$

The posterior mean is an average of the prior mean (η_j) and the maximum likelihood estimate ($\sum x_{ij}y_i/\sum x_{ij}^2$), inversely weighted by their respective variances, ϕ_j^2 and $\sigma^2/\sum x_{ij}^2$. This is the essence of a shrinkage estimator: the posterior distribution of β_j is shrunk towards its prior distribution. As the number of observations increases, the posterior distribution is weighted more heavily toward the observed data. With orthogonal data of moderate size, the observed data will quickly overwhelm anything but the strongest priors (i.e., those with very small ϕ^2), and estimates obtained from these parametric Bayesian models will be similar to the MLE. For concreteness, we generated a small (n=50) dataset with 5 orthogonal covariates, none of which have an effect. We estimate $\beta_1 \ldots \beta_5$ using MLE and using model P1 in expression 2 with $k = 5$ and $\eta_j = 0$ for all j. Figure 1 shows the resulting distribution of the MLE of β_1, as well as the Bayesian estimate with $\phi_j^2 = 0.5, 1.0$, and 2.0. Note that, on average, the MLE will be unbiased but in this single sample the results are far from the truth. The amount of shrinkage in the hierarchical models is a function of the prior variance: as the prior variance decreases (representing increasing certainty about the effect of β), the posterior distribution shrinks toward the prior mean.

Because ϕ^2 is so vital to model P1, it is wise to vary it in sensitivity analyses and see how estimates change with different plausible values. In Figure 1, for example, $\phi^2 = 1.0$ may have been the best guess of the variance of β with sensitivity analyses conducted for $\phi^2 = 2.0$ and $\phi^2 = 0.5$. However, although there may be uncertainty regarding the prior variance (leading to sensitivity analyses), estimates from model P1 cannot account for this uncertainty.

Model P2 explicitly accounts for uncertainty in the prior variance by placing a prior distribution on ϕ^2, resulting in estimates that are averaged over plausible values of ϕ^2. Unlike the fixed shrinkage of model P1, model P2 adapts the shrinkage of β_j based on the observed variability of $\beta_1 \ldots \beta_k$ from their prior mean. Additionally, when the prior mean is a function of covariates (e.g., $\eta_j = \sum \theta_l z_{lj}$), substantive information may exist for the effect of those variables and a prior distribution can be placed on those parameters. A typical specification for model P2 is:

$$\begin{aligned}
[y_i|\beta_j] &\sim N\Big(\sum_{j=1}^{k} \beta_j x_{ij}, \sigma^2\Big) \\
[\beta_j|\theta, \phi_j^2] &\sim N\Big(\sum_{l=1}^{p} \theta_l z_{lj}, \phi_j^2\Big) \\
[\theta_l] &\sim N(\mu_l, \omega_l^2) \\
[\phi_j^2] &\sim IG(\alpha_{1j}, \alpha_{2j}) \qquad (4)
\end{aligned}$$

Here, θ_l is the effect of a z_{lj} covariate and its prior mean, μ_l, is the prior knowledge regarding the size of θ_l's effect; the prior variance ω_l^2 represents uncertainty in that effect. The prior distribution for the ϕ_j^2 is chosen as an inverse gamma (IG) distribution with parameters α_{1j} and α_{2j}. The inverse gamma distribution is a common choice for the prior distribution of a variance term because of its flexibility and computational convenience. The prior mean of ϕ_j^2 is $\alpha_{2j}/(\alpha_{1j} - 1)$ and its variance is $\alpha_{2j}^2/((\alpha_{1j} - 1)^2(\alpha_{1j} - 2))$. Model P1 is a special case of model P2 in which the variance of θ and ϕ^2 goes to zero. In choosing values of α_1 and α_2 for an analysis, we suggest specifying a most likely value of ϕ^2 (call this E_{ϕ^2}) and a value for the variance of ϕ (call this V_{ϕ^2}) such that 95% of the reasonable ϕ^2 values would fall within the 95% confidence interval (CI) for $IG(\mathrm{E}_{\phi^2}, \mathrm{V}_{\phi^2})$. Solving the mean and variance equations for α_1 and α_2 gives: $\alpha_1 = \mathrm{E}_{\phi^2}^2/\mathrm{V}_{\phi^2} + 2$ and $\alpha_2 = \mathrm{E}_{\phi^2}^3/\mathrm{V}_{\phi^2} + \mathrm{E}_{\phi^2}$. It is useful to plot a large number (n $\approx$ 10,000) of samples from the prior to ensure that the shape of distribution conforms to prior knowledge.

The full conditional posterior distributions for the parameters in model P2, assuming ϕ^2 is the same for all β_j (for simplicity) are:

$$[\beta_j|Data, \sigma_2^2, \theta_j, \phi^2] \sim N\Big(\frac{\sum_l \theta_l z_{lj}/\phi^2 + \sum x_{ij}y_i/\sigma^2}{1/\phi^2 + \sum x_{ij}^2/\sigma^2}, \frac{1}{1/\phi^2 + \sum x_{ij}^2/\sigma^2}\Big) \quad (5)$$

$$[\theta_j|Data, \beta_j, \phi^2] \sim N\Big(\frac{\mu_l/\omega_l^2 + \sum z_{lj}\beta_j/\phi^2}{1/\omega_l^2 + \sum z_{lj}^2/\phi^2}, \frac{1}{1/\omega_l^2 + \sum z_{lj}^2/\phi^2}\Big) \quad (6)$$

$$[\phi^2|Data, \beta_j, \theta_j] \sim IG\Big(\alpha_1 + p/2, \alpha_2 + \frac{\sum_j(\beta_j - \sum_l z_{lj}\theta_j)^2}{2}\Big) \quad (7)$$

The conditional distribution of ϕ^2 in expression (7) is of particular interest. The adaptive shrinkage properties of model P2 are apparent from the $\sum_j(\beta_j - \sum_l z_{lj}\theta_j)^2$ term, that represents the variation of the β_j from their prior mean. As the variance of the parameters increases, the mean of ϕ^2 also increases and when the variance decreases, the mean of ϕ^2 decreases. Thus, if the data indicate that our prior specification of ϕ^2 is too small, the posterior mean of ϕ^2 is increased to reflect this. Because ϕ^2 determines the amount of shrinkage, β_j will be shrunk to a lesser extent. The converse is also true; when the data show little variability of the estimates from the prior mean, the posterior estimate of ϕ^2 will decrease and cause greater shrinkage of β_j to their prior distribution. This adaptive shrinkage is a potential improvement over model P1 that has a constant amount of shrinkage, regardless of the variability of the β_j from the prior mean that is observed in the data. Model P2 also allows inferences to be more data-driven and less sensitive to the prior specification of μ and ϕ^2.

The distribution of β_j in expression (5) is similar to the distribution in expression (3). However, the distribution from model P1 is conditional on known values, while the distribution from model P2 is conditional on random variables (ϕ^2 and θ_j). To average the distribution of β_j over these random variables, we use Gibbs sampling (a type of Markov Chain Monte Carlo) that proceeds by iteratively drawing parameter values from the full conditional distributions in expressions (5), (6) and (7), given the value of the other ran-

dom variables from the previous steps of the Gibbs sampler.[27,28] After running the Gibbs sampler for a large number of iterations and discarding some initial number of iterations to allow for a burn-in period, the mean and variance of β_j in the remaining samples are the mean and variance of the marginal posterior distribution of interest. For more information regarding burn-in period and convergence, consult Gelman et al.[29] We also note that these algorithms (which can be implemented in programs such as WinBUGS [MRC Bisotatistics Unit, Cambridge, UK]) generate the *exact* posterior distribution of the coefficients that is useful in small datasets. This result is an improvement over previous methods proposed for fitting model P1 that rely on asymptotic approximations.[30]

We analyze, under model P2, the dataset we previously examined for the model P1. The prior mean for all β_j is zero and the parameters for the prior variance, ϕ^2, are $\alpha_1 = 1$ and $\alpha_2 = 1$. We ran a Gibbs sampling algorithm for 50000 iterations and excluded the first 5000 iterations as a burn-in period. The marginal posterior distributions of β_1 and ϕ^2 are presented in Figure 2. The mean of $\beta_1 = -0.51$, which is between the mean of the estimates from model P1 under the assumption of a fixed $\phi^2 = 1$ ($\beta_1 = -0.56$) and $\phi^2 = .5$ ($\beta_1 = -0.43$). Although the mean of the prior variance was 1 in the model P2, $\beta_1 \ldots \beta_5$ exhibited less variability than the prior indicated, and the posterior mean of ϕ^2 (0.87) decreased to reflect this additional information. Thus, by incorporating information on ϕ^2 that is contained in the data, we adaptively allow greater shrinkage of β_1 towards its prior mean.

Although we have focused on linear regression with orthogonal data, the results can be generalized to correlated predictors and logistic regression. It is only for computational convenience that we have focused on linear models here. We implement logistic hierarchical models in simulations and the applied example presented later in this paper.

Semi-parametric models

As we demonstrate (Appendix I, available with the online version of this article), models P1 and P2 can offer a distinct improvement over MLE. However, results of these models may be sensitive to the assumed prior distribution of β_j and a non-parametric prior may be preferable. Further, when sufficient prior information exists, coefficients may be grouped into classes by incorporating second level coefficients; however, in many epidemiologic applications such prior knowledge may not exist. Instead, we explore a procedure that allows coefficients to be grouped into clusters based on similarity of effect sizes before shrinking them toward a prior distribution.

In Bayesian non-parametric inference, a common method to limit the dependence of a parameter on a prior distribution is to let the prior distribution be random. In the previous section we assumed $\beta_j \sim N(\mu, \phi^2)$. Instead, we could specify $\beta_j \sim D$, where D is a random distribution. Because D is random, we place a prior distribution on it; in this case we choose a Dirichlet process prior, $D \sim DPP(\lambda D_0)$, where D_0 is the base distribution (such as a normal distribution) and λ is a precision parameter determining how closely D follows D_0. As λ increases, D converges to D_0, and the non-parametric approach reduces to the parametric models of the previous section. Smaller values of λ indicate less certainty that $\beta_j \sim D_0$. Figure 3 presents two realizations of $DPP(\lambda D_0)$ with $D_0 \equiv N(0, 1)$ and λ equal to either 1 or 100. The larger value of λ yields a distribution that resembles the base distribution, while the sample with $\lambda = 1$ shows no similarity to the D_0.

A feature of the two distributions shown in Figure 3 is their discrete nature. Rather than being continuous, like the base distribution, a draw from a DPP is discrete, implying that any 2 (or more) coefficients have a nonzero probability of being clustered together and having the same effect size. The scale of the predictor may be important, and the predictors could potentially be rescaled to allow greater similarity among coefficients. The

clustering feature can be seen more clearly through the (identical) representation of the DPP as a mixture distribution: $\beta_j \sim w_0 D_0 + w_1 \sum_{i \neq j} \delta_{\beta_i}$, where w_0 and w_1 are weights determined by λ. The term δ_{β_i} indicates that, with probability w_1, β_j is clustered with coefficient β_i. The posterior probability of clustering coefficients depends on λ (smaller values of λ favor clustering) and the similarity of the magnitude of those coefficients (increased similarity favors clustering).

Consider two predictors, x_{im} and x_{in}, with effects β_m and β_n which follow some unknown distribution D that is assigned a DPP. This model estimates, based on prior knowledge and information in the data, a probability p_{mn} that $\beta_m = \beta_n$. In the extreme (and unlikely) case where $p_{mn} = 1$, coefficients x_{im} and x_{in} are estimating parameters with the same value ($\beta_m = \beta_n$). That is, the data contain twice as much information regarding the common effect, resulting in more precise effect estimates as well as less shrinkage toward the prior distribution. At the other extreme, if $p_{mn} = 0$, the two coefficients do not aid in each other's estimation. More commonly, p_{mn} will be between 0 and 1, allowing β_m and β_n to add some information to one another's estimation. This will result in different posterior distributions for the two coefficients, that can have lower MSE than model P1 or P2 (see Appendix I). In the sense that model P1 allowed for constant shrinkage of all coefficients toward the prior mean, and model P2 allowed for adaptive shrinkage of all coefficients toward the prior mean, the semi-parametric models (SP1 and SP2) allow individual coefficients to be adaptively shrunk toward the prior mean to different extents. The more likely coefficients are to be clustered together, the more information there is in the data regarding their common effect, and the less impact the prior specification will have.

This model is semi-parametric because the distribution of the outcome, y_i, is parametric, while the distribution of β_j is non-parametric. The first semi-parametric model (SP1)

is an extension of model P2 and can be specified as:

$$\begin{aligned} y_i &\sim N(\sum_{j=1}^{k} \beta_j x_{ij}, \sigma^2) \\ \beta_j &\sim D \\ D &\sim DP(\lambda D_0) \\ D_0 &\equiv N(\mu, \phi^2) \\ \lambda &\sim G(a, b) \\ \phi^2 &\sim IG(\alpha_1, \alpha_2), \end{aligned} \tag{8}$$

where G is a gamma distribution with mean ab and variance ab^2. Placing a prior distribution on the precision parameter, λ, serves the same function as placing a parameter on ϕ^2 in model P1; it allows the data to help guide inference rather than relying solely on prior knowledge. Generally, relatively noninformative values are chosen for a and b, such as $a = 1, b = 1$ or $a = 0.1, b = 0.1$. However, empirical Bayes methods are available to estimate this parameter as well.[31] As with the model P2, estimating these parameters requires a Gibbs sampling algorithm.[32,33]

In many instances it may be useful to exclude variables that have no effect on the outcome or there may be prior substantive knowledge that the exposure has no effect. In either case, modification of the base distribution D_0 in expression 8 allows a variable-selection prior distribution to be incorporated in a DPP model. Following the approach of Dunson et al.,[34] we specify a second semi-parametic model (SP2):

$$\begin{aligned} y_i &\sim N(\sum_{j=1}^{k} \beta_j x_{ij}, \sigma^2) \\ \beta_j &\sim D \\ D &\sim DP(\lambda D_0) \\ D_0 &\equiv \pi\delta_0 + (1-\pi)N(\mu, \phi^2) \end{aligned}$$

$$\begin{aligned} \lambda &\sim G(a,b) \\ \pi &\sim beta(c,d) \\ \phi^2 &\sim IG(\alpha_1,\alpha_2) \end{aligned} \tag{9}$$

where δ_0 indicates a point mass at the value 0. The base distribution has a value of 0 with probability π, and distribution $N(\mu,\phi^2)$ with probability $1-\pi$. This simple modification to the base distribution allows β_j to be equal to 0, in which case it is effectively removed from the model. This exclusion can help increase the precision of estimates, particularly in the presence of highly correlated variables or in small datasets. An important feature of allowing coefficients to be equal to 0 is that it allows for easy testing of a point hypothesis, such as $\beta_j = 0$. For instance, if a Gibbs sampling algorithm is run for R iterations and $\beta_j = 0$ for r of those iterations, the posterior probability that $\beta_j = 0$ is r/R.

When $\pi = 0$, model SP2 reduces to the model SP1. The coefficient π is given a $beta(c,d)$ distribution in order to allow the data to inform the probability that a coefficient is zero. Elicitation of c and d can proceed by specifying the expected probability, E_π, that a randomly selected coefficient is 0 and the variance surrounding that estimate, V_π. Solving the equations for the mean and variance of the beta distribution:

$$\begin{aligned} c &= \frac{\mathrm{E}_\pi^2 - \mathrm{E}_\pi^3}{\mathrm{V}_\pi} - \mathrm{E}_\pi \\ \mathrm{d} &= \frac{\mathrm{E}_\pi(\mathrm{E}_\pi - 1)^2}{\mathrm{V}_\pi} + \mathrm{E}_\pi - 1. \end{aligned}$$

Example: Application to Study of Pesticides and Retinal Degeneration

The Agricultural Health Study, which enrolled farmers who applied for pesticide licenses in Iowa or North Carolina between 1993 and 1997, has been described in more detail elsewhere.[1] Kirrane et al.[2] recently examined the association between pesticide exposure and retinal degeneration among the farmers wives. Wives filled out a questionnaire

with information on their medical and pesticide history. We analyzed the same data used by Kirrane et al. in their analysis (31,173 women, 281 of whom experienced retinal degeneration), but we limit our analysis to herbicides, of which there are 18 unique chemicals. These 18 chemicals exhibited a wide range of correlation, from 0.06 to 0.58. Table 2 shows the 4 hierarchical models used to analyze the data. Prior parameter values are based on prior knowledge and are similar to those used in Kirrane et al. There is little evidence of an effect of herbicides on retinal degeneration, so we center our prior distributions at OR=1.0. Gibbs sampling algorithms were programmed in Matlab (The Mathworks, Natick, MA) and run for 60,000 iterations, with the initial 5,000 excluded as a burn-in period.

To help illustrate the four hierarchical models, we present representations of the prior distributions for the effect of the herbicide imazethapyr (β_1) in Figure 4. Since the prior distributions for models P2, SP1 and SP2 depend on random variables, we evaluate their prior distributions at the posterior mean of all other random variables (ϕ^2 for model P2; ϕ^2, λ, and $\beta_2 \ldots \beta_{18}$ for model SP1; ϕ^2, λ, $\beta_2 \ldots \beta_{18}$, and π for model SP2). The prior distribution for model P1 is determined by our belief that herbicides most likely have no effect on retinal degeneration ($\exp(\mu)$=1.0) but that we are 95% certain the effect lies between approximately OR=0.3 and OR=3.1 ($\phi^2 = 0.35$). The prior distribution for β_1 in model P2 is more complicated since ϕ^2 is random. We observe little variability of the herbicide's effects about the prior mean, leading to a smaller posterior estimate of $\phi^2 = 0.11$. Thus the prior distribution for β_1 evaluated at $\phi^2 = 0.11$ is more concentrated around the null, leading to greater shrinkage of effects toward the prior mean. As indicated earlier, the prior distribution for model SP1 is a mixture of a normal distribution, with a mean OR=1.0 and posterior estimate of $\phi^2 = 0.17$, and a set of point masses at the posterior estimates of $\beta_2 \ldots \beta_{18}$. The mean posterior value of $\lambda = 1.8$ indicates that the data provide somewhat more evidence in favor of normally distributed effects than indicated by the

prior; this value implies that with probability 0.1, β_1 is distributed according to $N(0, 0.17)$, and with probability 0.9, β_1 is assigned the value of one of the other coefficients, $\beta_2 \ldots \beta_{18}$. The prior distribution for model SP2 is similar to model SP1, except for a large point mass at 0. The posterior mean of $\pi = 0.68$ and $\lambda = 1.5$ imply that β_1 is distributed according to $N(0, 0.18)$ with probability 0.03 or set equal to one of $\beta_2 \ldots \beta_{18}$ with probability 0.29 or set equal to 0 with probability 0.68.

The results of the models are presented in Table 3. Figure 5 shows the posterior distribution of the effect of imazethapyr from the four hierarchical models. Model P1 estimated an effect of imazethapyr that was no longer statistically significant but still markedly elevated (OR=1.7; 95% CI= 0.8-3.6). Models P2, SP1, and SP2 were all largely in agreement, indicating little evidence of effect of imazethapyr on retinal degeneration. The distribution of β_1 estimated through model SP2 is of particular interest. The large spike observed at 0 is the posterior probability that $\beta_1 = 0$ (p=0.63). Also of interest in the posterior distribution from model SP2 is the fact that the most likely non-null effect is virtually null, leading us to suspect this variable may have no association with retinal degeneration.

Discussion

Although highly correlated data are common in epidemiology, standard analyses can produce extremely imprecise confidence intervals or fail to converge altogether. We examined four Bayesian hierarchical models that perform well in this context. These models may have broad use in other areas, as well, such as in estimating models with a large number of predictors.

When deciding which of the four models to use in an analysis, consideration should be given to the properties of each model as well as the computational skill required to implement them. The two parametric models (P1 and P2) are the easiest computationally.

Either model can be implemented in WinBUGS, using the code we provide in Appendix II. The advantages of model P2 may justify its use in preference to model P1. Model P1 assumes a fixed prior variance, while model P2 updates the prior variance based on the observed data. This "Bayesian learning" allows for adaptive shrinkage and makes estimates more data-driven and less sensitive to prior specification. However, as the sample size increases, the difference between model; P1 and P2 (and ML) tends to decrease.

Although more computationally intensive than the parametric models, the two semiparametric models presented here have very desirable properties in many situations. These models may be particularly useful when the researcher is unaware how to group coefficients. When some coefficients have similar true values, the semiparametric models can decrease MSE by aggregating data within clusters. Indeed, even if the true values of the coefficients are not exactly identical, "soft clustering" can still reduce MSE (see Appendix I). However, as the probability of clustering coefficients increases, models SP1 and SP2 can perform remarkably well. The decision whether to implement model SP1 or SP2 should be made on substantive grounds. When researchers have a high prior probability that many of the effects in question may be zero, the selection prior in model SP2 can help estimation. Model SP2 may be particularly useful in situations where hypothesis testing is required. However, when the true value of most coefficients is zero and only a few coefficients are non-zero (but still close to zero), model SP2 performs slightly worse than model SP1.

In summary, the challenges of analyzing highly correlated data can be greatly diminished using the Bayesian framework. The 2 parametric and 2 semiparametric models we examine in this paper provide useful alternatives to current maximum likelihood techniques. The choice of model should be guided by careful thought regarding the likely magnitude of effects, as well as whether many effects of similar sizes may be seen.

Funding source: This research was supported by the Intramural Research Program of the NIH, and NIEHS.

References

1. Alavanja M, Sandler DP, McMaster SB, et al. The Agricultural Health Study. *Environ Health Perspect.* 1996;104:362-9.

2. Kirrane EF, Hoppin JA, Kamel F, et al. Retinal degeneration and other eye disorders in wives of farmer pesticide applicators enrolled in the Agricultrual Health Study. *Am J Epidemiol.* 2005;161:1020-9.

3. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics.* 1970;12:55-67.

4. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology.* 1999;10:37-48.

5. Greenland S. A semi-Bayes approach to the analysis of correlated multiple associations, with an application to an occupational cancer-mortality study. *Stat Med.* 1992;11:219-30.

6. Greenland S. Principles of multilevel modelling. *Int. J. Epid..* 2000;29:158-67.

7. Hoerl AE, Kennard RW. Ridge regression: applications to nonorthogonal problems. *Technometrics.* 1970;12:69-82.

8. De Roos AJ, Poole C, Teschke K, Olshan AF. An application of hierarchical regression in the investigation of multiple paternal occupational exposures and neuroblastoma in offspring. *Am J Ind Med.* 2001;39:477-86.

9. Engel SA, Erichsen HC, Savitz DA, Thorp J, Chanock SJ, Olshan AF. Risk of spontaneous preterm birth is associated with common proinflammatory cytokine polymorphisms. *Epidemiology.* 2005;16:469-77.

10. Engel SA, Olshan AF, Savitz DA, Thorp J, Erichsen HC, Chanock SJ. Risk of small-for-gestational age is associated with common anti-inflammatory cytokine polymorphisms. *Epidemiology.* 2005;16:478-86.

11. Greenland S. Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum-likelihood, preliminary-testing, and empirical-Bayes regression. *Stat Med.* 1993;12:717-36.

12. Greenland S. Hierarchical regression for epidemiologic analyses of multiple exposures. *Environ Health Perspect.* 1994;102 Suppl 8:33-9.

13. Greenland S, Poole C. Empirical-Bayes and semi-Bayes approaches to occupational and environmental hazard surveillance. *Arch Environ Health.* 1994;49:9-16.

14. Ferguson TS. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics.* 1973;1:209-230.

15. Ferguson TS. Prior distributions on spaces of probability measures. *The Annals of Statistics.* 1974;2:615-29.

16. Gopalan R, Berry DA. Bayesian multiple comparisons using Dirichlet process priors. *Journal of the American Statistical Association.* 1998;93:1130-1139.

17. Leamer EE. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley 1978.

18. Raftery AE. Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika.* 1996;83:251-66.

19. Draper D. Assessment and propagation of model uncertainty. *J. Roy. Statist. Soc. B.* 1995;57:45-70.

20. Efron B, Tibshirani R. Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology.* 2002;23:70-86.

21. Newton MA, Kendziorski CM, Blattner FR, Tsui KW. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology.* 2001;8:37-52.

22. Geweke J. Variable selection and model comparison in regression. in *Bayesian Statistics 5* (Bernardo JM, Berger JO, Dawid AP, Smith AFM. , eds.):609-620Oxford Press 1996.

23. Thomas DC, Siemiatycki J, Dewar R, Robins J, Goldberg M, Armstrong BG. The problem of multiple inference in studies designed to generate hypotheses. *Am J Epidemiol.* 1985;122:1080-95.

24. Ishwaran H, Rao JS. Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics.* 2005;33:730-73.

25. Lindley DV, Smith AFM. Bayes estimates for the linear model. *J. Roy. Statist. Soc (Ser. B).* 1972;34:1-41.

26. Browne WJ, Draper D. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Biostatistics.* 2006;1:473-514.

27. Geman S, Geman D. Stochastic relaxation, Giggs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 1984;6:721-41.

28. Casella G, George E. Explaining the Gibbs Sampler. *American Statistician.* 1992;46:167-74.

29. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis.* Chapman and Hall/CRC 2000.

30. Witte JS, Greenland S, Kim L. Software for Hierarchical Modeling of Epidemiologic Data. *Epidemiology.* 1998;9:563-6.

31. McAuliffe JD, Blei DM, Jordan MI. Nonparametric empirical Bayes for the Dirichlet process mixture model.. *Statistics and Computing, to appear.* 2005.

32. Escobar MD, West M. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association.* 1995;90:577-588.

33. Escobar MD, West M. Computing nonparametric hierarchical models. in *Practical Nonparametric and Semiparametric Bayesian Statistics* (Dey D., Muller P., Sinha D.. , eds.):1-22New York: Springer-Verlag 1998.

34. Dunson DB, Herring AH, Mulherin-Engel SM. Bayesian selection and clustering of polymorphisms in functionally related genes. *ISDS Tech Report, Duke University.* 2005.

35. Spiegelhalter DJ, Thomas A, NG Best. WinBUGS Version 1.2 User Manual. tech. rep.MRC Biostatistics Unit 1999.

36. Kamel F, Boyes WK, Gladen BC, et al. Retinal degneration in licensed pesticide applicators. *American Journal of Industrial Medicine.* 2000;37:618-628.

Table 1: Hierarchical models used to analyze Agricultural Health Study data on herbicides and retinal degeneration in wives of pesticide applicators, North Carolina and Iowa, 1993-1997.

P1			P2			SP1			SP2		
β_j	$\sim$	$N(0, 0.35)$	β_j	$\sim$	$N(0, \phi^2)$	β_j	$\sim$	D	β_j	$\sim$	D
			ϕ^2	$\sim$	$IG(2.13, 0.40)$	D	$\sim$	$DP(\lambda D_0)$	D	$\sim$	$DP(\lambda D_0)$
						D_0	$\equiv$	$N(0, \phi^2)$	D_0	$\equiv$	$\pi\delta_0 + (1-\pi)N(0, \phi^2)$
						λ	$\sim$	$G(1, 1)$	λ	$\sim$	$G(1, 1)$
						ϕ^2	$\sim$	$IG(2.13, 0.40)$	ϕ^2	$\sim$	$IG(2.13, 0.40)$
									π	$\sim$	$beta(1.5, 1.5)$

Table 2: Estimated effects of exposure to herbicides on retinal degeneration among the wives of pesticide applicators, Agricultural Health Study, North Carolina and Iowa, 1993-1997.*

Herbicide	**MLE** OR (95% CI)	**P1** OR (95% CI)	**P2** OR (95% CI)	**SP1** OR (95% CI)	**SP2** OR (95% CI)
Imazethapyr	2.6 (1.0- 6.3)	1.7 (0.8- 3.6)	1.3 (0.8- 2.5)	1.1 (0.6- 2.7)	1.1 (0.8- 2.3)
Chlorimuronethyl	1.9 (0.7- 5.0)	1.4 (0.6- 3.0)	1.2 (0.7- 2.1)	1.0 (0.6- 2.4)	1.0 (0.7- 1.9)
Alachlor	1.4 (0.6- 3.1)	1.2 (0.6- 2.2)	1.1 (0.6- 1.8)	1.0 (0.4- 1.5)	1.0 (0.6- 1.3)
Petroleum oil	1.4 (0.7- 2.9)	1.3 (0.7- 2.4)	1.2 (0.7- 1.9)	1.0 (0.4- 1.8)	1.0 (0.7- 1.4)
2,4,5-TP	1.3 (0.1- 11.2)	1.0 (0.3- 2.6)	1.0 (0.5- 1.9)	1.0 (0.4- 1.8)	1.0 (0.6- 1.6)
2,4-D	1.3 (0.8- 1.9)	1.2 (0.8- 1.8)	1.2 (0.8- 1.7)	1.0 (0.5- 1.6)	1.0 (0.6- 1.4)
Butylate	1.1 (0.3- 3.9)	1.0 (0.4- 2.3)	1.0 (0.5- 1.8)	1.0 (0.4- 1.8)	1.0 (0.6- 1.5)
Gylphosate	1.1 (0.8- 1.5)	1.0 (0.8- 1.4)	1.1 (0.8- 1.4)	0.9(0.5- 1.4)	1.0 (0.6- 1.3)
Dicamba	1.0 (0.4- 2.2)	1.0 (0.5- 1.9)	1.0 (0.6- 1.7)	1.0(0.5- 1.5)	1.0 (0.6- 1.4)
Trifluralin	1.0 (0.5- 2.1)	1.0 (0.5- 1.8)	1.0 (0.6- 1.7)	1.0 (0.4- 1.7)	1.0 (0.6- 1.5)
Cyanazine	0.9 (0.3- 2.5)	0.9 (0.4- 1.9)	0.9 (0.5- 1.6)	0.9 (0.4- 1.4)	1.0 (0.5- 1.3)
Metribuzin	0.9 (0.3- 3.1)	0.9 (0.4- 2.1)	1.0 (0.5- 1.7)	0.9 (0.4- 1.6)	1.0 (0.6- 1.4)
EPTC	0.8 (0.2- 3.4)	0.9 (0.4- 2.2)	0.9 (0.5- 1.7)	1.0 (0.4- 1.7)	1.0 (0.6- 1.5)
2,4,5-T	0.7 (0.1- 3.2)	0.9 (0.3- 2.0)	0.9 (0.5- 1.7)	1.0 (0.4- 1.6)	1.0 (0.6- 1.3)
Atrazine	0.6 (0.2- 1.4)	0.7 (0.3- 1.3)	0.8 (0.5- 1.3)	0.8 (0.2- 1.3)	1.0 (0.4- 1.2)
Metolachlor	0.5 (0.2- 1.4)	0.6 (0.3- 1.4)	0.8 (0.4- 1.4)	0.9 (0.2- 1.3)	1.0 (0.5- 1.2)
Pendimethalin	0.5 (0.2- 1.6)	0.7 (0.3- 1.6)	0.8 (0.4- 1.4)	0.9 (0.2- 1.3)	1.0 (0.4- 1.2)
Paraquat	0.3 (0.0- 2.1)	0.7 (0.3- 1.5)	0.8 (0.4- 1.4)	0.9 (0.3- 1.4)	1.0 (0.5- 1.3)

$*$ All models adjusted for state and age.

OR is odds ratio; CI stands for "confidence interval" for maximum likelihood and "credible interval" for Bayesian models; MLE is maximum likelihood estimate; 2,4,5-TP is 2,4,5-trichlorophenoxypropionic acid; 2,4,5-T is 2,4,5-trichlorophenoxyacetic acid; 2,4-D is 2,4-dichlorophenoxyacetic acid; EPTC is S-ethyl dipropylthiocarbamate

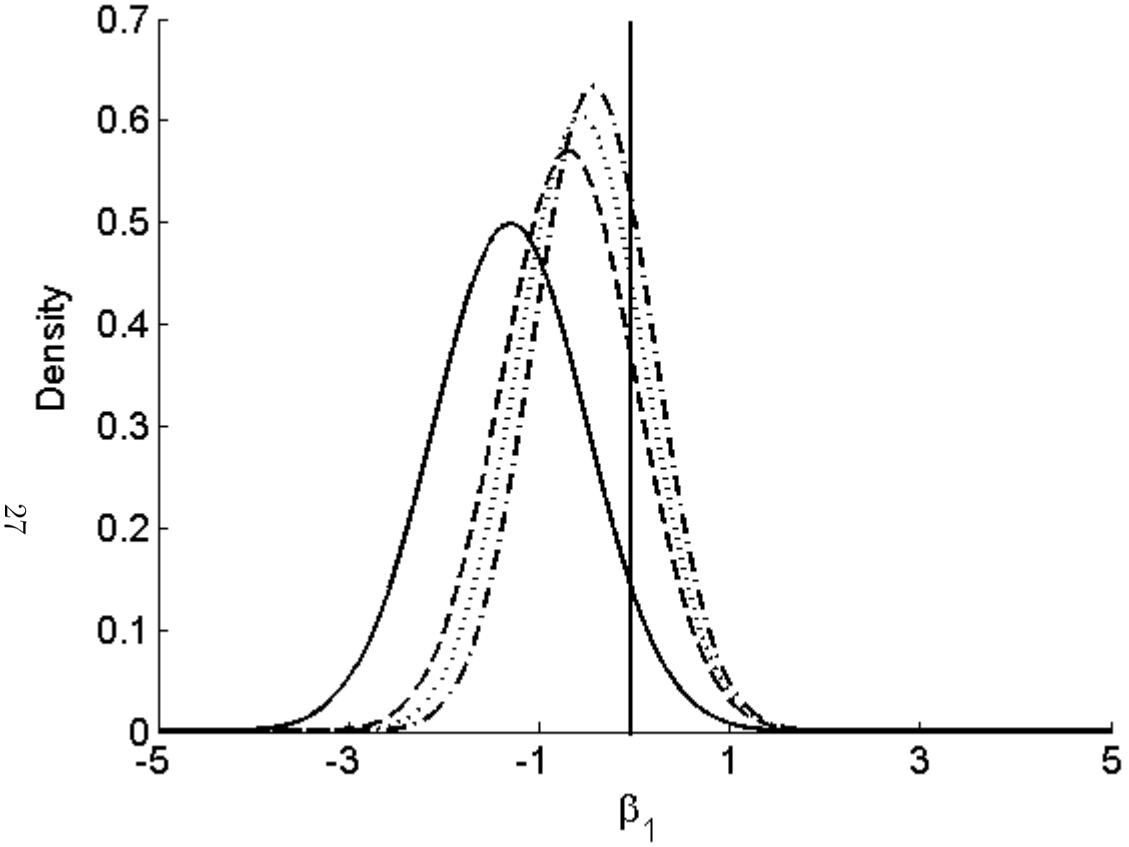
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Density
-5
-3
-1
1
3
5
β1

Figure 1: Estimates of β_1 obtained from maximum likelihood estimation and model P1 with prior variance of $\phi_j^2 = 2, 1, 0.5$. Data are a random sample with true $\beta_1 = 0$, n=50. Solid line indicates distribution of ML estimator; dashed line, distribution of β_1 with $\phi_j^2 = 2$; dotted line, distribution of β_1 with $\phi_j^2 = 1$; dash-dot line, distribution of β_1 with $\phi_j^2 = 0.5$; and vertical line, true value ($\beta_1 = 0$).

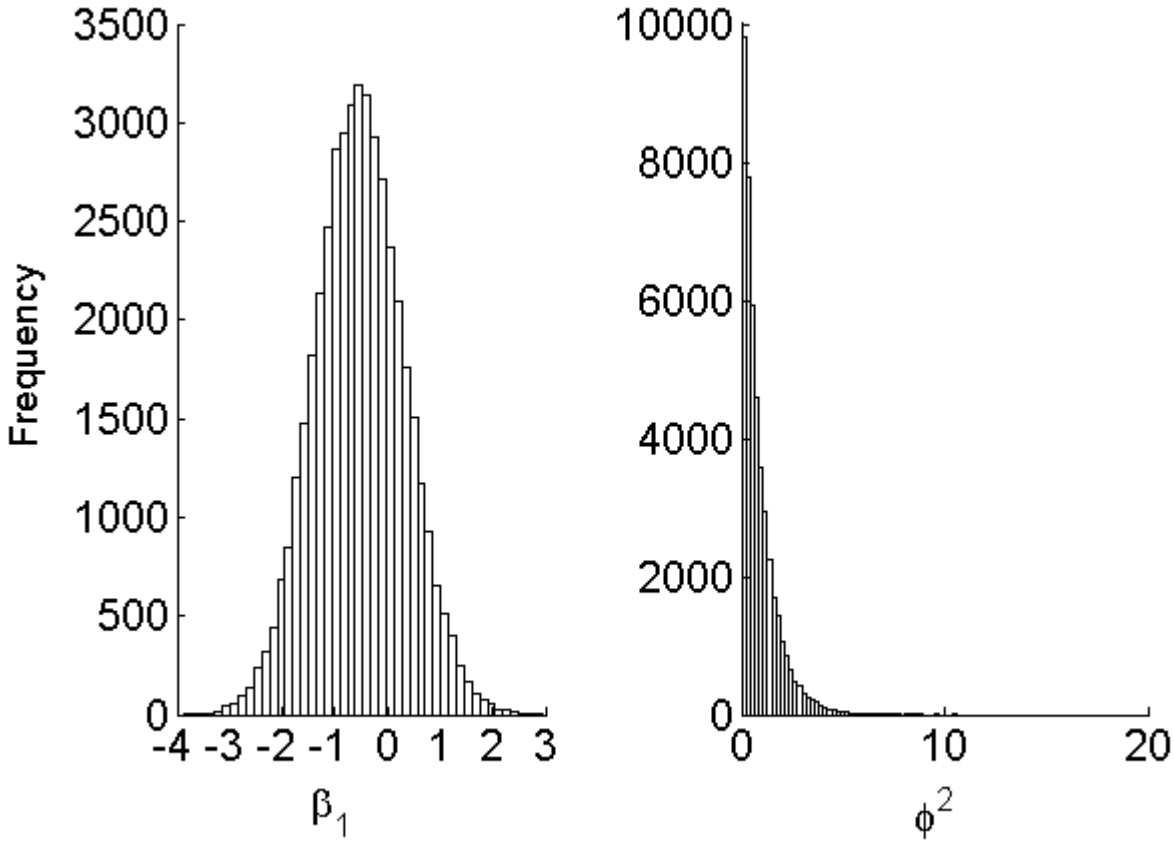
Frequency
3500
3000
2500
2000
1500
1000
500
0
-4 -3 -2 -1 0 1 2 3
β1
10000
8000
6000
4000
2000
0
0 10 20
ϕ2

Figure 2: Histogram of 45,000 draws from the posterior distributions of β_1 and ϕ^2 in model P2 ($\mu = 0$, $\alpha_1 = 1$ and $\alpha_2 = 1$).

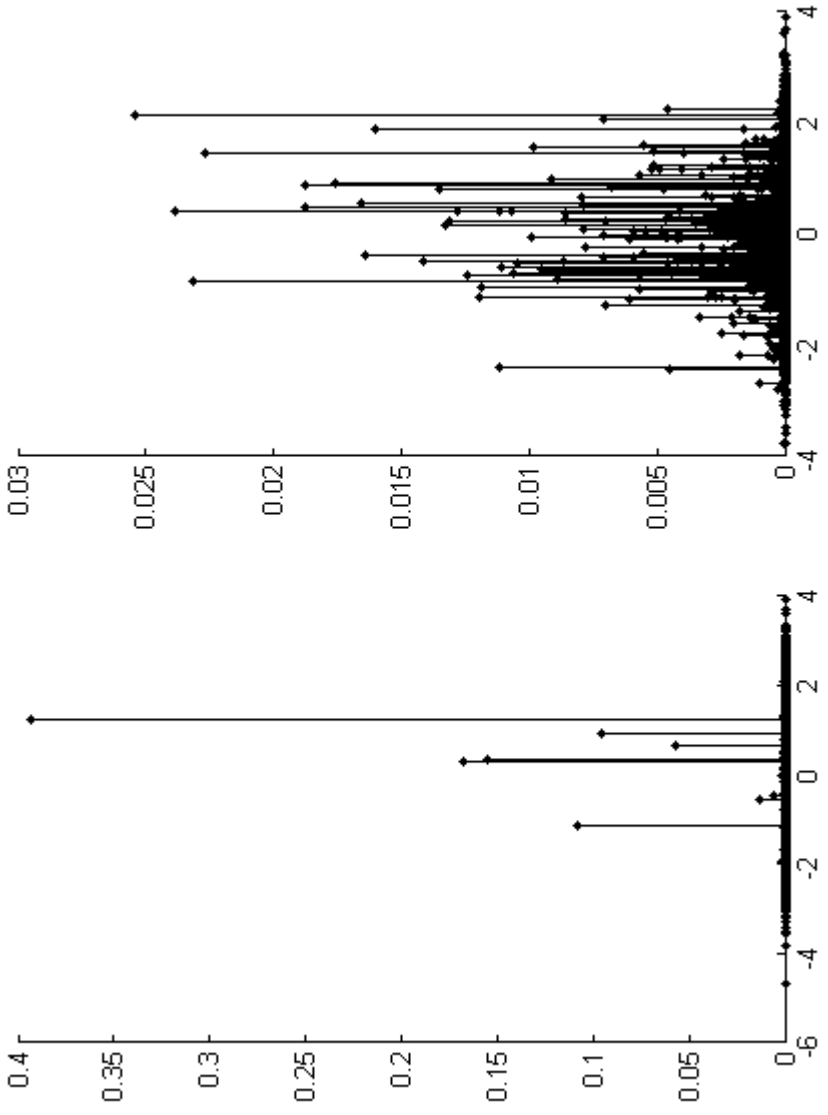
0.03
0.025
0.02
0.015
0.01
0.005
0
4
2
0
-2
-4
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
4
2
0
-2
-4
-6

Figure 3: Two simulations from $DPP(\lambda D_0)$ with $\lambda = 1$ (left) and $\lambda = 100$ (right). $D_0 \equiv N(0, 1)$.

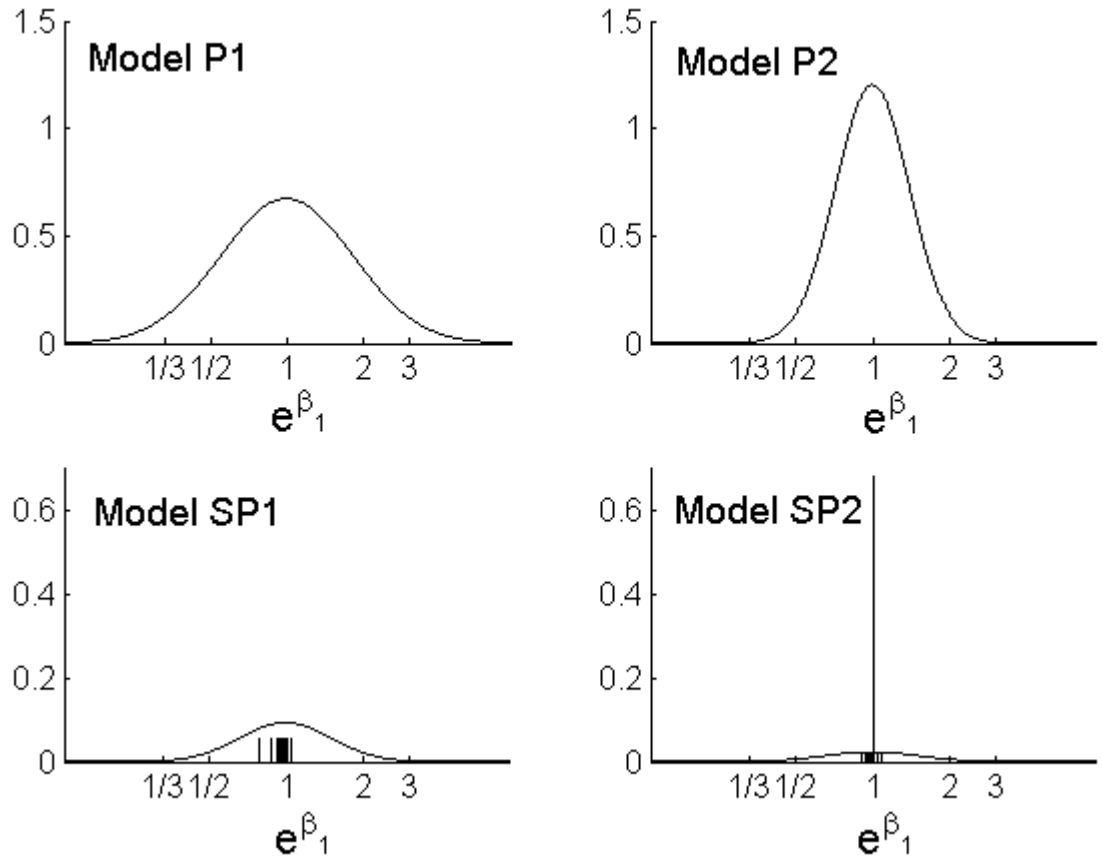
Model P1
Model P2
Model SP1
Model SP2
1.5
1
0.5
0
0.6
0.4
0.2
1/3 1/2 1 2 3
e^{β_1}

Figure 4: Prior distributions for the effect of imazethapyr, using the four hierarchical models applied to the Agricultural Health Study data, evaluated at the mean posterior of all other random variables.

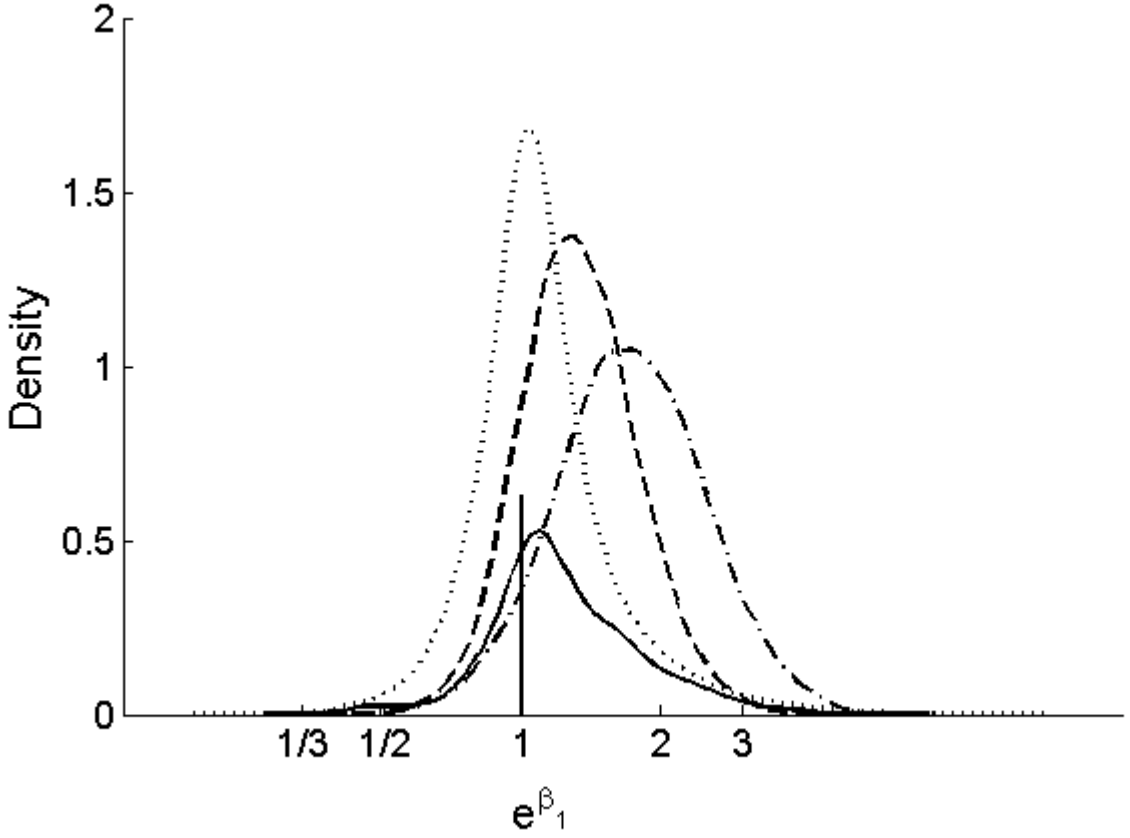
2
1.5
1
0.5
0
Density
1/3
1/2
1
2
3
e^{β_1}

Figure 5: Posterior distributions for the effect of imazethapyr, using the four hierarchical models applied to the Agricultural Health Study data. Solid line indicates SP2; dotted line, SP1; dashed line, P2; dash-dot line, P1.