Appendix I: Performance of Models in Simulated Datasets

To assess these models (P1, P2, SP1, and SP2) and their ability to estimate effects in a variety of scenarios, we examined their performance in simulated data. Data were simulated from the logistic model:

$$\mathrm{logitPr}(Y_i = 1\,|x_{i1}, \ldots, x_{i10}) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{10} x_{i10},$$

under a variety of scenarios for the true effect of $\beta_1 \ldots \beta_{10}$. Datasets were simulated under the assumption of no correlation between covariates and assuming a 90% pairwise correlation between all covariates. Datasets of 250 observations were simulated 500 times and analyzed using a maximum likelihood logistic model as well as a logistic regression with the priors specified in Table A1.1. Gibbs sampling algorithms to analyze each model were programmed and run in Matlab for 10,000 iterations. The initial 3000 iterations were discarded as a burn-in.

Figure A1.1 shows the mean squared error (MSE) of datasets simulated assuming all $\beta_j = 0$ (top row), and $\beta_1 = 0.05, \beta_2 = 0.1, \beta_3 = 0.15 \ldots \beta_{10} = 0.5$ (bottom row). When none of the coefficients have an effect the two semi-parametric models outperform the parametric hierarchical models (and the ML model). The decreased MSE of models SP1 and SP2 occurs because during most iterations of the Gibbs sampler, all coefficients are clustered together. Model SP2 has slightly lower MSE than model SP1 because it gives increased probability (π) to the true value of the β_j. When the data are orthogonal, little difference is observed between ML, P1 and P2 (i.e., the data tend to swamp out the prior knowledge); however, when the data are highly correlated model P1 has lower MSE than ML and model P2 has slightly lower MSE than model P1. The improved performance of the hierarchical models when data are highly correlated occurs because the high correlation implies less information is available to estimate the effects (similar to simply having less data in the orthogonal case), and prior knowledge is more important in estimating coefficients.

The bottom row of Figure A1 shows simulation results when the true effect of each coefficient is different. In these simulations, the true effects of the coefficients do not differ by much. Models SP1 and SP2 again have lower MSE than any of the other models, because the slight gain in bias (by occasionally assuming these coefficients have the same effect), is offset by the increase in precision. However, when the true effect is not zero Model SP2 tends to have slightly worse performance as the true effect moves further from the null. If the difference between the coefficients was larger, the semi-parametric models would perform somewhat worse than the two parametric models.

Figure A1.2 presents simulations in which only β_1 has an effect ($\beta_1 = 0.5$ in the top row and $\beta_1 = 1.0$ in the bottom row). When $\beta_1 = 0.5$, the models SP1 and SP2 perform somewhat worse than the other models in estimating β_1 (however they still perform better for the other 9 coefficients). The decreased performance results from the relatively small size of the effect, which the semi-parametric models are more likely to cluster with the other coefficients (which have no effect). When the effect is larger ($\beta_1 = 1.0$), the four hierarchical models perform similarly. With highly correlated data the semi-parametric models can perform better when the effect is small (because these models bias the coefficient by clustering it with the other coefficients, but substantially reduce the variability), but worse when the effect is larger (because the bias introduced through clustering becomes more severe).

Figure A1.3 presents simulations in which the coefficients fall into clusters having an effect of $\beta_1 \dots \beta_5 = 0.5$ or having no effect ($\beta_6 \dots \beta_1 0$). When the data are orthogonal, all estimation methods produce similar MSE, with a slight improvement of the two SP models. When data are highly correlated, as would be expected, the two semi-parametric models have lowest MSE. Model SP1 has lower MSE for $\beta_1 \dots \beta_5$ where the true effect is not null, while model SP2 has lower MSE for $\beta_6 \dots \beta_{10}$ where the true effect is null (since model SP2 gives additional prior probability to this hypothesis). Model P1 and P2 have roughly equivalent MSE in this setting.

Figure A1.4 presents simulations in which the first coefficient has no effect while $\beta_2 \dots \beta_{10} = 1.0$. When the predictors are not correlated (top panel), model SP2 routinely outperforms the other models since it is capable of clustering the last nine coefficients together while allowing the first (truly null) coefficient to fall into the zero cluster. Model SP1 performs well for coefficients $\beta_2 \ldots \beta_{10}$, but has an increased MSE for β_1. The decreased performance for this coefficient is due to model SP1 grouping β_1 with $\beta_2 \ldots \beta_{10}$ too often. Over the 500 simulated datasets, we observed false positive rates for β_1 of $5\%, 4\%, 4\%, 7\%, 2\%$ for the ML, P1, P2, SP1 and SP2 models, respectively. The bottom panel in figure A4 illustrates the effect of high correlation among the predictors. For $\beta_2 \ldots \beta_{10}$, model SP1 performs well generally clustering these coefficients to improve MSE. The 2 semi-parametric models perform poorly in estimating β_1. We observed false positive rates for β_1 of $7\%, 3\%, 2\%, 33\%, 28\%$ for the ML, P1, P2, SP1 and SP2 models, respectively. On the other hand, models SP1 and SP2 were far more likely to correctly flag any of the other 9 coefficients as significant (0.18, 0.17, 0.15, 0.60, 0.58 for ML, P1, P2, SP1, and SP2 respectively).

Figure A1.5 compares the performance of models P1, P2, SP1 and SP2 where $\beta_1 \ldots \beta_5 = 1.0$ and $\beta_6 \ldots \beta_{10} = -1.0$. Models P1 and P2 specify $\beta_1 \ldots \beta_5 \sim N(1.0, \phi_1^2)$ and $\beta_6 \ldots \beta_{10} \sim N(-1.0, \phi_2^2)$. For model P1, $\phi_1^2 = \phi_2^2 = 0.5$ while for model P2 these coefficients are random. The coefficients are given non-parametric specification in models SP1 and SP2, centered on a $N(0, .5)$ distribution. With no correlation, the models SP1 and SP2 are able to discover the clustered structure of the data and outperform models P1 and P2. With high correlation, however, models P1 and P2 perform far better than models SP1 and SP2. This set of simulations, while informative, is also potentially misleading. It compares two correctly specified models (models P1 and P2) with two incorrectly specified models (SP1 and SP2). The improved performance of models P1 and P2 is not surprising in that regard.

It is important to note that such simulations are inherently artificial and the relative performance of the models (assessed through MSE) could vary depending on the prior specification. For instance, in each of the simulations we present, the hierarchical models outperform the ML model. However, it is well known that this need not always be the case: if we had chosen a prior that was radically different from the true value of β_j, the ML model could have smaller MSE than any of the hierarchical models. Additionally, a squared error loss function is only one of many ways to assess the performance of these models and we make no guarantee that the performance of these models will be the same under different loss functions. Finally, it is important to note that these considerations are inherently frequentist. From a Bayesian perspective the expected loss is averaged over the prior distribution to find an estimator that yields the smallest Bayes risk. When a squared error loss function is used, that estimator is the posterior mean. This framework more naturally places the emphasis on specifying a prior distribution that most closely corresponds to prior knowledge.

Table A1.1. Hierarchical models used in analysis of simulated data in Figures A1.1-A1.5.

P1			P2		
β_j	$\sim$	$N(0,1)$	β_j	$\sim$	$N(0,\phi^2)$
			ϕ^2	$\sim$	$IG(3,2)$

SP1			SP2		
β_j	$\sim$	D	β_j	$\sim$	D
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	$\sim$	$IG(3,2)$	ϕ^2	$\sim$	$IG(3,2)$
			π	$\sim$	$beta(1,1)$

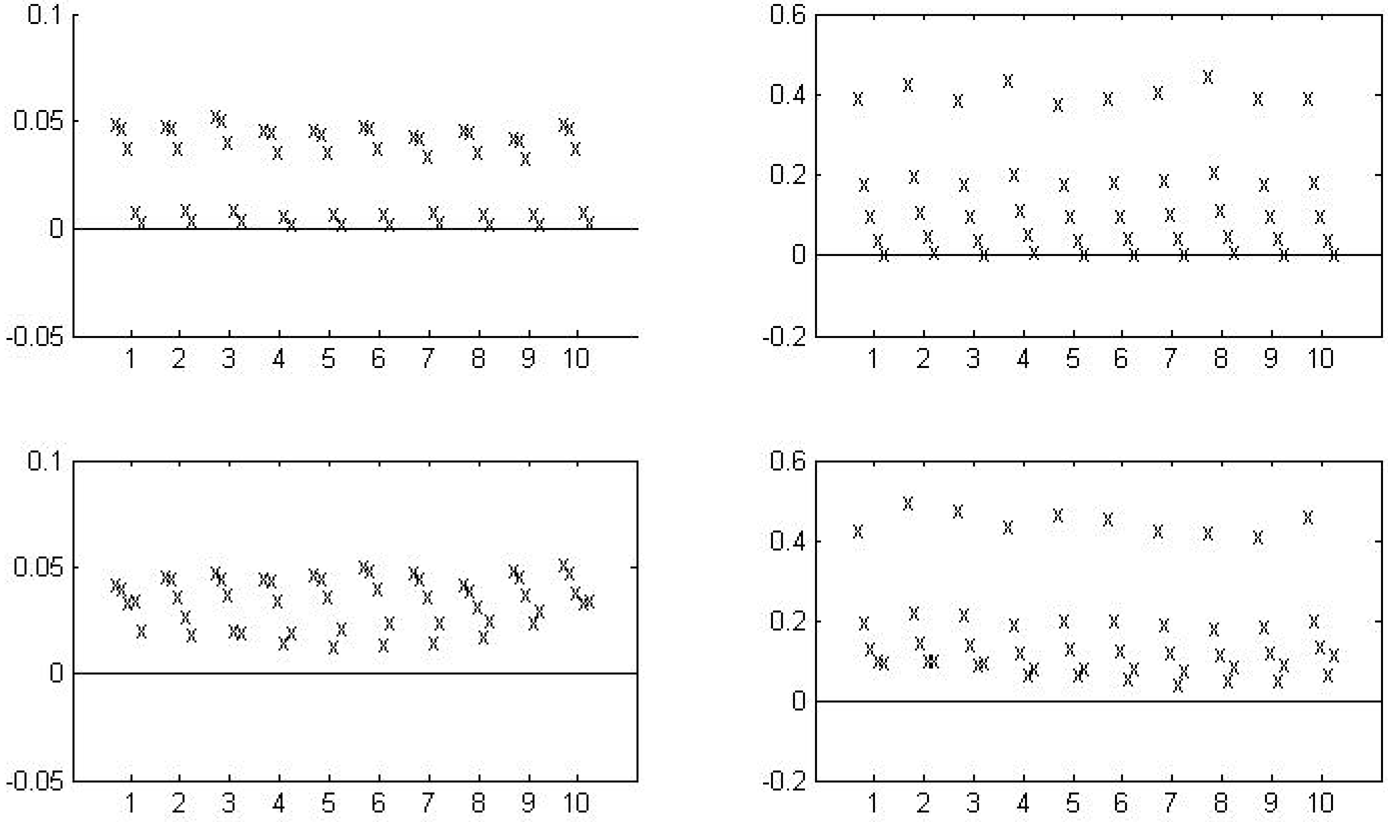

Figure A1.1: Mean squared error of parameter estimates ML, P1, P2, SP1 and SP2 models (in order from left to right, and grouped within coefficient). Top row: $\beta_1 \ldots \beta_{10} = 0$, all models are correctly specified. Bottom Row: $\beta_1 = 0.05, \beta_2 = 0.10 \ldots \beta_{10} = 0.50$, all models are incorrectly specified. Left Column: correlation=0, Right Column: correlation=0.9.

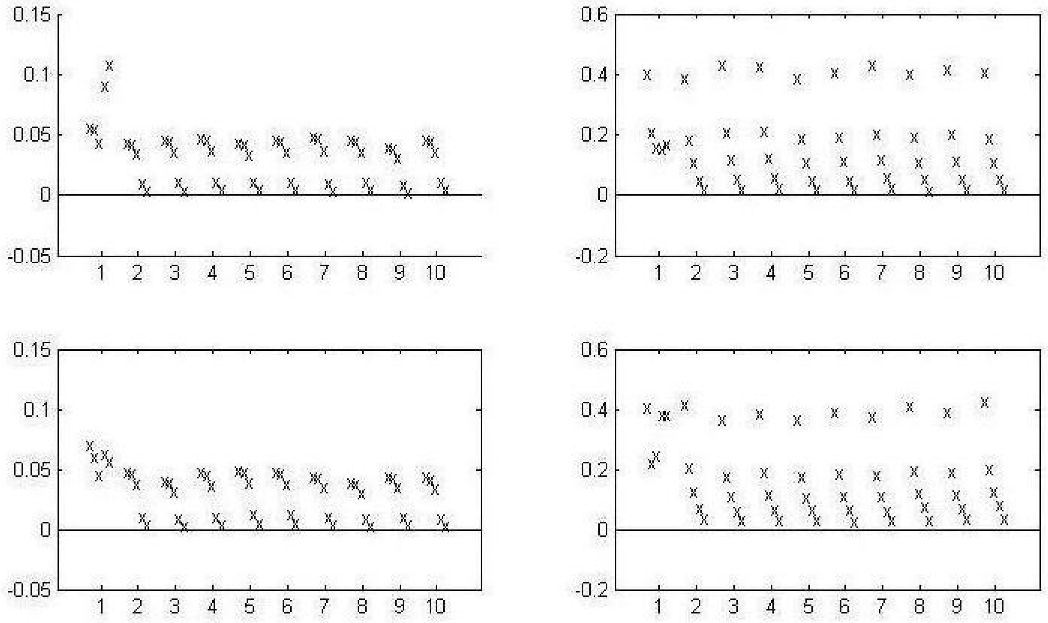

Figure A1.2: Mean squared error of parameter estimates ML, P1, P2, SP1 and SP2 models (in order from left to right, and grouped within coefficient). Top row: $\beta_1 = 0.5, \beta_2 \ldots \beta_{10} = 0$, all models incorrectly specified. Bottom Row: $\beta_1 = 1.0, \beta_2 \ldots \beta_{10} = 0$, all models incorrectly specified. Left Column: correlation=0, Right Column: correlation=0.9.

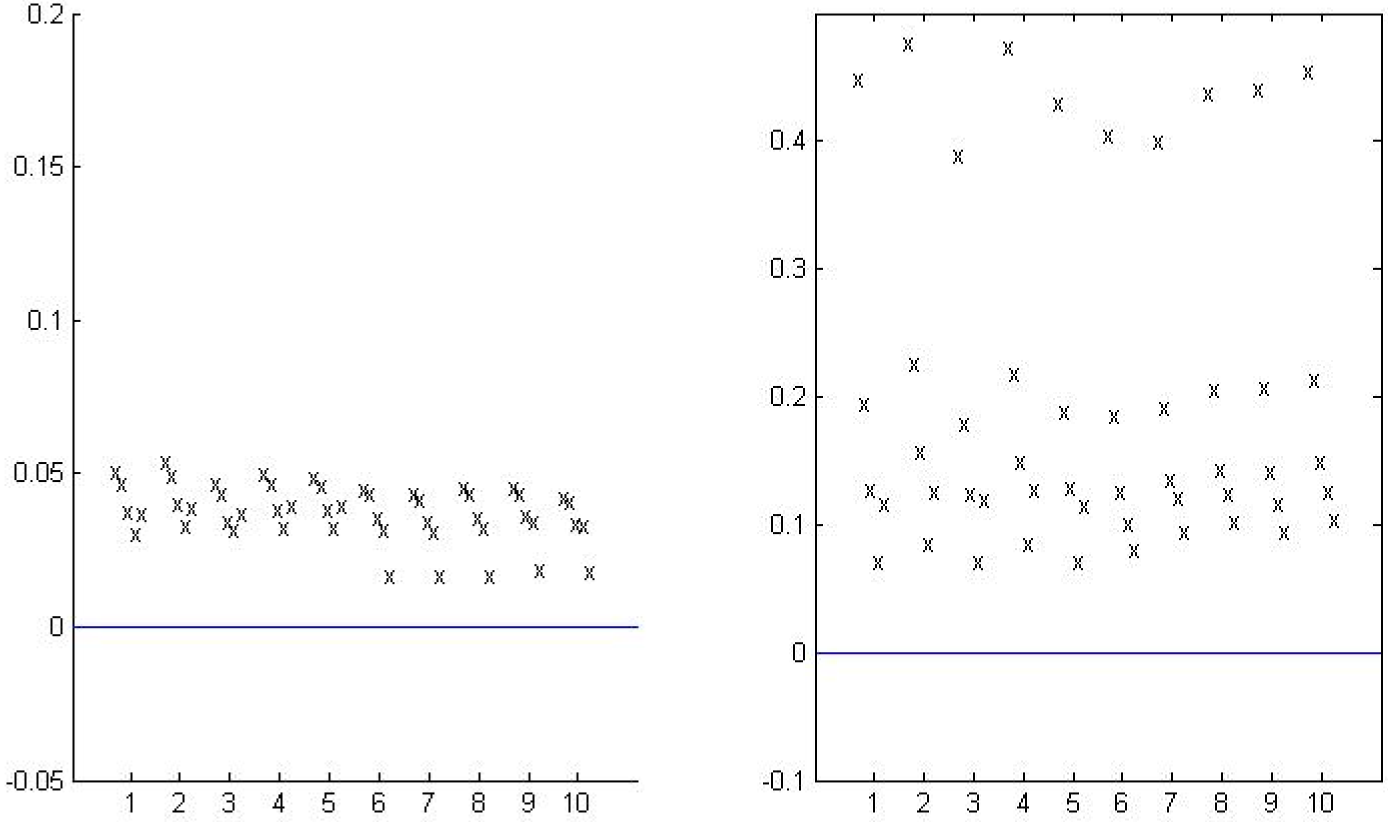

Figure A1.3: Mean squared error of parameter estimates ML, P1, P2, SP1 and SP2 models (in order from left to right, and grouped within coefficient). $\beta_1 \ldots \beta_5 = 0.5, \beta_6 \ldots \beta_{10} = 0.0$, all models incorrectly specified. Left Column: correlation=0, Right Column: correlation=0.9.

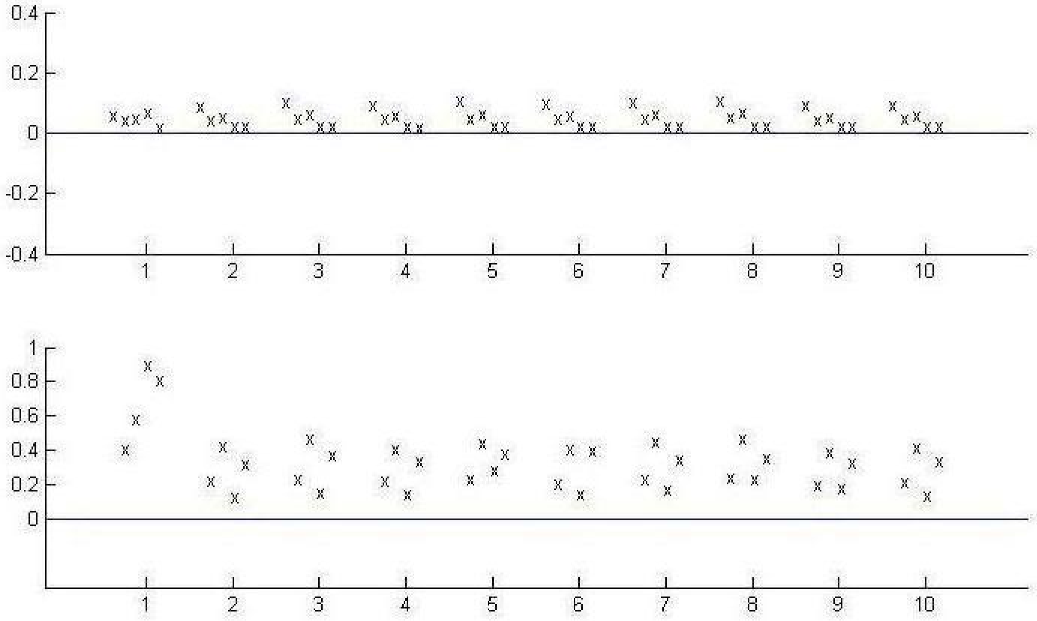

Figure A1.4: Mean squared error of parameter estimates (grouped within coefficient). $\beta_1 = 0, \beta_2 \ldots \beta_{10} = 1.0$. Top Row: correlation=0, Bottom Row: correlation=0.9.

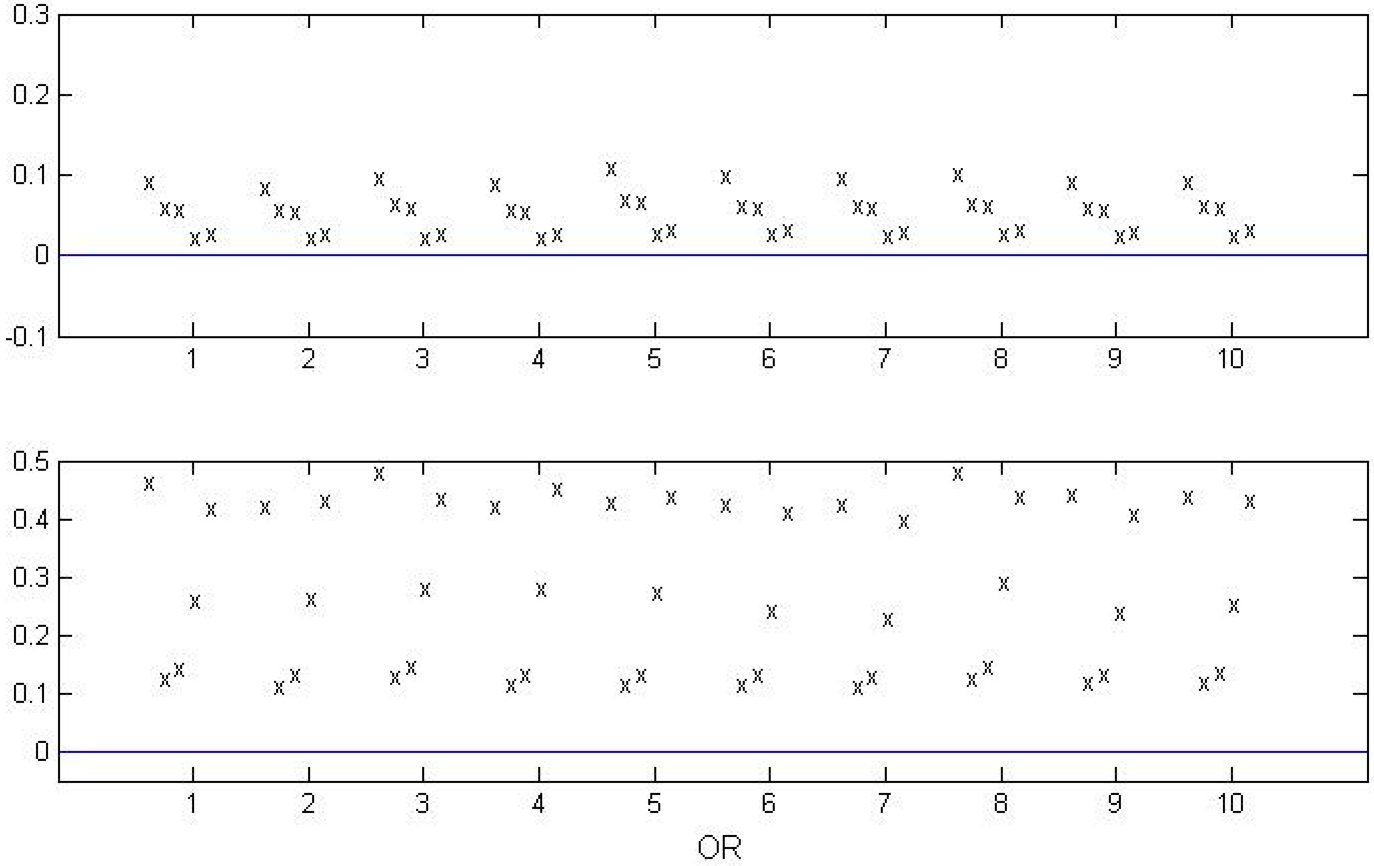

Figure A1.5: Mean squared error of parameter estimates (grouped within coefficient). $\beta_1 \ldots \beta_5 = 1.0, \beta_6 \ldots \beta_{10} = -1.0$. Models P1 and P2 correctly specify separate prior means for $\beta_1 \ldots \beta_5$ and $\beta_6 \ldots \beta_{10}$. Models SP1 and SP2 are incorrectly specified. Top row: correlation=0, Bottom row: correlation=0.9.

Appendix II: WinBUGS Code for Parametric Models

This appendix provides a generic template of WinBUGS code that can be used to implement models P1 and P2. We present code that can be used to analyze a hypothetical dataset with a binary outcome, y, and 7 dichotomous covariates $x_1 \dots x_7$. Information on how to read data into WinBUGS can be found in the WinBUGS manual. We use the following data:

```
list( x1=c(0,1,0,0,0,0,0,0), x2=c(0,0,1,0,0,0,0,0), x3=c(0,0,0,1,0,0,0,0),
x4=c(0,0,0,0,1,0,0,0), x5=c(0,0,0,0,0,1,0,0), x6=c(0,0,0,0,0,0,1,0), x7=c(0,0,0,0,0,0,0,1),
n = c(100,100,100,100,100,100,100,100), y = c(10, 8, 11, 12, 9, 13, 11, 14), N = 8, J=7)
```

The data are in aggregate form (i.e., there 100 people who are unexposed to $x_1 \dots x_7$ and 10 of them have the outcome. There are 100 people who are exposed to only x_1 and 8 of them have the outcome, etc). The following Winbugs code can be used to analyze this dataset using model P1:

Winbugs Code for Model P1

```
model {
for( i in 1 : N ) {
y[i] ∼ dbin(p[i],n[i])
logit(p[i]) ← alpha + b[1]*x1[i]+b[2]*x2[i]+b[3]*x3[i]
    + b[4]*x4[i]+b[5]*x5[i]+b[6]*x6[i]+b[7]*x7[i] }
for(j in 1:J) {
b[j] ∼ dnorm(0,.3) }
alpha ∼ dnorm(0.0,0.01) }
```

Note that dnorm(a,b) is a normal distribution with mean a and variance 1/b. We ran the SB model for 50,000 iterations and excluded the first 10,000 as a burn-in. The results from this model are given in the table below:

Coefficient	Posterior Mean	Standard Deviation
b[1]	-0.27	0.47
b[2]	0.08	0.43
b[3]	0.18	0.43
b[4]	-0.15	0.46
b[5]	0.27	0.42
b[6]	0.08	0.43
b[7]	0.36	0.41

WinBUGS Code for Model P2

Code for model P2, is only slightly more complex than code for model P1:

```
model {
for( i in 1 : N ) {
y[i] ~ dbin(p[i],n[i])
logit(p[i]) ← alpha + b[1]*x1[i]+b[2]*x2[i]+b[3]*x3[i]
    + b[4]*x4[i]+b[5]*x5[i]+b[6]*x6[i]+b[7]*x7[i] }
for(j in 1:J) {
b[j] ~ dnorm(0,phi) }
alpha ~ dnorm(0.0,0.01)
phi ~ dgamma(0.3,1)}
```

Because WinBUGS specifies the normal distribution in terms of precision (the inverse of variance), in model P2 a gamma prior is placed on precision parameter (which is equivalent to our earlier approach that placed an inverse gamma prior on the variance). In model P2, dgamma is a Gamma(α, β) distribution with mean= $\alpha\beta$ and variance=$\alpha\beta^2$. So our above specification gives a prior mean of 0.3 and prior variance of 0.3.

We ran the code for model P2 on the data above for 50,000 iterations of the Gibbs sampler, discarding the initial 10,000 as a burn-in. The results are given in the table below:

Coefficient	Posterior Mean	Standard Deviation
b[1]	-0.24	0.36
b[2]	0.03	0.34
b[3]	0.11	0.33
b[4]	-0.14	0.35
b[5]	0.19	0.33
b[6]	0.03	0.34
b[7]	0.27	0.33

Appendix III: Alternative analysis of Agricultural Health Study example

The Agricultural Health Study (AHS) enrolled farmers who applied for pesticide licenses in Iowa or North Carolina between 1993 and 1997 and has been described in detail elsewhere. Kirrane et al. recently examined the association between pesticide exposure and retinal degeneration among the wives of AHS farmers. Spouses of farmers filled out a questionnaire with information on their medical and pesticide use history. We analyzed the same data (31,173 women, 281 of whom experienced retinal degeneration) and controlled for the same covariates, but limit our analysis to the 6 fungicides. These chemicals exhibited a wide range of correlation up to 0.33. The literature on effects of fungicides on retinal degeneration is limited, amounting to one study which we used to inform our prior. Table A3.1 shows the 4 hierarchical models used for the analysis. Gibbs sampling algorithms were programmed in Matlab and run for 60,000 iterations with the initial 5,000 excluded as a burn-in period.

To illustrate the four hierarchical models, we present representations of the prior distributions for the effect, β_1, of benomyl in figure A3.1. Because the prior distributions for models P2, SP1 and SP2 depend on random variables we evaluate the them at the posterior mean of all other random variables (ϕ^2 for model P2; ϕ^2, λ, and $\beta_2 \dots \beta_6$ for model SP1; ϕ^2, λ, $\beta_2 \dots \beta_6$, and π for model SP2). The prior distribution for model P1 is determined by our prior belief that any fungicide has an effect on retinal degeneration of OR=1.8 and that we are 95% certain the effect lies between OR=0.8 and OR=4.0 ($\phi^2 = 0.16$). The prior distribution for β_1 in model P2 is more complicated since ϕ^2 is random. A larger than expected amount of variability is observed among the fungicide effects, leading to a posterior mean of $\phi^2 = 0.29$. Thus the prior distribution for β_1 evaluated at $\phi^2 = 0.29$ is less concentrated around the prior mean, which will lead to less shrinkage of effects toward OR=1.8. As indicated earlier, the prior distribution for model SP1 is a mixture of a normal distribution with a mean OR=1.8 and posterior estimate of $\phi^2 = 0.21$ and a set of point masses at the posterior estimates of $\beta_2 \ldots \beta_6$. The mean posterior value of $\lambda = 1.8$ (the data provide more evidence in favor of normally distributed effects than indicated by the prior), implies that with probability 26%, β_1 is distributed as $N(0.6, 0.21)$ and with probability 15%, β_1 is assigned the value of one of $\beta_2 \ldots \beta_6$. The prior distribution for model SP2 is similar to model SP1, except for a large point mass at 0. The posterior mean of $\pi = 0.40$ and $\lambda = 1.59$ imply that β_1 is distributed according to $N(.6, 0.16)$ with probability 14% or set equal to $\beta_2 \ldots \beta_9$ with probability 9% or set equal to 0 with probability 40%.

The results of the models are presented in Table A3.2. Figure A3.2 shows the posterior distributions of the effect of benomyl. Model P1 estimated a slightly elevated effect (OR=1.4, 95% CI (0.7, 2.6)). Because of the variability of the fungicide effects relative to their prior specification, model P2 had a larger posterior ϕ^2 than that of model P1 and therefore less shrinkage (OR=1.2, 95% CI (0.5, 2.8)) toward the prior mean than model P1. In model SP1, the coefficient for benomyl was the least likely of any effect to be clustered with any of the other coefficients (between 5% and 6% of the time). The relatively modest amount of clustering for benomyl resulted in posterior estimates similar the models P1 and P2. However, other coefficients (notably, Ziram) had greatly increased precision as a result of being clustered with the other effects as much as 20% of the time. The distribution of β_1 from model SP2 has a large spike at 0 which is the posterior probability that $\beta_1 = 0$ (p=0.57). The most likely non-null effect in model SP2 is very similar to the effects estimated by the other three models (OR=1.4). This indicates some uncertainty over whether benomyl is associated with retinal degeneration.

Table A3.1: Hierarchical models used to analyze Agricultural Health Study data on fungicides and retinal degeneration in wives of pesticide applicators, North Carolina and Iowa, 1993-1997.

P1			P2		
β_j	$\sim$	$N(0.6, 0.16)$	β_j	$\sim$	$N(0.6, \phi_1^2)$
			ϕ_1^2	$\sim$	$IG(2.03, 0.16)$

SP1			SP2		
β_j	$\sim$	D	β_j	$\sim$	D
D	$\sim$	$DP(\lambda D_0)$	D	$\sim$	$DP(\lambda D_0)$
D_0	$\equiv$	$N(0.45, \phi^2)$	D_0	$\equiv$	$\pi\delta_0 + (1-\pi)N(0.45, \phi^2)$
λ	$\sim$	$G(1, 1)$	λ	$\sim$	$G(1, 1)$
ϕ^2	$\sim$	$IG(2.03, 0.16)$	ϕ^2	$\sim$	$IG(2.03, 0.16)$
			π	$\sim$	$beta(0.2, 1.8)$

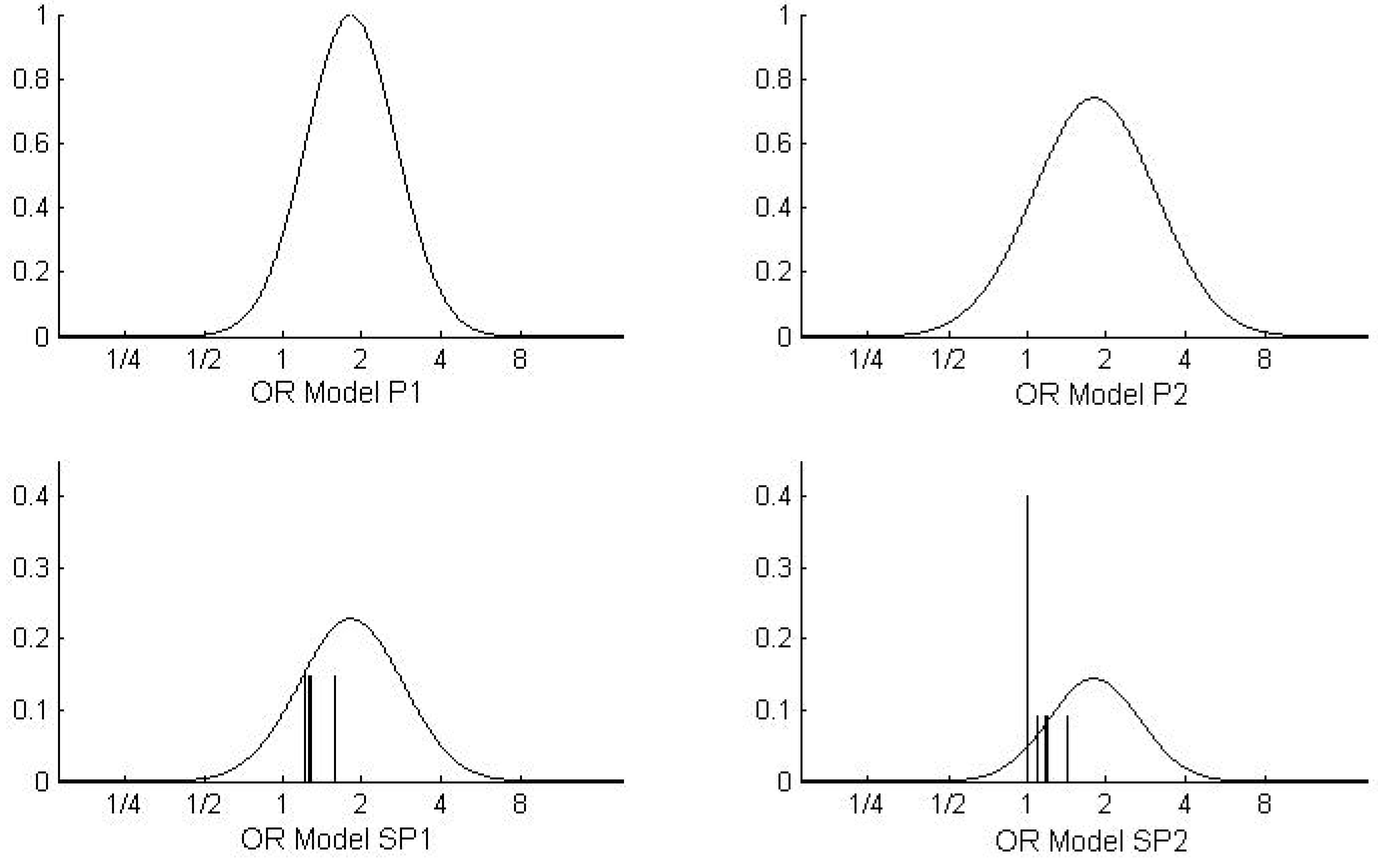

Figure A3.1: Prior distributions for the effect of benomyl using four hierarchical models used to analyze the Agricultural Health Study data, evaluated at the mean posterior of all other random variables.

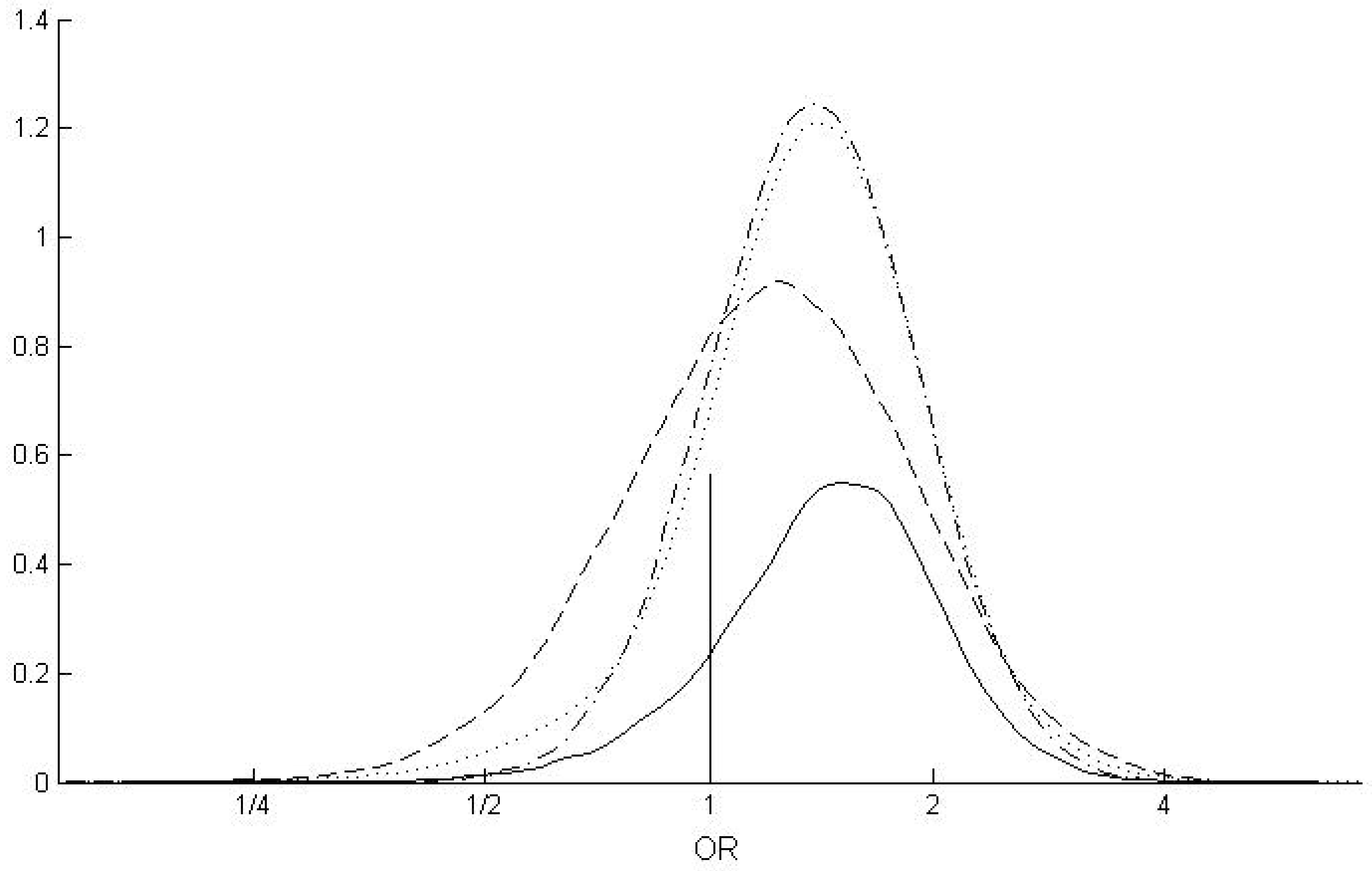

Figure A3.2: Posterior distributions for the effect of benomyl using four hierarchical models used to analyze the Agricultural Health Study data.

Figure A3.2: Posterior distributions for the effect of benomyl using four hierarchical models used to analyze the Agricultural Health Study data.

solid line: SP2

dotted line: SP1

dashed line: P2

dash-dot line: P1