Strategies for Model Mixing in Generalized Linear Models

MERLISE CLYDE

Institute of Statistics and Decision Sciences,

Box 90251 Duke University, Durham, NC 27708–0251.

1 Introduction

In linear regression models and generalized linear regression models (GLMs), there is often substantial uncertainty about the choice of covariates to include in the model. Both classical and Bayesian approaches that involve selecting a subset of covariates and making inferences conditional on that model choice ignore a major component of uncertainty in the problem. One approach for incorporating this form of model uncertainty into the analysis is by directly building into the model a vector of indicator variables γ that reflects which covariates are included in the model. If one assigns a prior distribution to the set of possible models, $\pi(\gamma)$, then Bayesian updating of the prior distribution leads to a posterior distribution given the data $\mathbf{Y} = (Y_1, \ldots Y_n)^T$ over the different models,

$$\pi(\gamma|\mathbf{Y}) = \frac{p(\mathbf{Y}|\gamma)\pi(\gamma)}{\sum_{\gamma'} p(\mathbf{Y}|\gamma')\pi(\gamma')} = \frac{q_\gamma}{\sum_{\gamma'} q_{\gamma'}} \tag{1}$$

where

$$p(\mathbf{Y}|\gamma) = \int p(\mathbf{Y}|\beta, \gamma)\, p(\beta|\gamma)\, d\beta \tag{2}$$

is the marginal distribution of the data $\mathbf{Y}$ given the model γ after integrating out model specific parameters β with respect to the prior distribution, $p(\beta|\gamma)$ on β. In many cases, posterior inference about various quantities of interest, such as predictive means, variances or predictive distributions, can be calculated from conditional expectations by a weighted average of the conditional model specific quantities,

$$\Delta = \sum_{\gamma} \Delta_\gamma \pi(\gamma|\mathbf{Y}) \tag{3}$$

see Leamer (1978, Ch 4), Raftery (1996), Raftery, Madigan and Volinski (1996), for examples. This Bayesian model averaging or model mixing provides a coherent method for making inferences and predictions that takes into account the uncertainty about the choice of covariates.

In practice, there are several difficulties in implementing this. One is that the integration required to obtain (2) may be analytically intractable, as is typically the case in GLMs. Proposals around this require using approximate methods of integration such as the Laplace method (Raftery (1993), Madigan, Raftery, and Volinsky (1996)) or Monte Carlo methods (George, McCulloch and Tsay 1994, Madigan and York 1993, Kuo and Mallick 1994, DeSimone 1996). Even if one can carry out the integration in (2), as in conjugate normal linear models (George and McCulloch, 1994; Clyde, DeSimone and Parmigiani, 1996; and Raftery,

Madigan and Hoeting, 1997), the number of models in the summation in (1) and (3) may be prohibitively large when there are many potential covariates, and it is often computationally infeasible to examine all possible models. Occam's window is a deterministic method of selecting a subset of models for use in model averaging (Madigan and Raftery 1994, Volinsky, Madigan, Raftery, and Krommal 1996). Markov Chain Monte Carlo (MCMC) samplers also provide a subset of models, where models are sampled based on their posterior probability. In addition to the previous references, Markov chain Monte Carlo methods for linear and generalized linear models have also been proposed by Carlin and Chib (1995), George and McCulloch (1993, 1994), Geweke (1994), Madigan and York (1993).

For normal linear regression models with an orthogonal design matrix, Clyde, DeSimone, and Parmigiani (1996) and Clyde, Parmigiani and Vidakovic (1995) developed efficient independent proposal distributions for sampling from large model spaces and for identifying models with high posterior probabilities. MCMC sampling in generalized linear models is computationally more intensive than in the conjugate linear regression framework, requiring samples from the posterior distribution on both the model space and parameter space. In problems where there are many potential covariates, it is important to develop efficient methods for exploring the model space and approximating the posterior model probabilities and other quantities of interest, based on a subset of models. As model uncertainty often dominates parameter uncertainty (Raftery, Madigan, and Volinski (1996), the focus here is on improved proposal distributions for the model space. In this article, we present several new approaches that extend the results of Clyde, DeSimone, and Parmigiani (1996) and Clyde, Parmigiani, and Vidakovic (1995) to generalized linear models, leading to improved convergence of MCMC methods.

2 Models

The response variables Y_i, $i = 1, \ldots, n$ are independent observations from an exponential family with canonical parameter θ_i of the form

$$f(y_i|\theta_i) = \exp\left\{\frac{w_i}{\phi}(y_i\theta_i - b(\theta_i)) + c(y, w, \phi)\right\}$$

for specific known functions $b()$, and $c()$ (see McCullagh and Nelder (1989)). We will assume that the positive weight w_i and scale parameter ϕ are both known. The mean of Y_i is

$$E(Y_i) = \mu_i = b'(\theta_i)$$

and the variance expressed as a function of μ_i is given by,

$$V(\mu_i) = b''(b'^{-1}(\mu_i))\frac{\phi}{w_i}.$$

The covariates are incorporated into the model through the linear predictor η_i and related to the mean μ_i via the link function g, where $\eta_i = g(\mu_i)$. Covariates are represented by the $n \times p$ design matrix $\mathbf{X}$. To

build in uncertainty about which covariates are included in the linear predictor η_i, a $p \times 1$ vector γ will be used to represent different models where γ is a sequence of binary random variables, each indicating whether the corresponding column of $\mathbf{X}$ is included in the model. If $\mathbf{\Gamma}$ is a $p \times p$ matrix with the elements of γ on the diagonal and zeroes elsewhere, then the $n \times 1$ vector of linear predictors under model γ is represented as

$$\eta_\gamma = \mathbf{X}\mathbf{\Gamma}\boldsymbol{\beta}.$$

To complete the specification of the model, we need to assign prior distributions to $\boldsymbol{\beta}$ and γ. Let $\pi(\gamma)$ denote the prior distribution on the model space. One approach is to assume the γ_j's are independent Bernoulli random variables,

$$\pi(\gamma) = \prod_{j=1}^{p} \pi(\gamma_j) \equiv \prod_{j=1}^{p} \rho_j^{\gamma_j}(1-\rho_j)^{1-\gamma_j}, \tag{4}$$

where ρ_j is the prior probability of covariate j being included. Let $p(\boldsymbol{\beta})$ denote the prior distributions for $\boldsymbol{\beta}$. We will take the prior distribution on $\boldsymbol{\beta}$ to be a $N(\mathbf{0}, \mathbf{C})$, where $\mathbf{C}$ is a diagonal matrix with the elements c_j^2 on the diagonal, as in Raftery (1996). The data augmentation prior distributions discussed by Bedrick, Christianson, and Johnson (1996) are also a natural choice and can be used to induce dependence in the distribution of $\boldsymbol{\beta}$. In their approach, the prior distribution has the same form as the likelihood, so that the prior information can be easily incorporated as additional observations.

3 Posterior Model Probabilities in Normal Models

The joint posterior distribution for $\boldsymbol{\beta}$ and γ can be represented as $p(\boldsymbol{\beta}|\mathbf{Y}, \gamma)\pi(\gamma|\mathbf{Y})$. However, these are usually known only up to the normalizing constants, so posterior inferences are made based on samples drawn from the posterior distribution via MCMC methods. MCMC methods are usually straightforward to implement in this case, however, when there are potentially many covariates, convergence is a real issue as the number of models in the model space often exceeds the number of samples that one is willing to entertain when running the MCMC algorithm.

In linear models with normal errors and conjugate prior distributions, one can integrate out the parameters in the model, but in moderate size problems one cannot enumerate all models to find the sum in the denominator in (1). With the previous prior specification and the additional assumptions that the design matrix $\mathbf{X}$ has orthogonal columns ($\mathbf{X}'\mathbf{X}$ is a diagonal matrix) and the error variance σ^2 is known, the posterior distribution of γ exists in closed form and factors analytically into a product of Bernoulli distributions. Let $\mathbf{x}_j$ denote the j-th column of $\mathbf{X}$. Under these assumptions, conjugate updating and straightforward

manipulations lead to the following expression for $\pi(\gamma|\mathbf{Y},\sigma^2)$

$$\begin{aligned}\pi(\gamma|\sigma,\mathbf{Y}) &= \prod_j^p p_j^{\gamma_j}(1-p_j)^{(1-\gamma_j)} \qquad (5)\\ p_j &= \frac{a_j(\mathbf{Y},\sigma)}{1+a_j(\mathbf{Y},\sigma)}\\ a_j(\mathbf{Y},\sigma) &= \left(\frac{x_j^T x_j+\sigma^2/c_j^2}{\sigma^2/c_j^2}\right)^{-1/2}\left(\frac{\rho_j}{1-\rho_j}\right)\exp\left\{\frac{1}{2}\frac{(\mathbf{x}_j^T\mathbf{Y})^2}{\mathbf{x}_j^T\mathbf{x}_j+\sigma^2/c_j^2}\right\}.\end{aligned}$$

If good estimates of σ^2 are available, these can be plugged in to the posterior model probabilities with very little loss of efficiency (Clyde, Parmigiani, and Vidakovic 1995). In what follows we will assume that the columns of $\mathbf{X}$ are orthogonal. This is often the case in designed experiments with balanced designs and contingency tables without missing cells. In the case of an arbitrary design matrix, one can reparameterize the problem so that there is an orthogonal basis for the linear predictor, and the prior distribution on $\boldsymbol{\beta}$ has a diagonal covariance matrix (Clyde and Parmigiani 1996). This approach may be advantageous when one is interested in prediction problems, as opposed to variable selection, and can lead to a reparameterization that improves the rate of convergence of the Markov chain (Clyde, DeSimone, and Parmigiani, 1996).

The distribution in (5) leads to a simple, and very efficient algorithm for sampling models. In many MCMC implementations, the proposal distribution for picking a new γ involves changing only one component at a time. The block proposal distribution in Clyde, Parmigiani, and Vidakovic (1995) allows for all components to change independent of the current model state. We will now show how these results can be used to construct computationally simple, yet efficient approximations to model probabilities in generalized linear models.

4 Approximate Model Probabilities for GLMs

The independent posterior model probabilities in (5) are calculated under the assumptions of 1) normality, 2) linearity, and 3) known constant variance. We will look at transformations of GLM data that improve these assumptions, and then apply (5) to the transformed data. Variance stabilizing transformations are used to achieve a constant known variance, and transformations that result in zero asymptotic skewness (approximate symmetry) or a quadratic log likelihood are considered as means to improve the normality assumption. Using (5) with the transformed data, the model probabilities under this approximate model are easily calculated and can serve as an independent proposal distribution for the model variables γ in importance sampling or MCMC. Note, we are using the these transformation only as a means to find a good proposal distribution, since the posterior distribution from the GLM is used in the acceptance step of the MCMC or used to reweight draws from the importance sampler. In section 5, we will examine an alternative approach for approximating the models probabilities by fitting a model of independence directly to the model space.

4.1 Variance Stabilizing Transformations

The first approach uses a variance stabilizing transformation, $h()$ applied to the data $\mathbf{Y}$ to achieve the assumption of constant variance for (5). For many exponential families there exists a variance stabilizing transformation h, so that the variance of the transformed response is approximately constant. Additionally, if Y_i is approximately normally distributed $N(\mu_i, V(\mu_i))$, then $h(Y_i)$ is approximately normally distributed $N(h(\mu_i), V(\mu_i)h'(\mu_i)^2)$, provided h is differentiable and $h'(\mu_i)$ is not zero. The approximate variance of $h(Y_i)$ does not depend on μ_i if $h'(\mu_i)^2 V(\mu_i) = k_i$, so the variance stabilizing transformation can be obtained as the solution to the differential equation,

$$h(\mu) = \int k^{1/2} V(\mu)^{-1/2} d\mu.$$

Bartlett (1936) found that for the Poisson distribution, h is the square root and that $k = 1/4$. For binomial proportions, the variance stabilizing transformation is the $\arcsin(\sqrt{Y_i})$ or equivalently, $-\arcsin(1 - 2Y_i)$ and $k_i = n_i/4$. Other cases are the natural exponential families with quadratic variance (Morris 1983) $V(\mu) = a_0 + a_1\mu + a_2\mu^2$, with a_0, a_1, and a_2 known. This includes one-parameter versions of the normal and gamma, negative binomial, and generalized hyperbolic secant, in addition to the binomial and Poisson. The variance stabilizing transformation, $h(\mu)$, is given by the following expressions:

$$\begin{aligned}
&\frac{k^{1/2}}{\sqrt{a_2}} \log(2\sqrt{a_2 V(\mu)} + 2a_2\mu + a_1) && \text{for } a_2 > 0 \\
&\frac{k^{1/2}}{\sqrt{a_2}} \sinh^{-1} \frac{2a_2\mu + a_1}{\sqrt{\Delta}} && \text{for } a_2 > 0, \Delta = 4a_2a_0 - a_1^2 > 0 \\
&\frac{k^{1/2}}{\sqrt{a_2}} \log(2a_2\mu + a_1) && \text{for } a_2 > 0, \Delta = 4a_2a_0 - a_1^2 = 0 \\
&-\frac{k^{1/2}}{\sqrt{-a_2}} \arcsin \frac{2a_2\mu + a_1}{\sqrt{-\Delta}} && \text{for } a_2 < 0, \Delta = 4a_2a_0 - a_1^2 < 0 \\
&\frac{2k^{1/2}}{a_1}\sqrt{a_0 + a_1\mu} && \text{for } a_2 = 0, a_0 + a_1\mu > 0
\end{aligned}$$

If we apply a variance stabilizing transformation h to Y, then

$$E(h(\mathbf{Y})) \approx h(\boldsymbol{\mu}) = h(g^{-1}(\mathbf{X\Gamma\beta})) \tag{6}$$

which is a nonlinear model in terms of the linear predictor $\boldsymbol{\eta}_\gamma$, depending on the specific link function and variance stabilizing function. In order to use the normal linear model methods in (5) to approximate the model probabilities, we can replace $h(g^{-1}(X\Gamma\beta))$ by the first two terms of the Taylor's series expansion about a point η_{0i},

$$h(g^{-1}(\eta_i)) \approx h(g^{-1}(\eta_{0i})) + \frac{h'(g^{-1}(\eta_{0i}))}{g'(g^{-1}(\eta_{0i}))}(\eta_i - \eta_{0i}). \tag{7}$$

Substituting the approximate linear regression (7) for the mean and rearranging terms, we have the required

normal regression for the transformed variable $\mathbf{W} = (w_1, \ldots, w_n)^T$ with a known diagonal variance Σ, where

$$\mathbf{W} \stackrel{\cdot}{\sim} N(\mathbf{X}\Gamma\beta, \Sigma) \tag{8}$$

$$w_i = \frac{g'(g^{-1}(\eta_{0i}))}{h'(g^{-1}(\eta_{0i}))}\left(h(Y_i) - h(g^{-1}(\eta_{0i}))\right) + \eta_{0i} \tag{9}$$

$$\Sigma_{ii} = k_i\left(\frac{g'(g^{-1}(\eta_{0i}))}{h'(g^{-1}(\eta_{0i}))}\right)^2. \tag{10}$$

When $\Sigma_{ii} = \sigma^2$ for $i = 1, \ldots, n$, we can use probabilities obtained from (5) to obtain an approximation to $\pi(\gamma|\mathbf{Y})$ for the GLM.

In the approximate normal problem, we can integrate out the other parameters β to obtain $\pi(\gamma|\mathbf{Y})$, which was not possible in the original GLM formulation of the problem. We are not necessarily suggesting to use the normal approximation for analysis of the model, but merely to provide a tractable approximation to the model probabilities that will allow us to easily identify high probability models. In a MCMC sampler or importance sampling, the models γ identified by the approximation can then be used to generate samples β given $\mathbf{Y}$ and γ based on standard proposal methods for the β's (George, McCulloch, and Tsay 1994, Madigan and York 1993, Madigan, Raftery, and Volinsky 1996, Gamerman 1994, West 1985).

4.2 Other Transformations for Normality

As the derivation of the approximate model probabilities relies on approximate normality, it is natural to look for other transformations such that the distribution of $h(\mathbf{Y})$ is "close" to a normal distribution. Hougaard (1982) discussed various transformations in one parameter exponential families where the transformations are obtained as the solution to

$$\left(\frac{\phi}{w_i}\right)^{\delta} \int \left\{\frac{d^2}{d\theta^2} b(\theta)\right\}^{\delta} d\theta,$$

and δ is a constant that determines various properties of the reparameterization. For example, $\delta = 0$ corresponds to the canonical parameterization, $\delta = 1/3$ corresponds to a quadratic loglikelihood parameterization (the third derivative of the log likelihood vanishes), $\delta = 1/2$ is the variance stabilizing transformation, $\delta = 2/3$ results in approximate zero skewness (symmetry), and $\delta = 1$ is the mean value parameterization. Under a change of variables we can re-express the results in terms of μ,

$$h(\mu) = \frac{\phi}{w_i} \int V(\mu)^{\delta - 1} d\mu. \tag{11}$$

Under the transformation $h(\mathbf{Y})$, the approximate mean is $h(\mu_i)$ and approximate variance is $V(\mu_i)^{2\delta-1}$.

If $h(Y_i)$ is approximately normal, then the above mean and variance determine the distribution. Again, it is necessary that the approximate mean is linear in $\mathbf{\Gamma}\beta$ so we will use a Taylor's series expansion as in (7) about a fixed point η_{0i}. The variable $\mathbf{W}$ is defined as before in (9) and is approximately normal with mean $\mathbf{X}\mathbf{\Gamma}\beta$, but Σ is diagonal with elements,

$$\Sigma_{ii} = \left(\frac{g'(g^{-1}(\eta_{0i}))}{h'(g^{-1}(\eta_{0i}))}\right)^2 V(g^{-1}(\eta_i))^{2\delta-1}. \tag{12}$$

The approximate variance Σ varies with η_i, which is unknown, so unless we substitute a fixed value for all cases, the normal model probabilities do not apply, as the nonconstant variance destroys the orthogonality required for the factorization of the model probabilities. Possibilities include: replacing Σ_{ii} by σ^2 equal to the median of Σ_{ii}, or the mean Σ_{ii} under the full model estimate of η, or evaluating Σ_{ii} at η_0, the point where the Taylor's series was evaluated. This latter method has been satisfactory in practice. The approximations in the variance may be outweighed, if the transformation results in a distribution "closer" to normality. While the least squares estimates obtained ignoring the nonconstant variance may be inefficient, the rescaled model probabilities may still be of the right magnitude.

4.3 Transformations for Linearity

While the above transformations may improve the normal approximations, they typically result in an approximate mean that is nonlinear in $\mathbf{\Gamma\beta}$. As linearity is also a key assumption of the normal model probability approximation, it may be more important to have a transformation that results in this directly. The expected value of $\mathbf{Y}$ is $g^{-1}(\eta_\gamma)$. Thus if we use the link function to define a transformation of $\mathbf{Y}$, then $E(g(\mathbf{Y}|\gamma))$ is approximately $\mathbf{X\Gamma\beta}$, and no additional expansions are needed of the mean vector in the transformed coordinates. In the case that the link function is the same as the variance stabilizing transformation, then the variance of $g(\mathbf{Y})$ is approximately constant. Otherwise, we must replace Σ_{ii}, as in (12), by a common value σ^2.

5 Estimation of Model Probabilities via Loglinear Models

The previous section relied on transformations of the data and model so that model probabilities from an appropriate normal model could be used to approximate the model probabilities under the generalized linear model. In this part we directly approximate the model probabilities via a log linear model representation for $\pi(\gamma|y)$.

One can view each model γ as a cell in a 2^p contingency table that represents the model space. The probability of being in the cell determined by γ is $\pi(\gamma|\mathbf{Y})$. We can use a saturated loglinear model to represent the model probabilities expressed as a function of the model vector γ:

$$\begin{aligned}\log(\pi(\gamma|\mathbf{Y})) &= \log(q_\gamma) - \log(\sum_{\gamma'} q_{\gamma'}) \\ &= \alpha_0 + \sum_j \alpha_j\gamma_j + \sum_{j,k} \alpha_{jk}\gamma_j\gamma_k + \sum_{j,k,l} \alpha_{jkl}\gamma_j\gamma_k\gamma_l + \ldots \alpha_{1\ldots p}\prod_{j=1}^{p}\gamma_j. \qquad (13)\end{aligned}$$

where the vector $\boldsymbol{\alpha} = (\alpha_0, \ldots \alpha_{1\ldots p})^T$ is a function of the data $\mathbf{Y}$.

In the normal linear regression model of section 3 with known σ,

$$\log(\pi(\gamma|\mathbf{Y})) = \sum_j (1 + a_j(\mathbf{Y}, \sigma))^{-1} + \sum_j a_j(\mathbf{Y}, \sigma)\gamma_j$$

where $\alpha_j = a_j(\mathbf{Y}, \sigma)$ for $j = 1, \ldots p$. This corresponds to a model of independence for the model space contingency table, where all the two-way (α_{jk}) and higher order interaction terms are zero. As our goal is to keep the proposal distribution in this form in the GLM framework, this suggests that another approach to obtain an approximation to the posterior model probability is to fit a model of independence to the table using estimated α_j's. Of course, we cannot estimate this from all models γ (that is the problem we started with). However, based on a sample of models, we can fit a model of independence to the model space contingency table. The second problem is that the exact posterior model probabilities of even this sample of models are unknown. We can instead use the Laplace estimates of the model probabilities in the sample for estimating $\boldsymbol{\alpha}$. Model probabilities can be accurately calculated using the Laplace approximation for integrals (Raftery 1996),

$$\begin{aligned} \pi(\gamma|\boldsymbol{y}) &\propto q_\gamma = \pi(\gamma) \int p(\mathbf{y}|\boldsymbol{\beta}\gamma) p(\boldsymbol{\beta}|\gamma)\, d\boldsymbol{\beta} \\ &\approx \pi(\gamma)\, (2\pi)^{p/2} |\psi_\gamma|^{-1/2} p(\mathbf{Y}|\tilde{\boldsymbol{\beta}}_\gamma, \gamma)\, p(\tilde{\boldsymbol{\beta}}_\gamma|\gamma) \end{aligned} \tag{14}$$

where $\tilde{\beta}_\gamma$ is the posterior mode given model γ and ψ_γ is the negative Hessian of the log posterior with elements

$$[\psi_\gamma]_{ij} = -\frac{\partial^2}{\partial\beta_i\, \partial\beta_j} \log(p(\mathbf{Y}|\boldsymbol{\beta}, \gamma) p(\boldsymbol{\beta}|\gamma)).$$

Raftery (1996) demonstrates empirically that (14) is accurate in generalized linear models. The difficulty of using (14) is that it requires finding the posterior mode $\tilde{\boldsymbol{\beta}}_\gamma$ and we do not know a priori which subset of models to examine. Volinsky et al.(1996) use a leap and bounds algorithm to search through models, and then estimate the Laplace probabilities for these models, but this implementation is currently limited to about 30 variables, which is smaller than problems we are interested in. By fitting this model of independence to at least $p+1$ models, we will have an alternative method to the leaps and bounds algorithm for identifying the high probability models.

We will determine q_γ approximately using (14) for a subset of l models, $l > p$. Let $\mathbf{Q}$ denote the vector of the Laplace approximations for these l models, and let $\mathbf{U}$ denote the $l \times p$ "design" matrix based on the l models, where the rows of $\mathbf{U}$ are the corresponding vectors γ. The loglinear model can be represented as

$$\log(\mathbf{Q}) = \mathbf{U}\boldsymbol{\alpha} + \mathbf{e} \tag{15}$$

where $\mathbf{e}$ represents an error term due to higher order terms not included in the model of independence. The least squares estimates of $\boldsymbol{\alpha}$ can be transformed to obtain estimates

$$p_j = \frac{\exp(\hat{\alpha}_j)}{1 + \exp(\hat{\alpha}_j)}$$

for estimating the model probabilities in (5). Note, there is nothing in this derivation that requires approximate normality or that the design matrix $\mathbf{X}$ for the observed data have orthogonal columns, thus we can actually use this method to approximate model probabilities in more general situations.

An important question is the choice of models to use in the design matrix $\mathbf{U}$. One approach is to use ideas from experimental design for 2^k-fractional factorial designs to choose the matrix $\mathbf{U}$. However in practice, we

have found that we can achieve better estimates by taking the best model under the transformed approach in section 4, and choosing the remaining p models by individually switching each γ_j to $1-\gamma_j$ for $j = 1, \ldots p$. This automatically ensures that each α_j is estimable. The Laplace approximations are then calculated for each of these models. This also allows one to check the agreement between the Laplace approximation and the independence approximations via the transformed variables.

6 Comparisons

We compare the different methods using a loglinear model example from Healy (1988, page 97) on deaths from tetanus. The problem is small enough, a 2^3 contingency table with factors Mortality (M), Severity of Tetanus (S), Antitoxin Indicator (S), so that we can calculate all 256 posterior model probabilities.

We will use a Poisson distribution for the cell counts. For Poisson regression with the canonical log link, $\boldsymbol{\mu} = \exp(\boldsymbol{\eta}_{\boldsymbol{\gamma}})$, so the link function $g(\cdot) = \log(\cdot)$. The variance stabilizing transformation is the square root transformation and the transformed variable $\mathbf{W}$ is

$$\mathbf{W} = \frac{2}{\sqrt{\bar{\mathbf{Y}}}}\left(\mathbf{Y}^{1/2} - \bar{\mathbf{Y}}^{1/2}\right) + \log(\bar{\mathbf{Y}}),$$

where we use a Taylor's series expansion about the point $\eta_{0i} = \log(\bar{\mathbf{Y}})$. For the Poisson model, the other normal transformations are given by $h(Y_i) = Y_i^{\delta}/\delta$ with an approximate mean of μ_i^{δ}/δ and approximate variance of $\mu_i^{(2\delta-1)}$ for $\delta > 0$ and is the log transformation for $\delta = 0$. We will evaluate Σ_{ii} at $\eta_i = \eta_{0i} = \log(\bar{\mathbf{Y}})$.

Figure 1 shows the log of the approximate model probabilities estimated using the methods from sections 4 and 5 for Poisson counts compared to the log of the model probabilities estimated using the Laplace approximation (Raftery 1996). The Kullback-Leibler divergence between the approximate posterior distributions of γ and the Laplace estimates are also calculated. All methods seem to provide close agreement, with the best agreement obtained by the transformation for symmetry. Even though the counts in the data ranged from 4 up to 22, the approximate model probabilities under normality seem very reasonable. Similar results have also been obtained with simulated data that include counts as small as 0 and 1. In the simulated data, the Laplace and transformed model probabilities agree for high probability models, but there is less agreement for lower probability models and more variability than what is exhibited in Figure 1. Visually the symmetry transformation provides the best overall agreement with the Laplace approximation, with a Kullback Leibler divergence of 0.08, although all methods give similar agreement for high probability models.

7 Discussion

The methods in sections 4 and 5 provide computationally simple approximations to the posterior distributions. The methods in section 4 rely on normality and linearity. If both of these assumptions are valid, under the constant variance assumption then the independent proposal distribution is appropriate. The method in

section 5 directly fits a model of independence. As these assumptions may not be tenable, we are exploring various diagnostics that can suggest when the independence methods may not be appropriate.

The methods using transformations to achieve normality in section 4, also required that we approximate the nonlinear regression, $h(g^{-1}(\mathbf{X}\Gamma\beta))$, by a first order linear approximation. If this approximation is poor, then the approximate model probabilities may be questionable. In nonlinear regression, there are many diagnostics based on second order expansions that can be used to address this; see Bates and Watts (1986) or Kass and Slate (1994). If there are indications that the linear approximation is not adequate, we may want to consider using some other transformation, or another approach to sample from the model space.

The approach in section 5, based on estimating the model probabilities via a model of independence for the model space, can be extended to include the second order or higher order interactions. This also provides a diagnostic check for whether the independence model is a reasonable approach, since one could proceed with estimating second order and higher terms and checking if they are "significant". We can informally test whether these additional terms are necessary before proceeding with the independence model as a proposal distribution. If it appears that the higher order terms are important, then we can actually use an estimated model with two-factor interactions as a proposal distribution or use other proposal distributions.

References

Bates, D.M. and Watts, D.G. (1986). *Nonlinear regression analysis and its Application.* John Wiley & Sons, New York.

Carlin, B.P. and Chib, S. (1995). Bayesian Model Choice via Markov chain Monte Carlo. *Journal of Royal Statistical Society - Series B*, 57, 473–484.

Clyde, M., DeSimone, H. and Parmigiani, G. (1996). Prediction via orthogonalized model mixing. *Journal of the American Statistical Association* **91** 1197-1208.

Clyde, M. and Parmigiani, G. (1996). Orthogonalizations and Prior Distributions for Orthogonalized Model Mixing. In: *Modelling and Prediction: Honoring Seymour Geisser*, edited by Jack C. Lee, Wesley O. Johnson and Arnold Zellner, Springer-Verlag. pp 206-227.

Clyde, M., Parmigiani, G., Vidakovic, B. (1995). Multiple Shrinkage and Subset Selection in Wavelets. ISDS DP95-37.

DeSimone, H. (1996). *Prediction Using Orthogonalized Model Mixing.* PhD Dissertation. ISDS, Duke University.

Draper, D. (1994). Assessment and propagation of model uncertainty (with Discussion). *Journal of the Royal Statistical Society* 56.

Gelfand, A.E. and Dey, D.K. (1992). Bayesian model choice: asymptotics and exact calculations. Journal of the Royal Statistical Society, Series B, 56, 501–514.

George, E.I. and McCulloch, R. (1993). Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, 88, pp. 881–889.

George, E.I. and McCulloch, R. (1994). Approaches for Bayesian Variable Selection. Graduate School of Business, University of Chicago.

George, E.I. and McCulloch, R. (1994b). Stochastic Search Variable Selection. In *Practical Markov chain Monte Carlo*, ed. W.R. Gilks, D.J. Spiegelhalter and S. Richardson.

George, E.I., McCulloch, R., and Tsay, R. (1994). Two Approaches to Bayesian Model Selection with Applications. In *Bayesian Statistics and Econometrics: Essays in Honor of A. Zellner*, ed. Berry D.A., Chaloner K.M., Geweke J.F, New York.

Geweke, J. (1995) Variable Selection and Model Comparison in Regression. In *Bayesian Statistics 5*, ed. J.M. Bernardo, J.O. Berger, A.P. David, and A.F.M. Smith., Oxford Press.

Healy, M.J.R. (1988). *GLIM: An Introduction* Clarendon Press, Oxford.

Hougaard, P. (1982). Parameterizations of non–linear models. *Jour. Roy. Statist. Soc. Ser. B.* **44** 244-252.

Kass, R.E. and Slate, E.H. (1994). Some diagnostics of maximum likelihood and posterior nonnormality. *Annals of Statistics* **22** 668-695.

Kuo, L. and Mallick, B. (1994). Variable Selection for Regression Models. Department of Statistics, University of Connecticut, and Dept of Math, Imperial College, London England.

Leamer, E.E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data* Wiley, New York.

Madigan, D.M. and York, J. (1995). Bayesian Graphical Models for Discrete Data. International Statistical Review, 63, 215-232.

Madigan, D.M. and Raftery, A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association.* 89, 1535–1546.

Madigan, D.M., Gavrin, J., and Raftery, A.E. (1994). Eliciting prior information to enhance the predictive performance of Bayesian graphical models. Communications in Statistics - Theory and Methods, 24, 2271-2292.

Mitchell, T.J. and Beauchamp, J.J. (1988). Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association* 83 , pp. 1023–1036.

Morris, C.N. (1983) Natural Exponential Families with Quadratic Variance Functions: Statistical Theory. *Annals of Statistics* **11**, 515-529.

Raftery, A.E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika* **83** 251:266.

Raftery, A.E., Madigan, D.M., and Hoeting, J. (1997). Model selection and accounting for model uncertainty in linear regression models. *Journal of the American Statistical Association* forthcoming.

Raftery, A.E., Madigan, D.M., and Volinsky C.T. (1996). Accounting for Model Uncertainty in Survival Analysis Improves Predictive Performance (with discussion). In *Bayesian Statistics 5*, ed. J.M. Bernardo, J.O. Berger, A.P. Dawid and Smith, A.F.M. Oxford Press. pages 323-350.

Volinsky, C., Madigan, D., Raftery, A.E., and Kronmal, R. (1996). Bayesian model averaging in proportional hazard models: Assessing stroke risk. Technical Report 302, Dept of Stat, Univ of Washington.

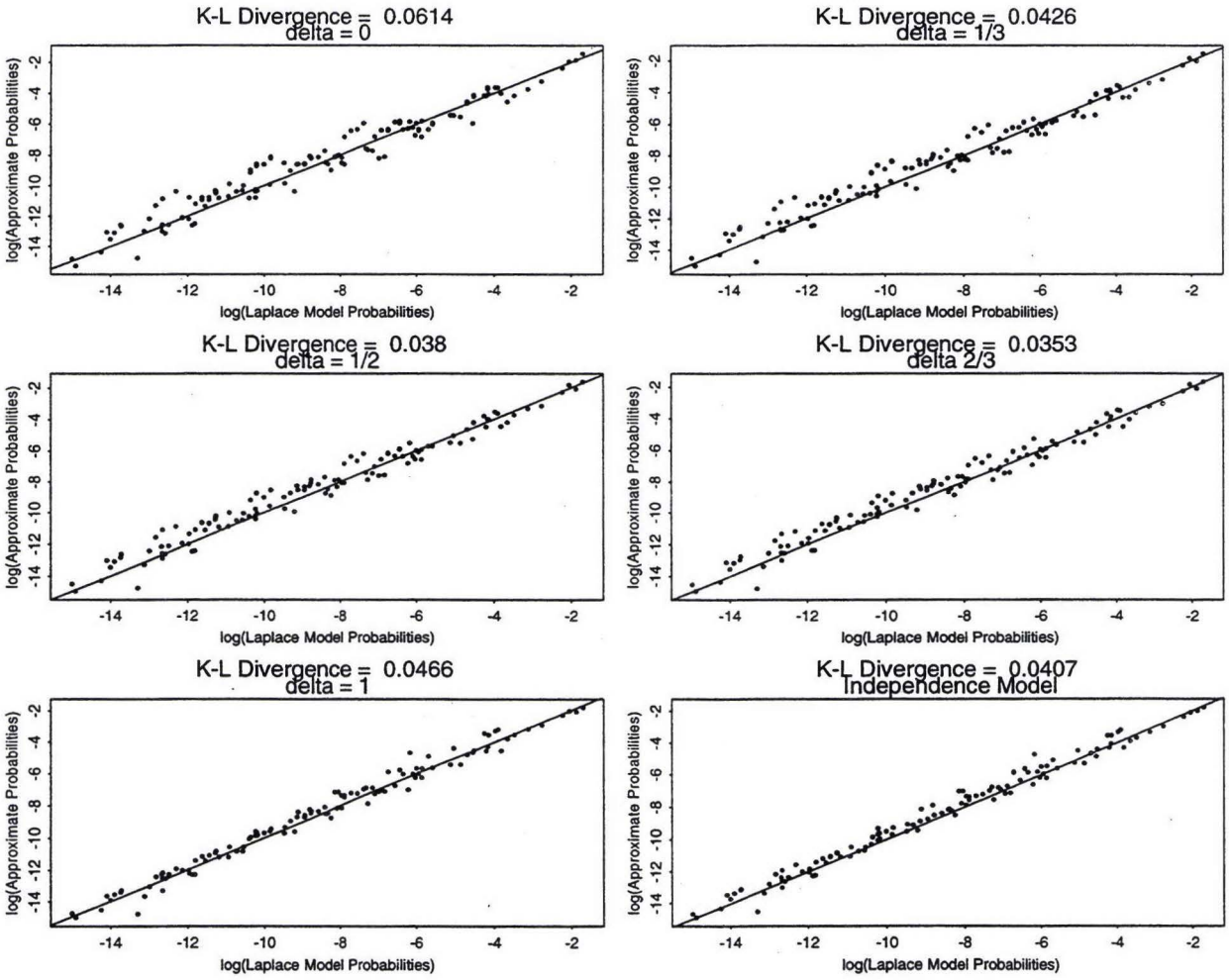

Figure 1: Comparison of model probabilities (estimated by the Laplace method) to approximate model probabilities under the different power transformations $\mathbf{Y}$ and the estimated model of independence for the Healy data.