Methods/Lit Review

Raven McKnight

Poisson Regression

Each model presented in this honors is built upon a simple Poisson regression. Poisson regressions are used to model count data. The model assumes that the natural log of the mean of response variable Y can be modeled by a linear combination of predictors. Poisson regressions are a form of generalized linear models with the natural log as its link function. For response variable Y and covariates $x_1, x_2, ..., x_k$, the basic Poisson regression can be written

$$Y_i \sim \text{Poisson}(E_i \lambda_i) \tag{1}$$

$$\eta_i = log(\lambda_i) = \beta_0 + \sum_{k=1}^{K} x_k^T \beta_k \tag{2}$$

where β_0 is an optional intercept term and E_i is an offset term.

The offset E_i is often termed "exposure" in applications such as disease risk modeling where the offset may be the number of observed cases in a previous year, etc. More technically, the offset term scales the Poisson output to be a rate rather than a count. This is appropriate when observations i have different potentials for response Y. For example, a county with higher population will naturally have a higher count of patients with asthma than a county with a smaller population. Mathematically, the offset is a covariate with parameter set equal to 1. Recall, the Poisson regression assumes we can model the mean of response Y, $\bar{Y}$ with a combination of linear predictors:

$$log(\bar{Y}) = \beta_0 + \sum_{k=1}^{K} x_k^T \beta_k$$

Therefore, if we wish to model the rate Y/E, the equation is rewritten

$$log(\bar{Y}/E) = \beta_0^{'} + \sum_{k=1}^{K} x_k^T \beta_k^{'} \tag{3}$$

$$log(\bar{Y}) = log(E) + \beta_0^{'} + \sum_{k=1}^{K} x_k^T \beta_k^{'} \tag{4}$$

The most obvious assumption made in a Poisson regression is that response variable Y follows a Poisson distribution. The primary assumption this makes about our data is that $\text{Var}(Y) = \text{E}(Y)$. When this assumption is not met, we can either use a generalization of the Poisson distribution, such as a Negative-Binomial distribution, or add a parameter to account for extra-Poisson variance.

For the purposes of this study, we can add a set of heterogeneous random effects θ to account for overdispersion (where $\text{Var}(Y) > \text{E}(Y)$). Following Morris et al (2019), the second line of **1** can be written

$$\eta_i = log(\lambda_i) = \beta_0 + \sum_{k=1}^{K} x_k^T \beta_k + \theta \tag{5}$$

where θ is given a vague normal prior centered around 0, such as $\theta \sim \text{Normal}(0, 1)$. The addition of these random effects improve fits on data with higher-than-expected variance. With sufficiently vague priors, θ can account for most or all overdispersion not accounted for by $\beta_0 + \sum_{k=1}^{K} x_k^T \beta_k$.

Horseshoe Priors and Variable Selection

Often, the coefficients $\beta_1, ..., \beta_k$ are given simple normal priors such as $\beta \sim \text{Normal}(0, 1)$ in the case of standardized parameters. When parameter vector x contains many possible covariates, we may reasonably assume that some entries in β are 0 – in other words, that some of the parameters in x have no effect on Y. In frequentist applications, we might use a method such as the LASSO (Least Absolute Shrinkage and Selection Operator) to identify relevant variables. There are several Bayesian alternatives for variable selection which generally consist of applying particularl priors to β.

Spike and Slab

In the Bayesian setting, there are two primary methods for variable selection. The first is called the "spike-and-slab" prior (Mitchell and Beauchamp, 1988; George and McCulloch, 1999) and has often been considered the "gold standard" for sparse Bayesian regression, or variable selection (Piironen and Vehtari, 2017). This prior is often expressed as

$$\beta_j | \lambda_j, c, \epsilon \sim \lambda_j \text{Normal}(0, c^2) + (1 - \lambda_j)\text{Normal}(0, \epsilon^2) \tag{6}$$

$$\lambda_j \sim \text{Bernoulli}(\pi) \tag{7}$$

for $j = 1, 2, ..., J$ where $\lambda_j \in 0, 1$ indicates whether β_j is from the "spike" (ie, near zero) or the "slab" (ie, nonzero). The "spike" is the area where most of the prior density for β is centered (generally around 0) while the "slab" refers to the low-density extent of the prior. The spike-and-slab prior encourages all β_j towards zero; only β_j sufficiently far from 0 will be estimated to be from the slab.

The primary challenge with the spike-and-slab (and other Bayesian shrinkage methods) is the selction of priors. Priors must be set for the width of the slab c and the prior inclusion probability π. Here, π reflects our prior understanding of the sparsity of β. In practice, we rarely have strong prior knowledge of the number of predictors we expect to be distinguishable from zero. This can make the implementation of sparse regression more challenging than, say, the frequentist LASSO.

Horseshoe

The second method for Bayesian sparse regression is to give β some continuous *sparsity inducing prior*. **Van Erp et al (2019)** provide an introduction to many such continous priors. Ridge regression and the Bayesian LASSO are two of the most approachable priors, but all sparsity inducing priors utilize the same logic: give β a prior with *most* of its mass near 0, to shrink irrelevant coefficients, and the rest of its mass "far" from 0. Identifying "far" depends more or less on the specific data and model in question.

The horseshoe prior, proposed by **Carvalho et all, 2010** is a popular choice in Bayesian literature, in part because it is similar to the "gold standard" spike-and-slab. Following the notation of **Vehtari & Pironeen**, the horseshoe prior can be expressed

$$\beta_k \mid \lambda_k, \tau \sim \text{Normal}(0, \tau^2 \lambda_k^2) \tag{8}$$

$$\lambda_k \sim \text{Half-Cauchy}(0, 1) \tag{9}$$

for $k = 1, 2, ..., K$. The horseshoe prior is so-named because for fixed values $\tau = \lambda_k = 1$, the prior resemebles a Beta(1/2, 1/2) distribution, or a horseshoe.

The horseshoe is often favored because it is a global-local shrinkage prior. This means that τ shrinks *all* parameters towards 0 while the local parameter λ_k and the heavy Cauchy tails allow larger coefficients to remain unshrunk. While this is precisely the goal of a sparsity inducing prior, the horseshoe prior can fail to regularize large coefficients *at all*, which can be a problem when parameters are weakly identified by data. With no regularization applied to the largest coefficients, there is a risk of overfitting. Additionally, there is no consensus in the literature for assigning priors to *tau*. Piironen and Vehtari (2017) introduced the *regularized horseshoe* to address both of these shortcomings of the horseshoe. The regularized horseshoe builds upon **need to figure out equation numbering** such that

$$\beta_k \mid \lambda_j, \tau, c \sim \text{Normal}(0, \tau^2 \tilde{\lambda}_k^2) \tag{10}$$

$$\tilde{\lambda}_k^2 = \frac{c^2 \lambda_k^2}{c^2 + \tau^2 \lambda_k^2} \tag{11}$$

$$\lambda_k \sim \text{Half-Cauchy}(0, 1) \tag{12}$$

$$c \sim \text{ Inverse-Gamma}(a, b) \tag{13}$$

where $c > 0$ helps to regularize β_k far from zero. When c approaches infinity, the regularized horseshoe becomes the standard horseshoe. Additionally, c in the regularized horseshoe corresponds to the slab width in the spike-and-slab prior.

As with the spike-and-slab, the selection of priors is a challenge with the regularized horseshoe...

Spatial Structure

When modeling data with a spatial component, we generally expect adjacent areas to be more similar than areas which are far apart. While this is a straightforward assumption, it can improve model fits in several ways. First, it can encode prior information about response Y not otherwise represented by covariates x (ie, that nearby observations are more similar than far flung ones). Second, it can provide geographic smoothing when observations are sparse or noisy. This is often the case in small areas or when observing events which can only occur at specific locations, such as air quality measured by sensors.

Conditional Autoregressive Priors

This whole section follows Morris et al (probably too) closely right now

We can encode spatial information into Bayesian models in several ways. Conditional Autoregressive (CAR) priors are one of the most widespread Bayesian methods for modeling spatial autocorrelation. Conditional Autoregressive models were introduced in Besag 1974 and remain perhaps the primary method for Bayesian areal data modeling. CAR models are used as *priors*, not as the actual observation model.

Areal data corresponds to finitely many discrete areal units, such as counties or census tracts. Counts in areal units tend to be noisy, particularly when the counted event is rare, the population of areal unit i is small, or the phsyical boundaries of areal units form unusual patterns in the data. Conditonal Autoregressive models combine information from areal unit i as well as its neighbors to smooth noisy counts.

Generally, CAR models use binary neighborhood relationships encoded in a neighborhood matrix. Areal units i and j are considered neighbors if they share a boundary. For strictly rectangualar lattice data, we often choose either a "Queens" or "Rooks" neighborhood structures **(include graphic)**. For irregularly shaped areal units, such as census tracts, we generally default to a queen neighborhood structure wherein areal units that share *any* points of contact are deemed neighbors. For M regions represented in a MxM neighborhood matrix W, $w_{ij} = 1$ if regions i and j are neighbors and 0 otherwise. Matrix W is symmetric. Note that regions i are not considered neighbors with themselves.

The spatial interactions between i and j are represented by random variable ϕ. Here, ϕ is a vector $\phi = \phi_1, \phi_2, ..., \phi_M$. The distribution of each ϕ_i is determined by the sum of its neighbors values such that

having a notation crisis

$$\phi_i \mid \phi_j, j \neq i \sim \text{Normal}(\sum_{i=1}^{M} w_{ij}\phi_j, \sigma^2)$$

Besag (1974) proved that the joint distribution of ϕ is a multivariate normal centered at 0 where its variance is given by a symmetric positive definite precision matrix Q. Matrix Q is simply the inverse of the covariance matrix $\sum$. This finding simplifies the above to

$$\phi \sim \text{Normal}(0, Q^{-1})$$

Precision matrix Q can be defined for multivariate normal ϕ using two other MxM matrices: the neighborhood matrix W as well as diagonal matrix D in which d_{ii} indicates the number of neighbors region i has. All off-diagonal entries are 0. Using these matrices, Q is defined

$$Q = D(I - \alpha W)$$

where I is the identity matrix and $\alpha \in (0, 1)$ determines the amount of spatial autocorrelation present in Y. Parameter α is a key component of CAR models. At $\alpha = 0$, the model assumes spatial independence and at $\alpha = 1$, perfect spatial autocorrelation.

Following Morris et al, the log probability density of ϕ is proportional to

$$\frac{M}{2}\log(\det(Q)) - \frac{1}{2}\phi^T Q\phi$$

Calculating this value is computationally expensive. For models with 1,000 areal units, for example, calculating the determinant requires 1 billion operations (Morris et al, 2019).

Intrinsic Conditional Autoregressive

The Intrinsic Conditional Autoregressive (ICAR) prior is a slight simplifation of a CAR prior. ICAR priors set $\alpha = 1$ which simplifies the definition of Q

$$Q = D - A$$

thereby setting $det(Q) = 0$. As a result, the first term in **equation numbering** can be simpliefied to

$$-\frac{1}{2}\phi^T Q\phi$$

This reduces the computational expense of computing CAR models significantly. Notably, the ICAR prior is improper but yields proper posteriors (Morris et al, 2019).

The ICAR specifies each ϕ_i to be normally distributed with its mean equal to the mean of its neighbors values. This is how the CAR/ICAR model "borrows strength" from geographic neighbors to smooth noisy estimates. Intuitively, the variance of each ϕ_i decreases as its number of neighbors d_i increases. The conditional distribution for each ϕ_i can be written

$$\text{p}(\phi_i \mid \phi_{i\sim j}) = \text{Normal}(\frac{\sum_{i\sim j}\phi_i}{d_i}, \frac{\sigma_i^2}{d_i})$$

Here, the variance σ is unknown. The conditional specification of each ϕ_i is helpful in building intuition about the assumptions the ICAR component makes and how it performs its "smoothing". The joint distribution can be rewritten

$$\mathrm{p}(\phi) \propto \exp(-\frac{1}{2}\sum_{i\sim j}(\phi_i - \phi_j)^2)$$

BYM and BYM2

The specific ICAR model used in this analysis is the Besag-York-Mollie (BYM) Poisson model. The BYM model contains random effects and a spatial ICAR term to account for both spatial and non-spatial heterogeneity. Note that in **section??*, we introduced a random effect term θ. Therefore, the BYM model can be though of as decomposing θ into spatial and non-spatial components.

Specifically, the BYM model is written

$$\eta_i = \beta_0 + \sum_{k=1}^{K} x_k^T \beta_k + \theta + \phi$$

where ϕ is the addition: an ICAR component.

The BYM model is appealing in its simplicity. However, the lack of hyperpriors specified for θ and ϕ can make the model incredibly challenging to fit. In theory, extra-Poisson variance could be explained 100% by θ, 100% by ϕ, or anywhere in between. In *practice*, we likely have limited information about where that ratio falls. Riebler et al (2016) further explain the sampling issues faced by the BYM model faced with no hyperpriors. In short, the sampler is forced to explore all possible combinations of θ and ϕ, no matter how unlikely a particular combination my be.

There are some suggestions in the literature for setting hyperpriors for θ and ϕ (Besag and Mollie, 1991; Clayton and Montomoli, 1995). The existing methods, however, tend to rely on the specific data at hand which makes the selection of hyperparameters time unnecessarily time consuming.

The BYM2 model proposed by Riebler et al (2016) aims to solve these sampling problems. The primary difference between the BYM and BYM2 models is the addition of a mixing parameter, ρ. The BYM2 model rewrites the BYM as

$$\eta_i = \beta_0 + \sum_{k=1}^{K} x_k^T \beta_k + ((\sqrt{\frac{\rho}{s}})\theta^* + ((\sqrt{1-\rho})\phi)\sigma$$

where $\rho \in (0, 1)$ determines how much variance/overdispersion is caused by spatial versus Non spatial error terms. Like the popular Leroux CAR prior, proposed by Leroux et al (2009), the BYM2 scales both θ and ϕ by σ, the standard deviation of the combined error terms (Morris et al, 2019). In this parameterization, s is a scaling factor such that $\mathrm{Var}(\theta_i) \approx \mathrm{Var}(\phi_i) \approx 1$. The equal unit variance is necessary for σ to truly be the standard deviation of the error terms.

Setting priors is somewhat more straightforward for the BYM2. The ICAR component, ϕ^* remains unchanged. Riebler and Morris recommend the prior $\theta \sim \mathrm{Normal}(0, n)$ where n is the number of connected graphs in the neighborhood graph. In many cases, such as modeling data for all counties in a state, $n = 1$. When $n > 1$, the variance is different in each subgraph which affects σ and s. It is possible to fit the model in this case, although computation time is much longer.

Stan

The models above can all be fit using the Stan Programming Langauge (cite). Stan is probabilistic programming language used to compute joint log probability densities.

HMC/NUTS Sampling

Stan utilizes the Hamiltonian Monte Carlo (HMC) and specialized No U-Turn Sampler (NUTS).

Include parameters we'll set : treedepth, adapt_delta, what they correspond to mathematically

Divergent Transitions and Other Diagnostics

Stan output provides several diagnostics which are helpful in confirming that the HMC/NUTS algorithm has converged and identifying problems with the models. There are a few we will refer to in the case study portion of this honors.

The quickest diagnostic Stan provides is $\hat{R}$, or `Rhat`. $\hat{R}$ indicates how well the Markov chains have mixed by comparing between- and within-chain estimates for each parameter. At model convergence, $\hat{R} = 1$. According to Stan best practices, $\hat{R} < 1.1$ generally indicates sufficient mixing. Mixing (and therefore, $\hat{R}$) can often be improved by running chains for more iterations.

Stan output `n_eff` is the estimated Effective Sample Size (ESS). ESS is the number of independent samples our sample is equivalent to. In other words, if `n_eff` = 1,000, we expect to gain the same amount of information from our sample as we would from 1,000 independent samples. Stan also provides Bulk ESS and Tail ESS estimates to determine how well Stan is exploring the most dense sections of the posterior (Bulk) versus the 5% and 95% quantiles (Tail). When `n_eff` is small, posterior estimates are unreliable. When `n_eff` is significantly less than the actual number of samples n, there is likely unusually high autocorrelation between samples. Through their GUI ShinyStan, Stan gives a warning when `n_eff` is less than 10% of n. However, if `n_eff` increases linearly with n, there is likely no "problem" with the model and a suitable effective sample size can be reached by increasing the number of iterations chains run for.

Diagnostics $\hat{R}$ and `n_eff` are both helpful "gut-checks" for model convergence and performance. However, neither are sufficient for diagnosing convergence and they should be treated only as initial checks.

Divergent transitions are perhaps the most important "diagnostic", or warning, Stan provides. Divergent transitions occur when the simulated Hamiltonian departs from the actual Hamiltonian trajectory **(need citation)**. When this occurs... In general, if Stan produces divergent transitions, model output must be discarded. Divergent transitions introduce bias which renders posterior estimates unusable. While $\hat{R}$ and `n_eff` can often be improved by running the chains for more iterations, divergent transitions will persist unless sampling parameters such as the target acceptance rate are changed. Often, divergent transitions call for reparameterizing the model.

Graphic Diagnostics, PPcheck

Various graphical checks of priors and posteriors are helpful tools in building, improving, and comparing Bayesian models. The application portion of this honors follows Michael Betancourt's "Towards a Principled Bayesian Workflow" as well as _____ in terms of using visual aids for diagnosing issues with models and comparing quality of fit between models. Only the most "useful" plots are included below, for more on using graphics throughout a Bayesian workflow see Betancourt or Gabry.