BYM2, Blending Error, and Covariate Bias

Background

Spatial Models in Epidemiology

The use of spatial models in epidemiology often tries to link observed spatial covariates with a health outcomes while adjusting for unobserved covariates that may or may not be spatially correlated. By adjusting for this unobserved correlation, the association between observed covariates and the outcome of interest may be better understood. Not accounting for this spatial correlation could lead to biased estimates of the parameter of interest or underestimation of the standard errors/posterior standard deviations(1,5).

On the other hand adding a spatially correlated random effect can also have negative consequences when the underlying process does not have spatially correlated error terms(4). The spatially correlated error term can account for much of the variation that the covariates would normally and there for bias results for those parameters towards the null. Justification for the use (or non use) of spatial terms can often be difficult to argue and their effects often have a non trivial effect on the outcome of the model at hand.

BYM2 and the PC prior

In an analysis done by Riebler et. al. (1) the authors propose a new model that is a modification of the Besag, York and Mollie (BYM) model that takes into account scaling so that both overdispersion and spatial random effects may be included and that they are on a similar scale. To illustrate this let us assume that

$$y_i|\theta_i \sim Poisson(E_i\theta_i)$$

Where y_i is the observed count of some noncontagious disease for an area, θ_i is the relative risk for the area and E_i is the expected value for a particular area. $log(\theta_i)$ may then be modeled using the form

$$log(\theta_i) = \beta \cdot x_i^T + b_i$$

Where x_i is a vector of explanatory covariates and β is an array of parameters that account for the relationships between the covariates and the relative risk. b_i is the random effect term that accounts for the error in the linear relationship between x_i and $log(\theta_i)$. It is in this term that we may also account for the spatial relationships between data points. Dean et al(3) proposes that the functional form of the term b may be written as

$$b = \frac{1}{\sqrt{\tau}}\big(\sqrt{1-\phi}v + \sqrt{\phi}u\big)$$

Where τ is the precision of the random effects, v is the error that occurs from simple overdispersion, u is the error that may be attributed to unobserved spatial correlation and ϕ determines how much each type of error contributes to the final estimate. One difficulty with this model is that it may falsely attribute variation of the errors to the spatial component u when in actuality it is not part of the underlying process. This over-complicates the model and could lead to biased parameter estimates. In addition setting hyperpriors on ϕ and τ becomes difficult across different scenarios as it is dependent on the structure of the spatial graph at hand(1). A way to adjust for this is proposed by Riebler et. al.(1) in which b is modeled using a modified $u_\star$ in which

$$u_\star \sim \mathcal{N}(0, Q_\star^-)$$

$Q_\star^-$ calculation is shown in detail in Riebler et. al.(1). The benefit of this model is that it allows generalized specifications of hyperpriors such that they may be applied to many models despite their geographic structure. In turn, this leads to the application of the penalized complexity (PC) prior. The PC prior allows for a model to include multiple terms however penalizes the model for their inclusion when it is unwarranted. In our case this permits the inclusion of a spatial term that will be essentially removed (i.e. $\phi = 0$) if their is not enough evidence for its inclusion in the data. From here on we will refer to the Dean et al. model with scaling and the PC prior as the BYM2 model. A more detailed description of the PC prior and its applications may be found in Simpson et al.(2).

Testing the BYM2 model

Past Research

In the proposal of the BYM2 Riebler et. al.(1) test how accurately the model is able to decompose to a simpler model (where either $\phi = 0$ or $\phi = 1$) when the data was simulated using such a specification. The BYM2 model was shown to perform well under both situations and perform as well or better than other spatial models at calculating τ. While this testing shows that the model is able to decompose well when their are no true or only spatial effects we would expect most model to have a blending of the two.

In order to test the ability of the BYM2 model to accurately maintain mixture of both spatial and non spatial components testing an array of values of ϕ where $\phi \in [0, 1]$. In this way we may test the ability of the model to maintain the proportionality of the two random effects u and v with out unnecessarily decomposing them. This is especially important for cases where $\phi \approx .5$ and there is equal division of contribution to relative risk from both u and v.

In addition, testing the BYM2 models ability to correctly asses covariate effects in the presence of different values of ϕ must also be assessed. As most epidemiologists are concerned with the correlation between the observed variables and the outcome, it is essential that the BYM2 model is able to accurately recover the true β parameters for a given covariate when the correct model form is used to estimate the parameters.

Methods

In order to test the BYM2 model a set of simulated data was created using the following form

$$log(\theta) = \beta \cdot x^T + (\frac{1}{\sqrt{\tau}}(\sqrt{1-\phi}\ v + \sqrt{\phi}\ u_\star))$$

For testing how well the BYM2 model is able to accurately reproduce ϕ we specify the variables such that β is a vector with one element 0, x is a matrix with a single column which is $N_{iid}(0,1)$, $\frac{1}{\sqrt{\tau}} = .5$, $v \sim N(0,1)$, $u_\star \sim N(0, Q_\star^-)$. We vary values of ϕ to be either
.25, .50, or .75. In this way we can look at how the estimated values of ϕ change as we vary the weight of the random effects to be present for both simple overdispersion and spatially correlated with even weights(.5), biased towards simple overdispersion, or biased towards spatial effects.

To test the effectiveness of the BYM2 model to return accurate estimates of covariates a similar form is used to generate data with the following adjustments. β is a vector of parameters with values $[0, 1, -1]$ and x is a matrix of values with a row for every observation and 3 columns for each beta. The elements within the matrix are created such that $x \sim N_{iid}(0,1)$.

With both of these models we will generate values for relative risk (θ) for each location and from it sample from the poisson distribution in order to have an observed value y while setting the expected value to 60 ($E = 60$). We will then fit a bym2 model with the correct number of covariates (0 for the first model and 2 for the second) and asses the models ability to recover the true parameters, both β and ϕ, from the simulated data.

The data generation process was applied to an area of the United States which covers the states of Texas and Louisiana. Each unit within the area was a county. The region was chosen due to the authors familiarity with the area and the spatial continuity. Each unique paramterization of the model, described above, generated 500 unique simulations resulting in a total of 6000 analysis (500 random effect simulations; 3 values of delta, covariates present or not).

All analysis was done in R (version 3.2.3) and a reproducible version of the code may be found at:

https://github.com/nmmarquez/spatial_epi/blob/master/project/project_outlay.r

Results

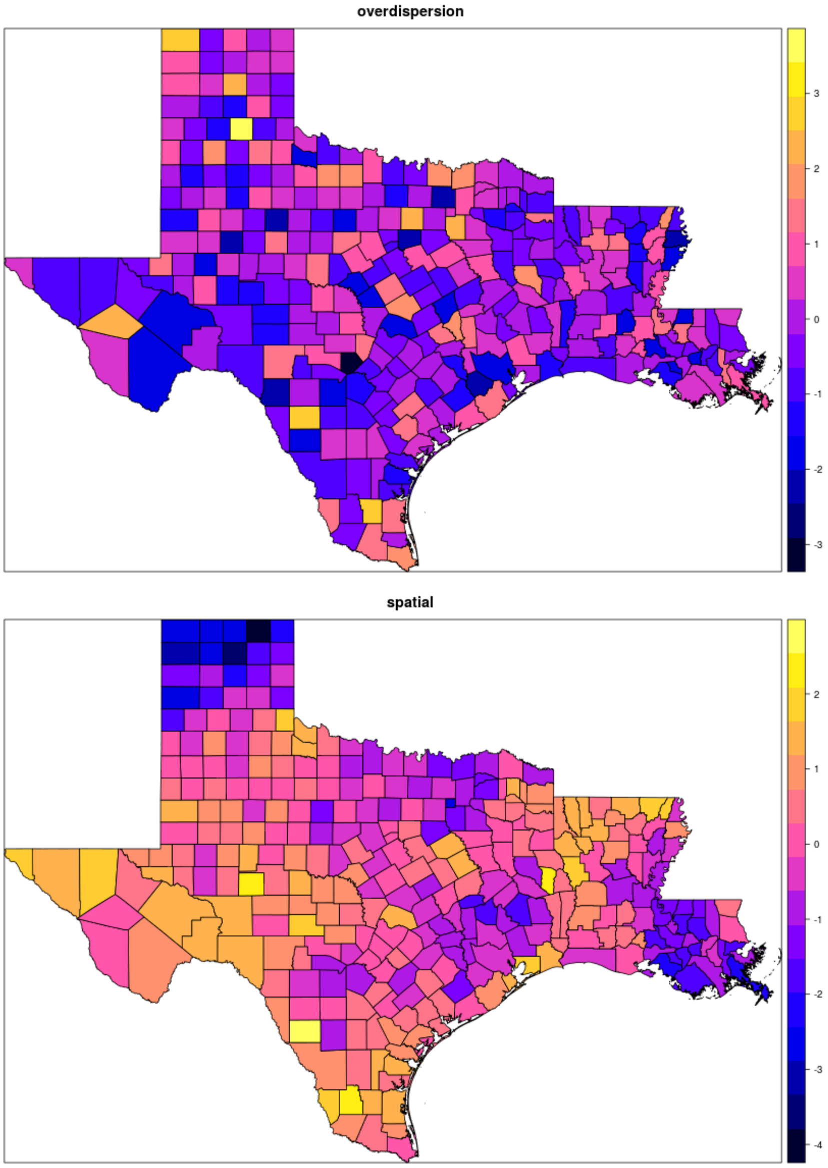

Figure 1: Simulated Overdispersion and Spatial Random Effects

Each simulation was run and the root squared error(RSE) was calculated for all parameters (β, τ, ϕ). The distributions of the RSE for the PC and the non PC BYM2 model are shown below for both covariate and no covariate inclusion in the form of density plots.

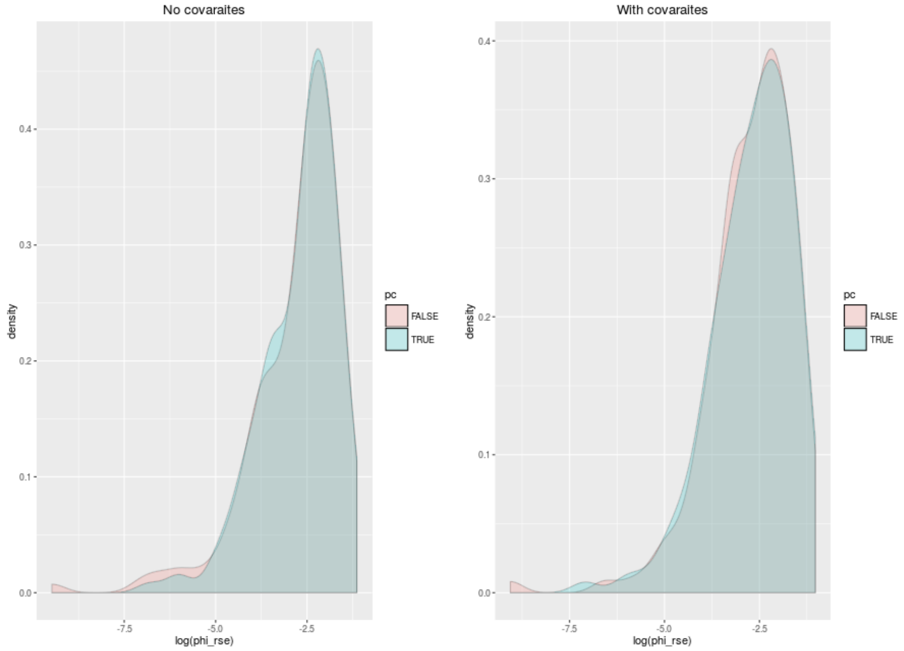

Figure 2: Density Plots of RSE Distribution by PC Inclusion

In addition to the value of ϕ being accurately recovered there appears to be minimal differences between the BYM2 model with and without the PC prior for the recovery of all parameters. The first graphs show the RSE of parameters when no covaraites are included.

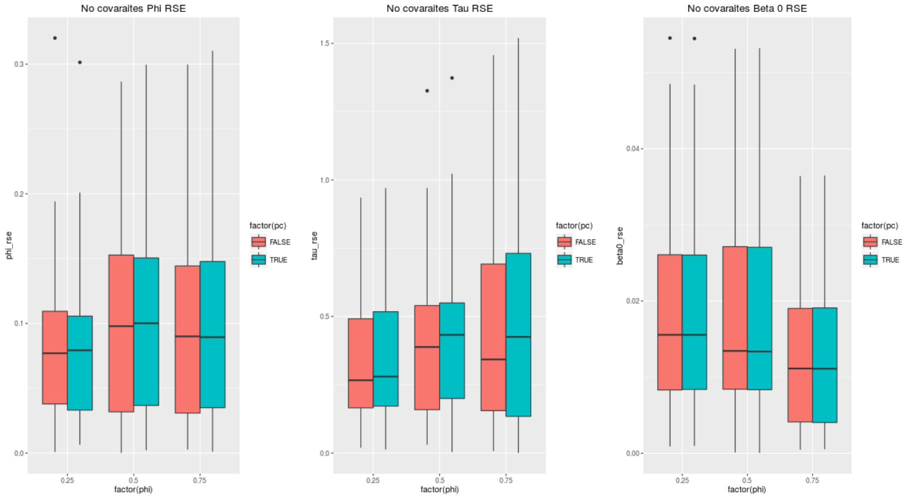

Figure 3: RSE of all Parameters In Covariate Free Simulations

The second graph shows the same Paramters estimated when covariates are included as well as the RSE of the additional β terms.

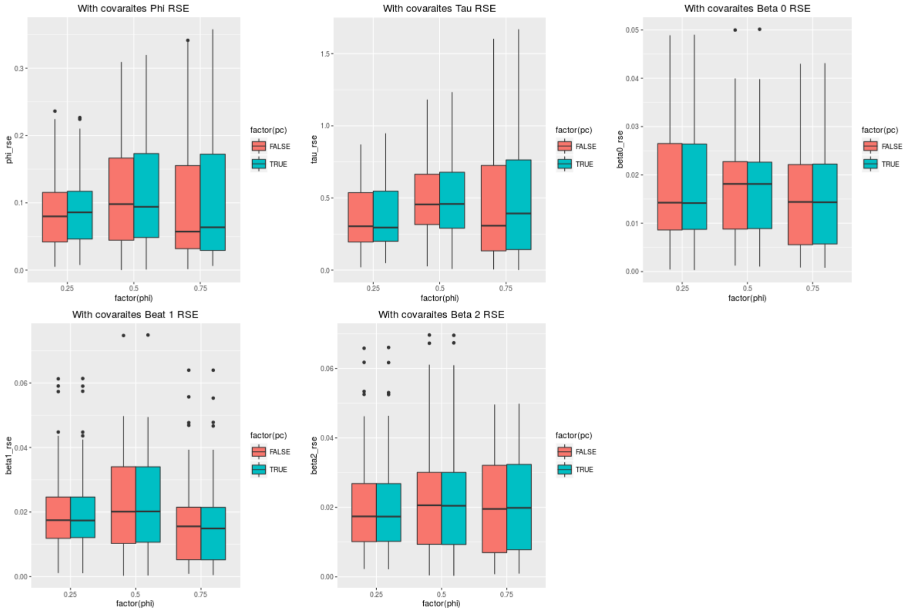

Figure 4: RSE of all Parameters In Covariate Simulations

Aside from both model appearing to have some strong outliers in the of both β_1 and β_2 both forms of the model perform relatively well in reconstructing parameters.

Discussion

We found that the use of the PC prior in conjunction with a modified bym model provided consistently accurate estimates of both ϕ and β. When ϕ was adjusted to represent mostly spatial, mostly overdispersion, or an even split between the errors the RSE was similar between the BYM2 model whether or not the PC prior was added. This showed that the PC prior did not have an adverse effect on the estimation of ϕ such that it would decompose into a simpler model.

In addition to this we found that the estimates for β were also accurate in the model estimation process. Though this was the likely the case given that the estimates of ϕ were accurate. This ability to reproduce the original beta parameter estimates lends to the body of evidence of the benefits of using the BYM2 model over the original BYM model. The BYM2 model will not only more accurately shrink to the base models when no evidence in the data suggests strong overdispersion or spatial effects but will also accurately recover a mixture of the two with decomposing to the base models i.e.($\phi = 0$ or $\phi = 1$).

While the BYM2 model has held up well to the tests that as specified in this paper and the analysis in Riebler et. al.(1) there are still are other approaches to test the model with. All test up to this point have been done using a data generation process that closely resembles the model. If we were to create spatially dependent variables in another way, such as using an extended neighborhood or cluster approach, then both ϕ and τ would offer little insight and the retrieval of the β parameters would be much more difficult. In addition the use of other generalized linear models, other than poisson, would show how flexible this model is in applications outside of estimating relative risk.

Citations

1. Reibler A, Sorbye SH, Simpson Daniel, Rue H. An intuitive Bayesian spatial model for disease mapping that accounts for scaling. arXiv preprint arXiv:1601.01180 2016
2. Simpson DP, Rue H, Martins TG et al. Penalising model component complexity: A principled, practical approach to constructing priors. arXiv preprint arXiv:14034630 2015
3. Dean C, Ugarte M and Militino A. Detecting interaction between random region and fixed age effects in disease mapping. Biometrics 2001; 57(1): 197–202.
4. Hodges JS, Reich BJ. Adding Spatially-Correlated Errors Can Mess Up the Fixed Effect You Love. The American Statistician Volume 64, Issue 4, 2010
5. Bruns SB, Ioannidis JPA. p-Curve and p-Hacking in Observational Research. PLoS ONE 11(2) 2016