Modelling spatiotemporal variance in epidemiological contexts

Michael Redman, CID: 00826863

December 2016

1 The problem

The correct identification of blah insert stuff here.

2 Naive/Other models

2.1 SATScan

2.2 CUSUM

3 Bayesian Model Construction

- Space-time seperability
- Identification issuses in mixture model
- Bayesian model selection
- Prior on mixture component probability
- Bayesian classification
- Compositional models
- Posterior simulations + ML classifier e.g. k - NN

3.1 Hyperparameter priors

- Gelman 2006 for variance parameters
- Stan wiki for others

3.2 Convergence statistics

- Trace plot
- Gelman-Rubin statistic
- Multiple-chains are run not for computational benefits but to assess convergence
- Gelman-Rubin-brooks plot
- Lack of divergences which are "incredibly sensitive to the kind of pathologies that can obstruct geometric ergodicity"

4 Smoothing

Ideally we wish to identify potential local risk factors in the aetiology of a disease, say, carcenogenic hazard from industrial polution. So it's clear that the ability to incorporate a high level of spatial granularity in our model is of value in these contexts. However this comes with the trade-off of greater variance in the counts, making identification of abnormal temporal trends difficult, especially for diseases with low incidence. Therefore we need to employ an element of smoothing over the local neighbourhoods of each region. This can be done in a variety of methods. One possibility is the use of splines such as in (source here) but in this project we will primarily look at using a conditionally autoregressive prior.

4.1 CAR models

Markov random field.
Conditionally autoregressive models can be best understood when specified in terms in terms of their conditional distribution.

$$v_i \mid v_j \; j \neq i \sim N(\alpha \cdot \bar{\mu_i}, \; \sigma_v^2/k_i) \tag{1}$$

where k_i is the number of neighbours adjacent to region i,

$$\bar{\mu_i} = \sum_{j \in \partial i} \frac{\mu_j}{k_i} \tag{2}$$

and α is a parameter measuring the degree of spatial dependence.
However as this specification is a markov random field and not a directed acyclic graph we can't use this definition in non-gibbs sampling methods we need the v_i to be jointly specified. Thankfully it is possible for it to be expressed in terms of a multivariate normal distribution as follows,

$$v \; \sim \; N(0, \; \sigma_v^2 \cdot [D(I_n - \alpha B)]^{-1}) \tag{3}$$

where

$$D = \text{diag}(k_i) \tag{4}$$

$$B = D^{-1}W \tag{5}$$

$$(W)_{ij} = \begin{cases} 1, & i \leftrightarrow j \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

It is intuitively clear that the precision matrix here is sparse and so naive calculations will be very inefficient see the section on computational considerations for some more sophisticated methods that we will use to simulate the distribution.

- Cite http://www.biostat.umn.edu/ brad/software/jbc.proofs.pdf
- Write in terms of join distribution for non-gibs samplers.
- Spatial dependence parameter α to prior or not to prior?
- Intrinsically autoregressive model improper as covariance matrix is semi-definite

4.2 BYM prior

4.3 Temporal smoothing

Similarly to in the spatial setting we can use a prior on the temporal component that assumes a level of similarity between adjacent regions - here consecutive time points. The prior prefered here is the one dimensional random walk prior, which we will denote by

$$\xi_{1:T} \sim \text{RW}(1) \tag{7}$$

where the dimensionality will often be infered from the context.

5 Individual trend model

The first Bayesian model we wil consider is similar to that which is formed in Someone et al. In this setting we construct two alternate hypothesis for each region, one where the counts at the region are broadly in keeping with some "global" temporal trend (subject to localised spatial deviations captured with a conditionally autoregressive prior), and in the other the region has its own individual temporal trend. Then by some method of classification we sort the regions into those deemed most likely to follow the global model and those exhibiting behaviour more typical of the second model - and label these regions "unusual".

5.1 Baystdetect and the cut function

In the original paper *Baystdetect* the use of the cut function in the *BUGS* language is employed to fit the two models to the data seperately and then the model selection is undertaken afterward. This method, which prevents the flow of information between the two models is defended in (Nicky Best presentation here) but has been met with some level of skepticism in the community, for example in Andrew Gelmans posts to the Stan mailing list here (insert link), as the analysis is not "truly Bayesian". Nethertheless, we examine this paradigm and compare it to fully Bayesian methods.

5.1.1 Model specification

We denote the counts at region i at time t by $Y_{i,t}$ and model them by a Poisson process

$$Y_{i,t} \sim \text{Poisson}(E_{i,t} \cdot \mu_{i,t}) \tag{8}$$

where E_{it} is the expected count based on population numbers, demographics etc and $\mu_{i,t}$ is the rate parameter by which we impute the two models behaviours. This rate variable we parameterize additively on the log scale for both models as follows

$$\log(\mu_{i,t}) = \begin{cases} \lambda_i + \gamma_t \ (+\ \alpha_0) & \text{Model 1 for all } i,t \\ u_i + \xi_{i,t} & \text{Model 2 for all } i,t \end{cases} \tag{9}$$

Here we see that for Model 1 we assume space-time seperability in the rate paramater with the components given the following priors

$$\alpha_0 \sim \text{Flat}(\mathbb{R}) \tag{10}$$
$$v_{1:N} \sim \text{CAR}(W, \sigma_v) \tag{11}$$
$$\lambda_{1:N} \sim \text{Normal}(v, \sigma_\lambda) \tag{12}$$
$$\gamma_{1:T} \sim \text{RW}(1) \tag{13}$$

We see here the BYM prior on the spatial component, imposing a smoothing constraint, and a one-dimensional random walk prior on the temporal component.
The variance hyperparameters are given (insert prior here) as recomended in (possibly Gelman 2006).

For the second model we drop the assumption of space-time seperability and each region gets its own temporal trend as follows

$$u_i \sim \text{Normal}(0, 1000) \qquad \text{for } i = 1, \ldots, N \tag{14}$$
$$\xi_{i,1:T} \sim \text{RW}(1) \qquad \text{for } i = 1, \ldots, N \tag{15}$$

6 Excess variability model

7 Computational considerations

7.1 Sampling methods

- Mainly using stan for the model
- Also used BUGS for mixing and speed comparison
- Discuss pymc3 too and compare with stan's parameter tuning/ initilization with ADVI

7.1.1 Gibbs sampling

7.1.2 Metropolis

7.1.3 Hamiltonian Monte-Carlo

- Rotational invariance
- Include some diagrams from MIchael Betancourt's papers

7.1.4 Autodiff/black-box variational inference?

7.2 Reparamterization of the models

Removing conditional dependencies e.g.

$$\lambda \sim N(v, \sigma_\lambda^2) = N(0,1) \cdot \sigma_\lambda^2 + v$$

7.3 Marginilizing over the mixture component

- Rao-Blackwellization?

7.4 Timing data?