Modeling 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?

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

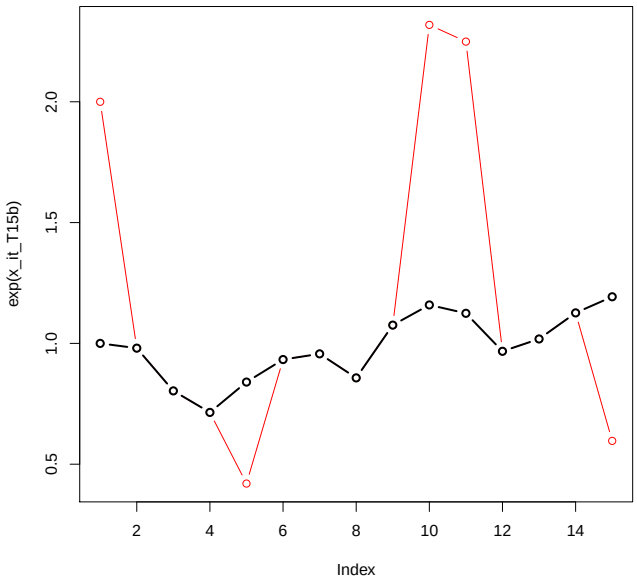
$$\xi_{1:T} \sim \text{RW}(1) \tag{7}$$

where the dimensionality will often be infered from the context. Note that this can be represented by a CAR model and is sometimes represented as such in the literature.

5 Simulated data

In order to test the accuracy of any potential models we need some data that exhibits trends that exhibit pathelogical behaviour that we wish to identify. The simulated data (include citation for Areti) is of asthma hospitalization counts across 221 clinical commisioning groups in England at 15 time points under a Poisson model with a BYM prior for the spatial component and a RW(1) prior for the temporal component.

A selection of 15 of the regions were chosen according to the 10th, 25th, 50th, 75th and 90th percentiles of the median expected counts over time (three regions per percentile) and were given a deviant temporal trend.



The temporal trends of these regions were manually changed, where – writing the original trend as $g(t)$ and the modified trend as $g^*(t)$ – we have:

$$g^*(t) = g(t) + \log(2) \qquad t = 1, 10, 11 \tag{8}$$

$$g^*(t) = g(t) - \log(2) \qquad t = 5, 15 \tag{9}$$

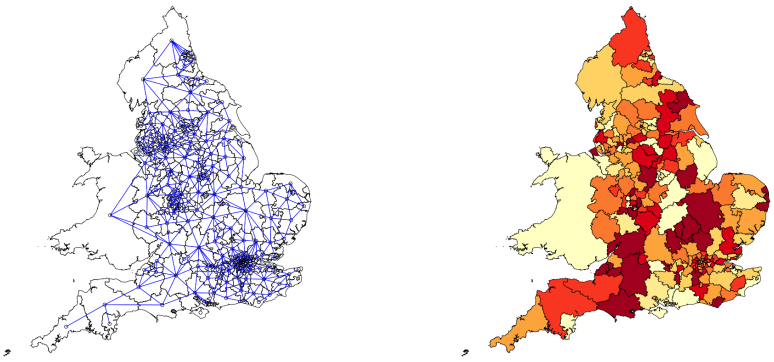
- Check this trend with Marta as the word document and pictured trend are different.

In addition to the observed counts at each point we have the expected number of cases for each regions based on true hospital admissions records for each CCG and shape files corresponding to the regions in question.

5.1 Cleaning and wrangling

Of the 211 regions one one had no neighbours, the Isle of Wight, so this region was removed to simplify calculations. An adjacency matrix calculated from the shape data. We had no data for wales so it was removed from the shapedata.

- Remove wales from the above pictures
- What to include here?



(a) Adjacency (b) Expected

Figure 1: Mapped data

6 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".

6.1 Baystdetect and the cut function

In the original paper [Li et al., 2012] 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.

- Include discussion of model averaging vs larger model
- Cite Gelman et al 2013

6.1.1 Model specification

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

$$Y_{i,t} \sim \text{Poisson}(E_{i,t} \cdot \mu_{i,t}) \tag{10}$$

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

$$\log(\mu_{i,t}) = \begin{cases} \lambda_i + \gamma_t \ (+\ \alpha_0) & \text{Model 1 for all } i,t \\ u_i + \xi_{i,t} & \text{Model 2 for all } i,t \end{cases} \tag{11}$$

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

$$\alpha_0 \sim \text{Flat}(\mathbb{R}) \tag{12}$$
$$v_{1:N} \sim \text{CAR}(W, \sigma_v) \tag{13}$$
$$\lambda_{1:N} \sim \text{Normal}(v, \sigma_\lambda) \tag{14}$$
$$\gamma_{1:T} \sim \text{RW}(1) \tag{15}$$

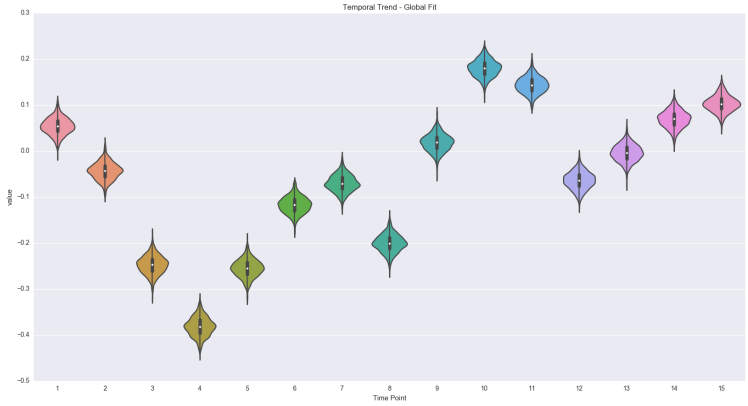
We see here the BYM prior on the spatial component, imposing a smoothing constraint, and a one-dimensional random walk prior on the temporal component.
The variance hyperparameters are given (insert prior here) as recomended in (possibly Gelman 2006).

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

$$i \in 1, \ldots, N \ \begin{cases} u_i \sim \text{Normal}(0, 1000) \\ \xi_{i,1:T} \sim \text{RW}(1) \end{cases} \tag{16}$$

6.1.2 Implementation

The two models were fit to the data using Hamiltonian Monte-Carlo in the *Stan* package – both for over 2000 samples over 4 chains after a "warm-up" period of 1000 samples. Gelman-Rubin statistics for all parameters were under 1.05 (find exact figure) indicating that the had chains converged. Additionally a visual inspection of the trace plots for some select parameters did not indicate any worrying pathalogical behaviour.
These two fits took under 5 minutes each to run after a short compilation which is a huge speed increase over typical *BUGS* implementation.



6.1.3 Classification and accuracy

First looking at the general trend model we see that it has accurately identified the global temporal trend.

6.2 Fully Bayesian model

7 Excess variability model

The second Bayesian model we will look at also attempts to classify the regions, into typical and atypical sets, using a mixture model. Here our abnormal regions, rather than coming from an individual temporal trend, simply have excess inseperable spatio-temporal variance. Some potential advantages of such a model compared to Baystdetect are immediately clear:

- It's relative simplicity allows for easier computation
- It is not immediately clear that an abnormal temporal trend is symptomatic of an endemic problem — variance is a more straightforwardly interpretable parameter
- Identification issues

7.1 Model specification

The model is similar to that of [Abellan et al., 2008], but with a mixture component that's hierachical at the regional level. Like before we model the counts as a Poisson process, $Y_{i,t} \sim \text{Poisson}(E_{i,t} \cdot \mu_{i,t})$, with a rate parameter defined additively on the log-scale

$$\log(\mu_{i,t}) = \lambda_i + \gamma_t + \psi_{i,t} \tag{17}$$

where, as before,

$$v_{1:N} \sim \text{CAR}(W, \sigma_v) \tag{18}$$
$$\lambda_{1:N} \sim \text{Normal}(v, \sigma_\lambda) \tag{19}$$
$$\gamma_{1:T} \sim \text{RW}(1). \tag{20}$$

We see the new component ψ captures a level of space-time inseperability to the counts. At every point (i, t) each component is modeled as coming from a mixture of two normal distribution, with the mixture component at the regional level,

$$\psi_{i,t} \sim z_i \cdot \text{Normal}(0, \tau_1^2) + (1 - z_i) \cdot \text{Normal}(0, \tau_2^2) \tag{21}$$

where

$$z_i \sim \text{Bernoulli}(q_i) \tag{22}$$

and the q_i are given Uniform priors on $(0, 1)$. Here we are marginilizing out the z_i and instead performing inference on the q_i which confers the advantages set out under the computational considerations section. The variance parameters are given half-normal priors, one a vague prior (representing the abnormal regions) and the other an informative prior which restricts the inseperable variance of the 'normal' regions to be very limited. Identification issues and the label switching problem are avoided by defining the larger of the variances addivitely in terms of the smaller:

$$\tau_1 \sim \text{Normal}(0, 0.01) \cdot I(0, \infty) \tag{23}$$
$$k \sim \text{Normal}(0, 100) \cdot I(0, \infty) \tag{24}$$

$$\tau_2 = \tau_2 + k \tag{25}$$

Here I is the indicator function.

7.2 Classification and accuracy

8 Computational considerations

8.1 Basic 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

8.1.1 Gibbs sampling

8.1.2 Metropolis

8.2 Hamiltonian Monte-Carlo

The previous methods, while possessing the desired asymptotic properties, can—and do—take a long time to converge to the target distribution (infinity is a long time!). The aim of *Hamiltonian Monte-Carlo* is to reduce the correlations between successive samples and so dramatically increase the effective sample size with minimal computational overhead. It does this by introducing an auxilliary variable and exploiting the interplay between this and the variables of the posterioir distribution within the framework of Hamiltonian mechanics.

8.2.1 Background

We write q for the variables of the posterior distribution and introduce a new variable p which we will call the *momentum* of the system. Drawing from statistical mechanics, we know that for a given energy function of a system $E(\theta)$ we have the canonical ensemble

$$p(\theta) = \frac{1}{Z} e^{-E(\theta)}. \tag{26}$$

So defining a Hamiltonian of our system where the "kiniteic energy" is dependent on our momentum and the potential energy is dependent only on the posterioir:

$$H(p,q) = \underbrace{T(p|q)}_{\text{Kinetic energy}} + \underbrace{V(q)}_{\text{Potential energy}} \tag{27}$$

looking at the canonical ensemble of the join distribution

$$\pi(p,q) \propto e^{-H(p,q)} \tag{28}$$

$$= e^{-[T(p|q)+V(q)]} \tag{29}$$

$$= e^{-T(p|q)} \cdot e^{-V(q)} \tag{30}$$

$$\propto \pi(p|q) \cdot \pi(q) \tag{31}$$

we see that the posterior and distribution of the momentum are seperable and so independent. This means that if we can sample from $\pi(p,q)$ then our choice of distirbution for the momentum will not effect the calculation of any expectations based on the samples of q.

Note that here we need a definition of potential energy which will give us back the posterior. Clearly the following satisfies this requirement

$$V(q) = -\log \pi(q) \tag{32}$$

So draws from $\pi(p,q)$ will allow us to make inferences on the q, but what advantages does the introduction of this auxilliary variable confer? Well, continuing

with our intuition regarding a physical Hamiltonian system, we know that p and q are related by Hamilton's equations

$$\frac{dq}{dt} = \frac{\partial H}{\partial p} = \frac{\partial T}{\partial p} \tag{33}$$

$$\frac{dp}{dt} = -\frac{\partial H}{\partial q} = -\frac{\partial T}{\partial q} - \frac{\partial V}{\partial q} \tag{34}$$

- Finish off this bit more gracefully

8.2.2 The Method

Hamiltonian monte-carlo works, not by applying a transition over q, but by giving our starting point "a kick" in the form of a random momentum and sampling along the level set of constant energy (defined by the Hamiltonian) using Hamilton's equations to evolve the joint system. Then—after some number of steps—we stop, sample a new momentum, and explore the new phase space that this defines.

The questions now are obvious:

- *How* do we evolve the joint system in accordance with Hamilton's equations?
- Do we need to work out all of the partial deriavtives $\partial V/\partial q_i$ by hand?
- For how long should we explore each level set before sampling a new momentum?
- What proposal density should we use for the momentum?
- Why does this method lead to samples with less autocorrelation?

Other things to talk about:

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

8.2.3 Auto differentiation

8.2.4 Symplectic Integration

- Leap-frog integration
- Metropolis correction

8.2.5 Integration time

- Naive implementations will not preserve detailed balance
- Short-time will give large autocorrelations
- Long-time will face diminishing returns
- Integration time obviously linear in time, so compare to linear
- Exhuastive monte-carlo
- *Identifying the Optimal Integration Time in Hamiltonian Monte-Carlo* (Betancourt 2016)

8.3 Autodiff/black-box variational inference?

8.4 Reparameterization of the models

Removing conditional dependencies e.g.

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

8.5 Marginilizing over the mixture component

As *Stan* used Hamiltonian Monte-Carlo, which is a gradient based sampler, we can't directly specify a discrete prior as used in mixture models. Instead we can "marginalize out" the mixture component to obtain a purely continuous distribution. This also has significan computational benefits (expand on this).

Take for example a mixture of two Normal distributions

$$k \cdot \text{Normal}(\mu_1, \sigma_1^2) + (1-k) \cdot \text{Normal}(\mu_2, \sigma_2^2) \tag{35}$$

$$k \sim \text{Bernoulli}(\lambda)\,. \tag{36}$$

In Monte-Carlo we simply need to be able to calculate the posterior at a specific point. Most software, including *Stan*, rather than multiply the posterior by a chain of conditional probability density functions works on the log scale where to calculate the log posterior we simply need to increment the log posterioir by the log conditional of the hierachical components. So for our mixture of Normals we need to implement the following

$$\begin{aligned}\text{log\_posterior} = \text{log\_posterior} + \log(&\lambda \cdot \text{Normal\_pdf}(x|\mu_1, \sigma_1^2) \\ &+ (1-\lambda) \cdot \text{Normal\_pdf}(x|\mu_2, \sigma_2^2))\end{aligned} \tag{37}$$

which we can see is continuous. The way we can specifically implement this in *Stan* is by using the following manipulation

$$\begin{aligned}
\log(p_X(x|\lambda,\mu_i,\sigma_i^2)) &= \log(\lambda\cdot\text{Normal_pdf}(x|\mu_1,\sigma_1^2) \\
&\quad + (1-\lambda)\cdot\text{Normal_pdf}(x|\mu_2,\sigma_2^2)) && (38)\\
&= \log(\exp(\log(\lambda\cdot\text{Normal_pdf}(x|\mu_1,\sigma_1^2))) \\
&\quad + \exp(\log((1-\lambda)\cdot\text{Normal_pdf}(x|\mu_2,\sigma_2^2)))) && (39)\\
&= \text{log_sum_exp}(\log(\lambda) + \log\text{Normal_pdf}(x|\mu_1,\sigma_1^2), \\
&\quad \log(1-\lambda) + \log\text{Normal_pdf}(x|\mu_2,\sigma_2^2)) && (40)
\end{aligned}$$

- Cite Stan manual
- Rao-Blackwellization?

8.6 Timing data?

References

[Abellan et al., 2008] Abellan, J. J., Richardson, S., and Best, N. (2008). Use of space-time models to investigate the stability of patterns of disease. *Environmental Health Perspectives*, 116(8):1111.

[Li et al., 2012] Li, G., Best, N., Hansell, A. L., Ahmed, I., and Richardson, S. (2012). Baystdetect: detecting unusual temporal patterns in small area data via bayesian model choice. *Biostatistics*, page kxs005.