

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 issues 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, carcinogenic hazard from industrial pollution. 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.

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

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

$$\bar{\mu}_i = \sum_{j \in \partial i} \frac{\mu_j}{k_i} \quad (2)$$

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,

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

where

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

$$B = D^{-1}W \quad (5)$$

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

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

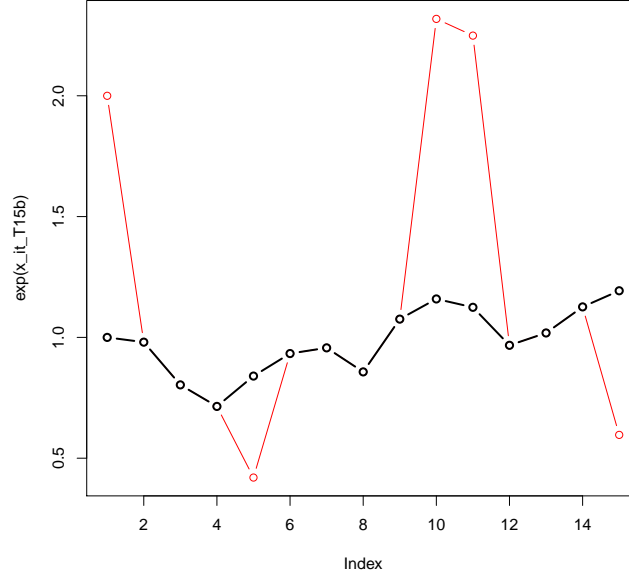
Similarly to in the spatial setting we can use a prior on the temporal component that assumes a level of similarity between adjacent regions – here consecutive time points. The prior preferred here is the one dimensional random walk prior, which we will denote by

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

where the dimensionality will often be inferred 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 pathological behaviour that we wish to identify. The simulated data (include citation for Areti) is of asthma hospitalization counts across 221 clinical commissioning 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:

$$g^*(t) = g(t) + \log(2) \quad t = 1, 10, 11 \quad (8)$$

$$g^*(t) = g(t) - \log(2) \quad t = 5, 15 \quad (9)$$

- 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.

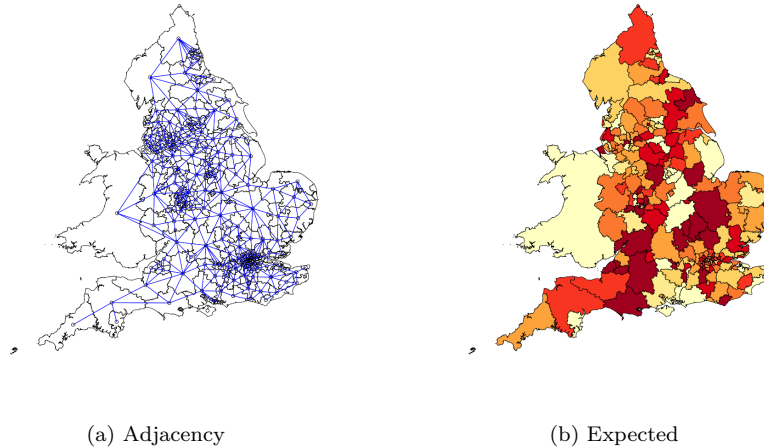


Figure 1: Mapped data

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

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 *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.

- 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

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

where $E_{i,t}$ 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

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

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

$$\alpha_0 \sim \text{Flat}(\mathbb{R}) \quad (12)$$

$$v_{1:N} \sim \text{CAR}(W, \sigma_v) \quad (13)$$

$$\lambda_{1:N} \sim \text{Normal}(v, \sigma_\lambda) \quad (14)$$

$$\gamma_{1:T} \sim \text{RW}(1) \quad (15)$$

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.

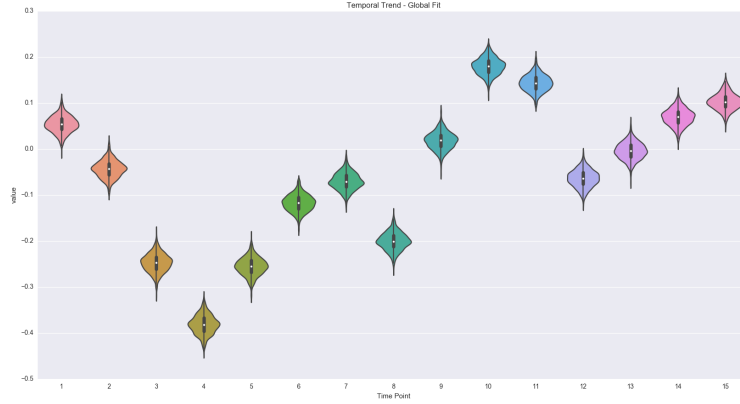
The variance hyperparameters are given (insert prior here) as recommended in (possibly Gelman 2006).

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

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

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 pathological 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.

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 inseparable 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 hierarchical 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

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

where, as before,

$$v_{1:N} \sim \text{CAR}(W, \sigma_v) \quad (18)$$

$$\lambda_{1:N} \sim \text{Normal}(v, \sigma_\lambda) \quad (19)$$

$$\gamma_{1:T} \sim \text{RW}(1). \quad (20)$$

We see the new component ψ captures a level of space-time inseparability 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,

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

where

$$z_i \sim \text{Bernoulli}(q) \quad (22)$$

and q is given a Uniform prior on $(0, 1)$. 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 inseparable variance of the ‘normal’ regions to be very limited. Identification issues are avoided by defining the larger of the variances additively in terms of the smaller:

$$\tau_1 \sim \text{Normal}(0, 0.01) \cdot I(0, \infty) \quad (23)$$

$$k \sim \text{Normal}(0, 100) \cdot I(0, \infty) \quad (24)$$

$$\tau_2 = \tau_1 + k \quad (25)$$

8 Computational considerations

8.1 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/ initialization with ADVI

8.1.1 Gibbs sampling

8.1.2 Metropolis

8.1.3 Hamiltonian Monte-Carlo

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

8.1.4 Autodiff/black-box variational inference?

8.2 Reparamterization of the models

Removing conditional dependencies e.g.

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

8.3 Marginilizing over the mixture component

- Rao-Blackwellization?

8.4 Timing data?