Detecting pathologies in spatiotemporal data

Michael Redman, CID: 00826863

December 2016

Contents

Chapter 1

Introduction

The identification of unusual patterns of disease is of obvious importance in public health. Discrepencies in the relative risk of disease incidence between regions will both inform the allocation of medical resources and motivate investigations over possible environmental exposure.

The detection of such discrepencies is often done by the use of disease mapping studies which aggregate the recorded incidents of the disease of interest over time, and compare the total counts between regions. Public health information systems now record this data indexed to such a fine degree of granularity that the detection of risk factors which act at a very local level, such as hazardous industrial waste, can be feasably indentified. However this fine-grained approach will leave the examined regions with very low incidence, which can hinder the isolation of the signal from the noise, hence the use of time aggregated values. However, in doing this we relinquish any potential information that a temporal trend may afford us. In fact, as described in Abellan et al. (2008), the nature of the temporal trend may suggest anomalities related to very different kinds risk factors. A region which conforms to a similar temporal trend as the others—for example a seasonal factor—but at a higher rate would imply a disparity which acts over the whole time period, such as an enviromental issue or a sociodemographic difference. While a region which acts wildly without regard to the general trend might suggest a more acute issue is at play.

Consequently, in this project we will be looking at data that is indexed in both space and time. Additionally in any model that we implement, we wish to be able to control for known covariates which may influence the risk within a given region. For example some regions will be more populous, while others may contain proportionally more at-risk peoples, so it would be wise to have some method of setting the expected counts by region.

A variety of methods have been investigated in these contexts. One of the most simple is that of scan statistics. This procedure, implemented in software such as SaTScan, seeks

to identify clusters which can not be explained by the base process – in our case typically an inhomogenous Poisson process. It does this by evaluating all possible "windows" in turn, where a given window will include the region in question and its "neighbours" both spatially and temporally. This forms a kind of cylinder which acts to smooth the observed counts in space and time, allowing the sparse data as described earlier to be examined without the noise overwhelming any underlying spatial heterogenicity. This method is descibed in detail and applied to epidemiological data in Kulldorff (1997). While offering an improvement over pointwise evaluation, spatial scan statistics by nature of this smoothing construction are limited in their ability to detect abnormalities at the individual level, i.e. isolated to a particular region (Li et al. 2012).

Control chart based methods offer both relatively fast computation times and the ability to specifically identify abnormal temporal trends. These, in the simplest of terms, are graphs we construct combined with a decision rule to classify whether a process is in 'in-control' or 'out-of-control'. Versions such as CUSUM are often designed for step detection – the identification of abrupt jumps in rate of the process. When looking at longer time periods, this is exactly the type of pathology one might expect. The regions which now show atypical trends should still have exhibited normal behaviour most of the time, the problem is to decide when such heterogeneity in the process is substantive. Another advantage is, as the techinique is sequential in nature, the sample size of the items in consideration need not be fixed in advance.

The most natural way to frame our problem is by the construction of a hierachical model. In this method we perform inference on a latent variable—within a complex model—which indicates the 'unusalness' of a point or group of points. This Bayesian approach compells us to express our assumptions about the data generatively, which confers a number of advantages. Firstly, it is often the case in discriminative models that the choice of method is implicitly performing shrinkage, without this being made explicit or even realized by the modeler. For an intereting example see Mandt et al. (2017). Thinking generatively forces us to be clear about our assumptions. Secondly, real world epidemiological data will presumably not supply the true response variable, that is, whether or not a region is genuinely anomalous. This abscence of training means we need our model to provide reasonable estimates immediately, rather than at some asymptotic limit. This advantage can be seen in, for example, Ng et al. (2002). Thirdly, the parameters of a Bayesian model are far more readily interpretable that those in a non-generative setting, which could allow for further inference beyond the identification of abnormal regions, e.g. the nature and extent of the deviation from what was expected. And lastly, the Bayesian framework naturally allows the addition of futher assumptions and complexity. For example, Banerjee et al. (2003) extends spatiotemporal models to the modelling of multiple diseases via the introduction of a shared 'frailty' term.

The examination, implementation and comparison of these Bayesian methods with discriminative models, specifically control charts, is the focus of this project.

Working on simulated data, we first implement a CUSUM based classifier and suggest a novel(?) specification of the 'in-control' process. With the efficiency of the method being of great importance, we implement the computations from scratch in Fortran rather than modifying the existing statistical routines in R.

Then, turning to Bayesian methods, we examine the general structure of the models employed in the literature and how this standard is well suited to the particular challenges of our problem. In particular, the favored priors for the smoothing of the data are reviewed. The difficulties and subtleties of computaion are addressed in detail, as are some of the superior techniques that have been developed in the past few years.

We examine BaySTDetect, one of the existing models in the literature, discussing the rationales behind the construction of the model and the 'cutting of feedback' that's employed – where we demonstrate that this particular choice is fundamental to it's success. Rather than use the standard MCMC software in the field, we implement the model in Stan using Hamiltonial Monte Carlo and combine the submodels directly through the likelihoods. We show that this can lead to order of magnitude speed-ups over the original implementation in Li et al. (2012).

We then suggest a new model, based on Abellan et al. (2008), that works fully within the Bayesian framework by identifying an excess in the variance which can not be explained by the shared random effects. Implementing this model, again in Stan, we show that this model obtains good results, despite not being the way in which our simulated data were originally created. However, as suggested in Best et al. (2005), models of this kind may oversmooth the counts spatially. We demonstrate this by simulating a new set of data where the regions to be 'unusual' are selected sequentially via a preferential attachment scheme and showing that our model performs poorly in this context.

Finally we compare the use three models on both sets of data with respect to their speed, false postive rate and power.

Chapter 2

Simulated data

In order to test the accuracy of any potential models we need some data that exhibits the type of pathological trends that we wish to identify. The following data was graciously supplied by Areti Boulieri.

2.1 Specification

A real dataset of asthma hospitalisation counts was used to contruct the expected rates, adjusted by regional demographics by age and sex, for the 221 Clinical Commisioning Groups in England. These are the expected counts that we will standardize against when attempting to identify abnormalities. We denote these values throughout by E_i, for $i = 1, \ldots, 221$. A plot that we generated to show the distribution of these values across the regions can be seen in fig **??**

A binomial random variable might be the most principled choice of response but the expected counts we have calculated are large enough for using a Poisson response to be essentially equivalent. Indeed this is the most popular choice in the literature. The data was generated across 15 time points, so our values are drawn from

$$Y_{i,t} \sim \text{Poisson}(E_i \cdot \mu_{i,t}), \quad i = 1, \ldots, 221 \quad t = 1, \ldots, 15 \tag{2.1}$$

where $\mu_{i,t}$ allows us to add random effects to our model.

These random effects are used to add additional natural variation to the counts, where the direction of this variation is in some sense shared between 'similar' regions or time points. We seperate these effects into those that describe the temporal trend and those which affect the spatial heterogeneity. So for the temporal trend we of course expect adjacent time points to be similar. This is achieved in this data by sampling from a one-dimensional Gaussian random walk

$$\xi_{1:15} \sim \text{RW}_1(\sigma_\xi^2). \tag{2.2}$$

Additionally, we would expect regions which are close to each other to be similar. The data generation incorporated this assumption by the use of the BYM prior, first defined in Besag et al. (n.d.),

$$\lambda_{1:211} \sim \text{BYM}(W, \sigma_\lambda^2, \sigma_v^2) \tag{2.3}$$

which is a multivariate normal distribution

$$\lambda_{1:211} \sim \mathcal{N}(v_{1:211}, I_{211} \cdot \sigma_\lambda^2) \tag{2.4}$$

where the mean vector is drawn from an *intrinsically autoregressive* process

$$v_{1:211} \sim \text{IAR}(W, \sigma_v^2) \tag{2.5}$$

which is a Markov random field where areas that are adjacent (according to an adjacency matrix W) are more similar to those further away. The precise specification of the BYM prior is discussed in section 5.1.

The presumed sources of these random effects will determine how we wish for them to influence the generation of the data. In this dataset it is assumed that the temporal and spatial effects will work multiplicatively, where increases in one will magnify the effect of the other. Therefore, given our samples from the aformentioned distributions, we calculate $\mu_{i,t}$ additively on the log scale

$$\log(\mu_{i,t}) = \lambda_i + \xi_t \quad i = 1, \ldots, 221 \quad t = 1, \ldots, 15 \tag{2.6}$$

In total, this scheme describes a method of generating random variables that has the features one would expect of spatiotemporal disease incidence. We now need to modify a selection of the regions to resemble a regions we would expect to be unusual.

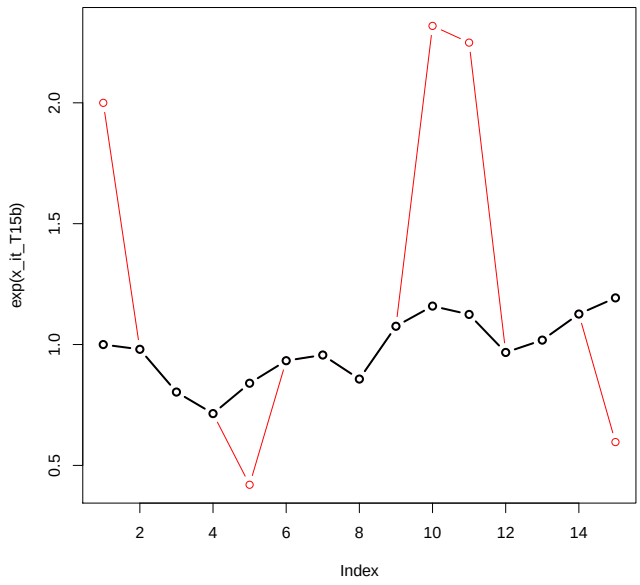
Fifteen of the regions were chosen according to the 10th, 25th, 50th, 75th and 90th percentiles of the median expected counts over time, with three regions per percentile. At each percentile the three regions were selected corresponding to one of three levels of spatial risk, i.e. the generated spatial effect $\lambda_{1:211}$, low (10th-30th percentiles), medium (45th-55th percentile) and high (70th-90th percentiles). This was done to ensure that the unusual regions were spread evenly given the underlying levels of risk.

To make these regions unusual they were given a deviant temporal trend. This was done by modifying the temporal trends of these regions manually, where—writing the original trend as $g(t)$ and the modified trend as $g^*(t)$—we have:

$$g^*(t) = g(t) + \log(2) \qquad t = 1, 10, 11 \tag{2.7}$$

$$g^*(t) = g(t) - \log(2) \qquad t = 5, 15 \tag{2.8}$$

This has the effect of either doubling or halving the expected counts depending on the time point. The resulting temporal trend compared to the original can be seen in figure



??.

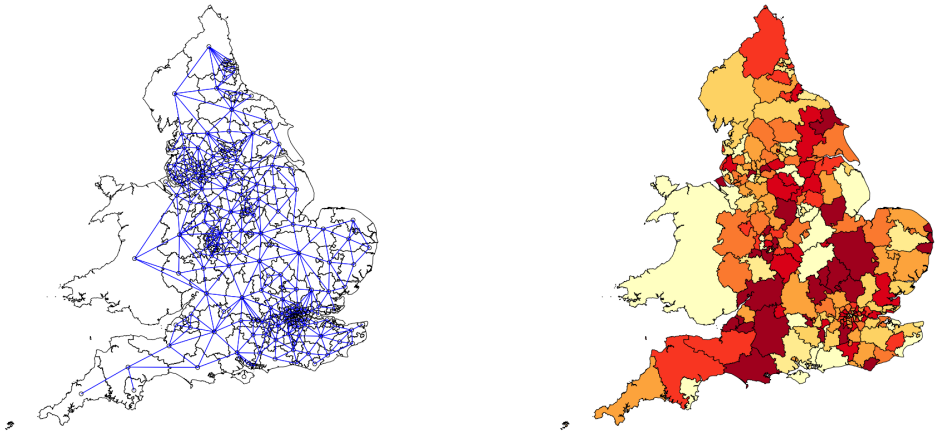
N.B. All of the unusual regions exhibit the same deviation, which is probably an unrealistic assumption. However, by the way that they're defined, it is clear that the models used in this project should not perform better because of this fact.

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

2.2 Cleaning and wrangling

Of the 211 regions only one one had no neighbours, the Isle of Wight. As this will complicate some of the smoothing calculations we need to perform for the Bayesian procedures (and the Isle of Wight was not one of those selected as unusual), the region was removed to simplify calculations.
Shape data for the 211 Clinical Commisioning Groups plus wales was provided by CITE HERE. The regions of Wales and the Isle of Wight were of course removed. The shapefiles were imported and edited using the package *rgdal* created by Bivand et al. (2015). Then, using the package *spdep* created by Bivand et al. (2013), the shapefile was used to generate an object which calculates all of the adjacent regions for each region. Then this object was converted into into an adjacency matrix W such that

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



(a) Adjacency (b) Expected

Figure 2.1: Mapped data

where the double arrow signifies adjacency. The result of this procedure can be seen over the original shapedata in fig **??**.

Chapter 3

Control charts

The use of control charts are widespead in the area of statistical quality control and have recently increasingly found use in fields such as epidemiology, see for example Mei et al. (2011). Their relative simplicity and ease of computation compared to most other methods makes them an ideal model to compare against, in order to observe any gains we might obtain from more complex modelling. Indeed, as expressed by Vapnik et al. (1998) in the context of generative vs discriminative models, "one should solve the problem directly and never solve a more general problem as an intermediate step".

3.1 CUSUM

We will specifically be looking at the use of CUSUM models, which seek to identify a qualitative change in the nature of a time series via the evaluation of the cumulative sum of a relavant test statistic. In the setting of the inhomogenous Poisson CUSUM, we assume that the data follows a null model defined by a series of rate parameters which define the 'in-control' Poisson process. Likewise we have a complementary set of parameters that define the rates of an 'out-of-control' process. We wish to detect the point in time at which there is sufficient evidence that the process was switched from the null to the 'out-of-control' process, as quickly as possible while not exceeding some measure of how often we err.

To do this, define S_t as the value of the control chart at time t and progress its value as follows

$$S_0 = 0 \tag{3.1}$$

$$S_{t+1} = \max(0, S_t + K_{t+1}). \tag{3.2}$$

where K_t is the value of a statistic calculated at time point t.

As first proposed in Page (1954), we will use

$$K_t = \log\left(\frac{f_1(Y_{i,t})}{f_0(Y_{i,t})}\right) \tag{3.3}$$

$$= \log\left[f_1(Y_{i,t})\right] - \log\left[f_0(Y_{i,t})\right]. \tag{3.4}$$

where f_0 is the likelihood of the 'in-control' rate and f_1 that of the 'out-of-control'.

That is, the log of the ratio between the likelihoods of the data, at the time point, under the 'in-control' and 'out-of-control' Poisson process. So if likelihood of the 'out-of-control' rate is greater than the 'in-control' then the value of S_t will increase and vice-versa. We then signal that the process is pathological if the value of S_t exceeds some threshold value h.

Note that, for a Poisson process, the likelihood is

$$f_j(x) = \frac{\lambda_j^x e^{-\lambda}}{x!}. \tag{3.5}$$

So for the log-likelihood we have

$$\ell_j(x) = x\log(\lambda_j) - \lambda_j - \log(x!) \tag{3.6}$$

giving the value of our test statistic as

$$K_t = Y_{i,t}(\log\lambda_{1,t} - \log\lambda_{0,t}) - (\lambda_{1,t} - \lambda_{0,t}). \tag{3.7}$$

This leaves us to come up with sensible methods of generating the 'in-control' and 'out-of-control' rates and the threshold value that we trigger at. With the data that we are examining in this project we have the expected values of the disease rates in each region and could use these for the values of the 'in-control' rate but this would neglect the fact that we expect the incident counts at each region to vary over time and—importantly—this is not neccasarily pathological behaviour. It is instead a temporal trend which differs from some general (here country-wide) trend which we wish to detect. Therefore the method we propose here is to construct a general temporal trend from an average across the regions and to weight this trend per region by that regions expected counts.

So for the data $Y_{i,t},\ i = 1,\ldots,R,\ t = 1,\ldots,T$ we normalize the temporal pattern of each regions by its expect count E_i as follows

$$\tilde{Y}_{i,t} = \frac{Y_{i,t}}{E_i}, \qquad i = 1,\ldots,R,\ t = 1,\ldots,T. \tag{3.8}$$

We then construct the 'in-control' rate for each region $I_{i,t}$ as described

$$I_{i,t} = E_i \cdot \frac{1}{R}\sum_{i=1}^{N}\tilde{Y}_{i,t}, \qquad i = 1,\ldots,R,\ t = 1,\ldots,T. \tag{3.9}$$

For simplicity we only consider the case of abnormally high count rates and ignore artificially low counts although this could be incorporated into the CUSUM model if desired. So to define the 'out-of-control' rate it makes sense to follow the same temporal trend but with a modified general rate. We do this by calculating the 'out-of-control' rate $O_{i,t}$ as a multiple of the 'in-control' as follows

$$O_{i,t} = \alpha \cdot I_{i,t}, \quad i = 1, \ldots, R, \ t = 1, \ldots, T. \tag{3.10}$$

with $\alpha > 1$.

Now for a given value of α we need a threshold level at which to signal. This can be done by caluclating the average-run length (ARL) which is the expected length of the series between flags of the model. The ARL on the 'in-control' rate will then give a measure of the false positives for a given threshold h, which we can optimize with respect to. As the length of the series we are examining is fixed and the calculation of an ARL would not be clearly defined for a series of varying expectation we will not use the ARL metric here. Instead we simply simulate the 'in-control' sequence for each region a large number of times and find the smallest value of h for which the the false-postive rate is below some desired rate.

The setup as described was implemented as a Python class acting as a wrapper to a series of subroutines in Fortran via F2PY[1]. The code can be found in the appendix.

This CUSUM process was applied to the asthma data with the ratio α set to 1.5, indicating we consider the "out-of-control" rate to be 50% greater than the "in-control", with each regions temporal pattern being simulated 10,000 times to determine the optimal threshold values. This identified 25 regions as "out-of-control" out of the total 210, giving a false-negative rate of 0% and false-positive rate of 5.13%.

The power of the test is clearly excellent, but the false-positive rate is clearly much greater than our desired limit of 1%. Why is this?

Note that the expected value of K_t under the 'out-of-control' model (H_1) will be the Kullback-Leibler divergence between the two models (Jiang et al. 2011),

$$\begin{aligned}
\mathbb{E}[K_t|H_1] &= \int_{-\infty}^{\infty} f_1(x_t) K_t \, dx_t \\
&= \int_{-\infty}^{\infty} f_1(x_t) \log\left(\frac{f_1(x_t)}{f_0(x_t)}\right) dx_t \\
&= D_{\mathrm{KL}}(H_1 \| H_0).
\end{aligned}$$

This gives an intuitive explanation of the procedure: under the alternate ('out-of-control') model, K_t measures the information lost by attempting to approximate the

1. I originally wrote the entire process in Python but the calculation of the h values was far too slow to be practical. The original code can be on my Github.

(true) alternate model by the null. When the cumulative information lost by this approximation becomes intolerable, we reject the null.

It can be shown that, under some assumptions, using this metric with the CUSUM method is in some sense optimal (Ritov 1990). However, the issues with the assumptions here are clear. The null model as we have defined attempts to incorporate the general temporal trend but we would also expect a level of random spaital variation which masks the underlying abnormalities and our procedure ignores this. Therefore for regions which are not unusual (call these H_G), the null model will underestimate the likelihood ($f_0(x_t)$) of their counts relative to the true likelihood $f_G(x_t)$. So, under H_G one would expect $\log(f_1(x_t)) - \log(f_0(x_t))$ to be large more often than under the null, underestimating the value of h which would be required to keep the FPR under 1%.

Additionally, even if the true data is distributed as we assume in the null, the parameters of the temporal trend that defines the 'in-control' models are still an estimation from the data. The errors between this hypothetical 'true' temporal trend and our estimate from the data will add additional noise into the process. In particular, this will affect the estimation of the optimal value of h for a given FPR – potentially increasing the number of false positives relative to what we might otherwise expect.

A number of methods exist that attempt to account for this additional estimation error. For example, following the method devised in Gandy et al. (2013), we could estimate the 'in-control' parameters as usual and then draw samples with replacement from the 'in-control' model with these parameters. Then, calculating estimates of the 'in-control' parameters from these samples, we obtain a series of bootstrap estimators. The optimal values of h for each of these bootstrap estimators can be used to devise a threshold which posseses our desired properties with a certain probability. Of course, the data from which we estimate the temporal trend is not all 'in-control', so our estimates will be even more biased than this would suggest.

3.2 Spatial CUSUM

While our method accounts for a varying temporal trend, it falls short most saliently by not accounting for the spatial effects we assume our data posseses. The approach suggested by Raubertas (1989) attempts to take the spatial information into account by replacing the counts at each region by a weighted sum

$$x'_{i,t} = \sum_{j=1}^{R} s_{i,j} x_{j,t}$$

where we pool within neighbourhoods by assigning the weights based on the 'closeness' of the regions.

For example this method was used in Dassanayaka (2014) with the Poisson CUSUM, where the counts at each region were replaced by

$$x'_{i,t} = \sum_{x_{j,t} \in \partial x_{i,t}} x_{j,t}$$

where $\partial x_{i,t}$ are all the regions including and adjacent to $x_{i,t}$.

The Poisson CUSUM statistics were calculated for these modified regions and the Benjamin-Hochberg procedure was used to determine the level at which we signal for a given false discovery rate. Despite this methods simplicity, the author uses a set of simulated data to show that it can provide a significant improvement on considering the regions individually.

This approach could be used in conjunction with that of the previous section to account for both spatial and temporal effects, while still enjoying the benefits of CUSUM. Note that, in this case, the expected counts would also need to be summed across neighbourhoods to provide the 'in-control' means.

Chapter 4

The Bayesian Framework

A sensible way of improving on this would be to use a generative model which incorporates both a flexibility that describes our uncertainty around the parameterization of the model and makes explicit our beliefs as to the structure of the data. This is most naturally done in the Bayesian framework, where we specify the model in terms of a likelihood which describes how we simulate the data based on a set of parameters and a collection of—possibly hierarchical—priors which describe our uncertainty about these parameters ex ante.

1. Advantages of generative models
2. Advantages over frequentist approaches, e.g. mean/mode often terrible approximations, in part due to concentration of measure
3. Level of shrinkage attributable to closeness of prior structure to some "true prior"
4. Bayes formula
5. Difficulty in evaluating integrals/expectations
6. Samples will converge to expectation (use betancourts conceptual into here)
7. Proposal density and preserving the pdf
8. Metropolis-hastings
9. Theory and issues including: large/small step sizes, large autocorrelation, slow exploration of the sample space, divergences
10. Gibbs sampleing and blocked gibbs: overcomes some element of dimensionality but poor for highly correlated variables, compounded by issues with
11. Asymptotics not followed in finite time often
12. Quick talk about existing software

13. Hamiltonian monte carlo provides some solutions to many of these problems
14. In last few years (get exact) emergence of better algorithms, tuning and software have made using HMC practical. Not used in epidemiology much yet though. Different fields different amounts due to issues with discrete paramters etc. that we will address later.
15. Convergence statistics

- Space-time seperability
- Identification issuses in mixture model
- Bayesian model selection
- Prior on mixture component probability
- Bayesian classification

4.0.1 Hyperparameter priors

- Gelman 2006 for variance parameters
- Stan wiki for others

4.0.2 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"

4.0.3 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.

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

$$p(\theta) = \frac{1}{Z} e^{-E(\theta)}. \tag{4.1}$$

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:

$$H(p,q) = \underbrace{T(p|q)}_{\text{Kinetic energy}} + \underbrace{V(q)}_{\text{Potential energy}} \tag{4.2}$$

looking at the canonical ensemble of the join distribution

$$\begin{aligned} \pi(p,q) &\propto e^{-H(p,q)} && (4.3)\\ &= e^{-[T(p|q)+V(q)]} && (4.4)\\ &= e^{-T(p|q)} \cdot e^{-V(q)} && (4.5)\\ &\propto \pi(p|q) \cdot \pi(q) && (4.6) \end{aligned}$$

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

$$V(q) = -\log \pi(q) \tag{4.7}$$

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

$$\begin{aligned} \frac{dq}{dt} &= \frac{\partial H}{\partial p} = \frac{\partial T}{\partial p} && (4.8)\\ \frac{dp}{dt} &= -\frac{\partial H}{\partial q} = -\frac{\partial T}{\partial q} - \frac{\partial V}{\partial q} && (4.9) \end{aligned}$$

- Finish off this bit more gracefully

The Method

Hamiltonian monte-carlo works, not by applying a transition over q_i, 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? (Leapfrog)
- 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:

- *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?
- Parameter tuning vs burn-in period

Other things to talk about:

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

4.0.4 General Model Form

- Use best et al to summarise basic model form

Chapter 5

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 Crainiceanu et al. (2007) but in this project we will primarily looking at the use of autoregressive priors. For a comparison of these two approaches see Goicoa et al. 2012.

5.1 CAR models

One of the most natural ways of describing the proximity of regional data is by their adjacency relationships. These can be described mathematically as an undirected graph where the edges connect those regions which share a border. When the probabilistic dependence of the regions on this graph extent only to those that share an edge (i.e. the Markov property), then we call the network a Markov random field.

In the Bayesian setting, the most popular way of modelling this interdependence is through the use of a conditionally autoregressive (CAR) prior. These CAR models can be best understood when specified in terms in terms of their conditional distribution

$$v_i \mid v_j \; j \neq i \sim N(\alpha \cdot \bar{\mu_i}, \; \sigma_v^2/k_i) \tag{5.1}$$

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

$$\bar{\mu_i} = \sum_{j \in \partial i} \frac{\mu_j}{k_i} \tag{5.2}$$

and $\alpha \in (0, 1)$ is a parameter measuring the degree of spatial dependence.

However, while this defines a Markov random field, it is not a directed acyclic graph and we can't use this definition in non-gibbs sampling methods – we need the v_i to be jointly specified. Thankfully it is possible (Banerjee et al. 2014) for this to be expressed in terms of a multivariate normal distribution as follows,

$$v \sim N(0,\ \sigma_v^2 \cdot [D(I_n - \alpha B)]^{-1}) \tag{5.3}$$

where

$$D = \text{diag}(k_i) \tag{5.4}$$

$$B = D^{-1}W \tag{5.5}$$

$$(W)_{ij} = \begin{cases} 1, & i \leftrightarrow j \\ 0, & \text{otherwise} \end{cases} \tag{5.6}$$

Note that it is intuitively clear that the precision matrix here in high dimensions will be 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.

5.1.1 Intrinsically autoregressive priors

When we take the CAR distribution under the limiting case of $\alpha \to 1$ we obtain the intrinsically autoregressive prior (ICAR/IAR). The ICAR prior is very popular throughout disease mapping, due to its use in the BYM prior described in the next section, but its important to note that it does not define a proper probability distribution as the integral of 5.3 with $\alpha = 1$ over the real line is infinite. Indeed the precision matrix

$$\tau = \sigma_v^2 \cdot [D - W] \tag{5.7}$$

is singular.

Not only this, but the mean is also undefined. However, the density function can still be evaluated pointwise (up to a constant), and if we constrain the variates to sum-to-zero then a posterior distribution using it as a prior should still be well defined.

As an aside, the conditional distributions of each region in the intrinsic CAR are given by

$$f(v_i|\kappa) \propto \exp - \sum_{j \sim i} \omega_{ij} \phi(v_i - v_j) \tag{5.8}$$

where

$$\phi(z) = \frac{z^2}{2\kappa}.$$

Although, as discussed in Besag et al. (n.d.), the form of 5.8 with other choices of $\phi(z)$ may be appropriate in different situations. For example if discontinuities in the risk surface are expected then

$$\phi(z) = \frac{|z|}{\kappa}$$

would provide shrinkage consistent with that belief.

5.1.2 BYM prior

In reality, the parameter α in 5.3 will often be close to its maximum value to confer even a moderate amount of spatial dependence (Lunn et al. 2012). Therefore, the prior first suggested in Besag et al. (n.d.), often now refered to as the BYM prior after its authors, is widely employed in disease mapping. The prior consists, first, of an intrisic CAR variable

$$v \sim \text{IAR}(W, \sigma_v^2)$$

which captures the spatial dependence between regions. However, this imposes the spatial structure on v a priori and there is no 'Bayesian learning' about the strength of the dependence. Therefore, in the BYM prior a second, unstructured random effect, is added as a degree of freedom which captures the effect of any unobserved unstructured latent variables (Lunn et al. 2012).

So the base level of the BYM prior takes the form

$$\lambda \sim \mathcal{N}(v, I_n \cdot \sigma_\lambda^2) \tag{5.9}$$

where I_n is the identity matrix of size n.

As noted in the *BUGS book*, one of these two effects will tend to dominate the other in the posterior. Although we don't know which until the model is fit.

In this project, the BYM prior is the preferred to capture the underlying spatial effects but in practice we substitute the first layer ICAR for the non-intrinsic CAR model with a reasonably high value of α. Computation of the ICAR model *is* possible is Stan (Joseph 2016) but, due to the relative recentness of its implementation, we avoid it *just in case* there are any unknown computational irregularities with respect to HMC.

This alterated prior, in Stan, was tested against the standard implementation, in BUGS, on the classic Scottish lip cancer dataset from Clayton et al. (1987) and the results were essentially identical.

5.2 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

$$\xi_{1:T} \sim \mathrm{RW}_1(\sigma_\xi^2) \tag{5.10}$$

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.

Explicitly, this is defined as

$$\xi_1 \; N(0, \sigma_\xi^2) \xi_{t+1} \; N(\xi_t, \sigma_\xi^2), \quad t = 2, \ldots, T. \tag{5.11}$$

Chapter 6

Individual trend model

The first Bayesian model we wil consider is the BaySTDetect model, first introduced in Li et al. (2012). In this setting, we construct two competing 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 CAR prior), and another where the region is free to vary with 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 – labeling these regions 'unusual'.

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, with model selection undertaken afterward. This method, which prevents the flow of information between the two models, has been met with some level of skepticism (Gelman 2013) as the analysis is not 'truly Bayesian'. Nethertheless, we examine this paradigm and compare it to the fully Bayesian methods.

6.1 Model specification

Proceeding as in Li et al. (2012), we denote the counts at region i at time t by $Y_{i,t}$ and model them by a Poisson process

$$Y_{i,t} \sim \text{Poisson}(E_{i,t} \cdot \mu_{i,t}) \tag{6.1}$$

where $E_{i,t}$ are the expected counts by region 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

$$\log(\mu_{i,t}) = \begin{cases} \lambda_i + \gamma_t \ (+ \ \alpha_0) & \text{Model 1 for all } i,t \\ u_i + \xi_{i,t} & \text{Model 2 for all } i,t \end{cases} \tag{6.2}$$

For the first model, *Model 1* we split any heterogeneity in spatial and temporal components, assuming that the variation is at least mostly seperable. The intercept term is

given an uninformative (ill-defined) flat prior over the real line

$$\alpha_0 \sim \text{Flat}(\mathbb{R}). \tag{6.3}$$

The spatial component is given a BYM prior as described in 5.1.2,

$$v_{1:N} \sim \text{CAR}(W, \sigma_v) \tag{6.4}$$
$$\lambda_{1:N} \sim \mathcal{N}(v, I_N \cdot \sigma_\lambda). \tag{6.5}$$

With the temporal component following a 1D random walk,

$$\gamma_{1:T} \sim \text{RW}_1(\sigma_\gamma^2)$$

Here our BYM prior on the spatial component imposes a smoothing constraint which allows adjacent regions to share information. This helps to combat the issues, described in Chapter 1, raised by any sparse counts at the individual level. Likewise, the random walk prior captures the general temporal variation that we believe is typical/non-anomylous.

The variance hyperparameters are given half-normal, weakly informative, priors

$$\sigma_v, \sigma_\lambda, \sigma_\gamma \sim N(0,1) \cdot I(\sigma_k > 0) \tag{6.6}$$

as recomended in Gelman (2006).

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

$$\xi_{i,1:T} \sim \text{RW}_1(\sigma_{\xi,i}^2), \quad i = 1, \ldots, R$$

and an individual intercept term with an uninformative prior

$$u_i \sim \text{N}(0, 1000), \quad i = 1, \ldots, R.$$

The temporal trends for each region are taken to have individual variance parameters, allowing for greater diversity in the forms of abnormal trends detected. These variance parameters are given log-normal priors

$$\log(\sigma_{\xi,i}^2) \sim N(a, b^2), \quad i = 1, \ldots, R$$

where a and b^2 are defined hierarchically

$$a \sim N(0, 1000)$$
$$b \sim N(0, 2.5^2) \cdot I(b > 0).$$

This prior structure represents the belief that 10% of the regions will have local temporal variance 10 times greater than the median. That is,

$$P(\sigma_{i,\xi}^2 > 10 \cdot \exp a) \approx 0.1.$$

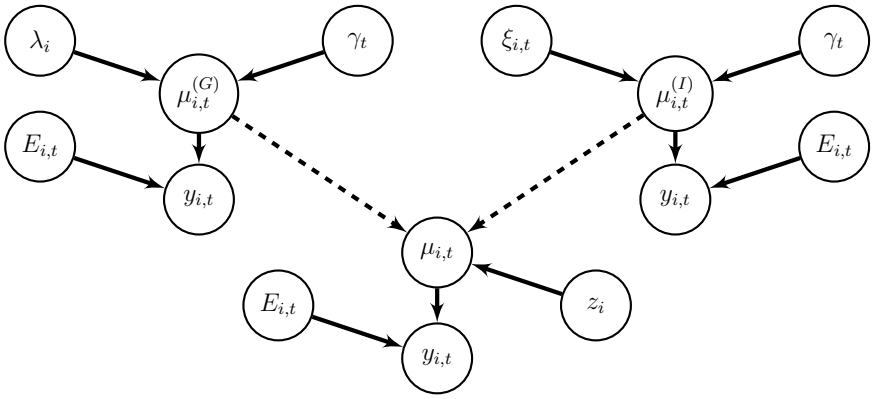


Figure 6.1: Directed acyclic graph showing conditional dependencies between model variables for the BaySTDetect model. Dashed lines indicte one way flow of information. Based on figure in Li et al. (2012).

Finally these two models are combined via the indicator

$$z_i \sim \text{Bernoulli}(0.95) \tag{6.7}$$

which is a latent variable that assigns a prior probability of 5% that a given region is 'unusual'.

It is important to note that this is **not** a mixture model in the traditional sense. The two models are fit *seperately* and only *then* are they combined

$$\mu_{i,t} = z_i \cdot \mu_{i,t}^{(G)} + (1 - z_i) \cdot \mu_{i,t}^{(I)} \tag{6.8}$$

where $\mu_{i,t}^{(G)}$ is the rate parameter implied by Model 1 (the general model) and $\mu_{i,t}^{(I)}$ is that which is implied by Model 2 (the individual trend model). Here, information is not allowed to flow back from this combination back to the model fits themselves. Graphically this can be represented by the Bayesian network in fig **??** which shows the conditional dependencies between variables in the model. Note that we essentially copy the data for the fitting of the two seperate models and then the final inference step.

6.2 Combining the models

The original computational specification of BaySTDetect was in *BUGS*, using the cut function[1]. The cut function, essentially, makes a copy of the model structure and data *downstream* from the variable and uses standard Gibbs to produce samples of the variable, isolated from the rest of the model. The samples of this variable are then used

1. The original BUGS code can be difficult to find on the Journal website: a copy can be found on my Github under /model/bugs/

downstream in the model as specified but are treated as though they came from some fixed distribution.

However, this implementation can be improved upon in a number of ways. Firstly, as found by (Plummer 2015), the method that is applied in the BUGS software will not converge to a well defined limiting distribution, although there may exist modifications to the algorithm that resolve this. Secondly, as described in section 2545, Gibbs sampling will be very slow to converge to true result – where Hamiltonian Monte-Carlo may improve on this. And thirdly, the base level of the model is simply a mixture on the rate component of a Poisson random variable – this can be simplified to a direct calculation by hand.

In our implementation we first fit the two models individually in Stan, the code can for this be found in codeblahref. Four chains, with dispersed starting positions, were run for x iterations each for model 1 and y iterations for model 2. Where these numbers was chosen to be a short chain length where the $\hat{R}$ values (the Gelman-Rubin statistic) were consistently < 1.01 over multiple runs, indicating good convergence. Additionally a visual inspection of the trace plots for some select parameters did not indicate any worrying pathalogical behaviour.

We then extracted a sample (without replacement) of the $\mu^{(G)}$ and $\mu^{(I)}$ from the chains and used these to contruct two models

$$P(y_{i,1:T}|M_1, \mu^{(G)}) = \prod_{t=1}^{T} \frac{\left(\mu_{i,t}^{(G)}\right)^{y_{i,t}} \cdot \exp\left\{-\mu_{i,t}^{(G)}\right\}}{y_{i,t}!} \tag{6.9}$$

$$P(y_{i,1:T}|M_2, \mu^{(I)}) = \prod_{t=1}^{T} \frac{\left(\mu_{i,t}^{(I)}\right)^{y_{i,t}} \cdot \exp\left\{-\mu_{i,t}^{(I)}\right\}}{y_{i,t}!}. \tag{6.10}$$

That is, the likelihood of the data at each region under each model is that of a series of independent Poisson variables where the rate parameters are decided by one of the samples from either the general or independent trend model.

We calculate both these likelihoods, for each region, for every sample of the μ. Then, using these likelihoods we can perform Bayesian model comparison as

$$P(M_j|y_{i,1:T}) = \frac{P(y_{i,1:T}|M_2)P(M_2)}{P(y_{i,1:T})} \tag{6.11}$$

where our prior for the probability of each model being correct follows from the definition of z_i as

$$P(M_1) = 0.95 \quad P(M_2) = 0.05 \tag{6.12}$$

We can then, ignoring the denominator as it is the same for both models, compute the unnormalised posteriors for the two models. Comparing these two values, we can assign

an indicator to the z_i

$$z_i = \begin{cases} 1, & \text{if } 0.95 \cdot P(y_{i,1:T}|M_1) > 0.05 \cdot P(y_{i,1:T}|M_2) \\ 0, & \text{otherwise} \end{cases} \tag{6.13}$$

So, z_i is 1 if the posterior probability of the general model is greater than that of the specific trend model. Finally, we an average of the z_i over the samples and call this $\bar{z}_i$.

To test if this procedure works as expected, we take the samples from our *Stan* chains and use them in a BUGS model with the true Bernoulli indicator. Taking the MCMC based posterior mean of the z_i in this gave very similar values to our method.

6.3 Classification and accuracy

We can use the $\bar{z}_i$ values to classify regions as either 'normal' and 'unusual'. We could simply say, any regions which are more likely to be 'unusual' are classed as such. However given we are performing this for 210 regions we run into issues of multiple comparison. Therefore, Li et al. (2012) uses the Benjamini-Hochberg procedure to find at which level probability we should classify all with lower probabilities as being 'unusual'. This procedure ensures, under some assumptions, that we keep the false discovery rate (FDR) under a given value.

With the values we found, explicitly implementing this procedure was not neccesary as a set of 16 regions had z_i values of 0 while the others had much higher values. Of these 16 regions, 15 were the correctly identified unsual regions. Giving a power of 100% and an FDR of about 0.51%.
As a sanity check, we look at the samples of γ_t from model 1 to see if it approximates the true general temporal trend. A violin plot of the samples can be seen in fig **??**. Comparing this trend to that in fig **??**

6.4 Fully Bayesian model

The BaySTDetect method, as originally applied is clearly very accurate in its identification of the 'unusual' regions. Though, the use of the cut function is puzzling. The model, when specified as a true mixture model, essentially describes how the data was generated originally. To test this, we combined the two models as specified without the cut function and fit the entire model in *Stan*. This was done by specifying the probability of the Bernoulli as the Beta prior with mean 0.90 – allowing the use of HMC and gaining computational efficiency as described in section 73826.

The resulting values of q_i, however, was barely more accurate than random in ordering the probabilities of each region being 'unusual'. In fact, looking at the posterior mean of some the individual temporal trends we found they did not resemble the regions true

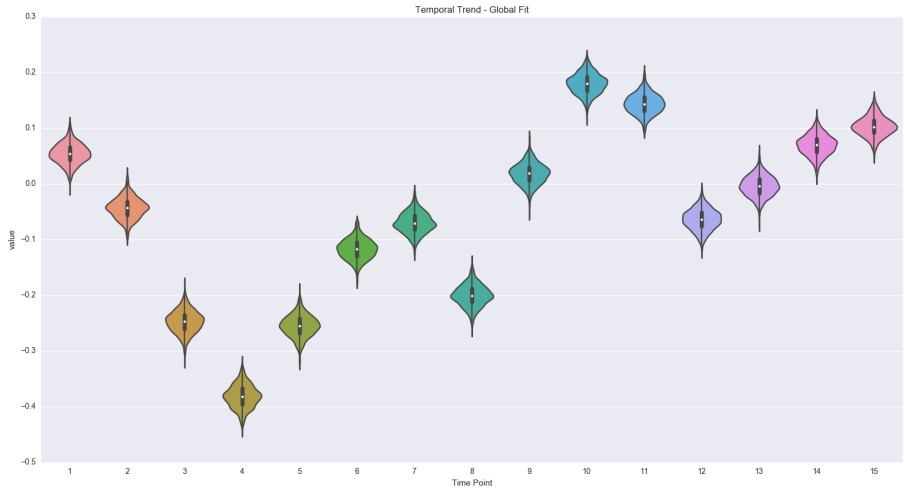


Figure 6.2: Violin plot of the samples from γ_t in model 1 of BayeSTDetect.

trend – even for the 'unusual' regions. The following section discusses why this is the case.

6.5 The cut function

Chapter 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. I this model, the pathology that defines an atypical areas is not an individual temporal trend, but possesing excessive inseperable spatiotemporal variance.

In this paradigm, we use seperate temporal and spatial random effects to capture most ot the heterogeneity between counts. Then any significant space-time interactions that are not common to the shared random effects would indicate a departure from what we would consider typical behaviour.

This model primarily based is on the one proposed in Abellan et al. (2008), with two modifications. First, our latent variable works at the regional level – as opposed to classifying regions based on a decision rule for a mixture component for each count (i.e. we use z_i rather than their $z_{i,t}$). And secondly, we take the response as Poisson distributed rather than Binomial as we do not know the population sizes explicitly.

7.1 Model specification

Like in the previous model 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

$$\log(\mu_{i,t}) = \lambda_i + \gamma_t + \psi_{i,t}. \tag{7.1}$$

where, as before, the spatial effects are given a BYM prior

$$v_{1:N} \sim \text{CAR}(W, \sigma_v^2)$$
$$\lambda_{1:N} \sim \mathcal{N}(v, I_N \cdot \sigma_\lambda^2)$$

and the

$$\gamma_{1:T} \sim \text{RW}_1(\sigma_\gamma^2).$$

With the new component ψ we wish to capture any residual space-time inseperable heterogeneity in the counts. However, we would expect that the is *some* level of interaction at each region, so we define ψ as the mixture of two normal distributions, with the mixture component at the regional level

$$\psi_{i,t} \sim z_i \cdot \text{Normal}(0, \tau_1^2) + (1 - z_i) \cdot \text{Normal}(0, \tau_2^2) \tag{7.2}$$

where the first normal mixture captures only minor interactions, to which we do not assign any epidemiological significance, and the second mixture captures that we would consider atypical.

The mixture components are each assigned Bernoulli priors

$$z_i \sim \text{Bernoulli}(q_i) \tag{7.3}$$

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:

$$\tau_1 \sim \text{Normal}(0, 0.01) \cdot I(0, \infty) \tag{7.4}$$

$$k \sim \text{Normal}(0, 100) \cdot I(0, \infty) \tag{7.5}$$

$$\tau_2 = \tau_2 + k \tag{7.6}$$

where I is the indicator function.

The structure of this model, based on the conditional dependencies between variables can be seen in the Bayesian network in fig fweufno.
Unlike the BaySTDetect model, when we classify the regions in the two sets, we can not use the numerical values of the posterior means of the q_i to define a decision rule by region. This is because the q_i are given uniform priors on the unit interval which does not adequetly convey our prior beliefs that the typical regions should far outnumer the atypical ones. Such a prior could be constructed but a simple Beta prior or similar will not accurately represent this belief.

Instead, we select a preselected number of regions ordinally based on these posterior means of the q_i.

7.2 Classification and accuracy

This model was fit entirely within *Stan*, as before, and the fifteen regions with the greatest values of q_i were selected as the atypical regions. This procedure correctly identified

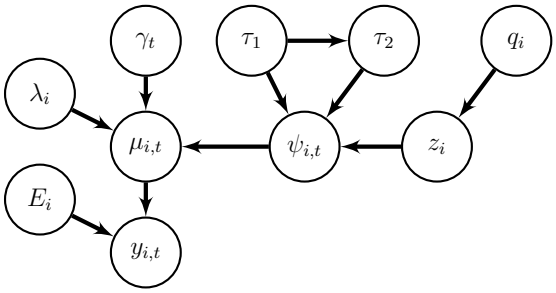


Figure 7.1: Directed acyclic graph showing conditional dependencies between model variables for the excess variability model. Dashed lines indicte one way flow of information. Based on figure in Li et al. (2012).

11 of the fifteen 'unusual' regions giving a false negative rate of 26.7% and a false positive rate of 2.05%. This improves on the false positive rate of the CUSUM model but demonstrates much lower power than either CUSUM or BaySTDetect. The performace the method can be seen graphically in fig 1984.

Despite the reasonable performance of the Abellan et al. (2008) model, it has a number of stark disadvantages compared to BaySTDetect. Firstly, as noted in Li et al. (2012)

> "However, by construction, this model may not be particularly sensitive when the departures exhibit particular time patterns, for example, higher risks occurring at some consecutive time points.".

And secondly, the BYM prior has a tendancy to oversmooth the pointwise data – leading some to develop semiparametric models to counteract this (Best et al. 2005).

7.3 Over-smoothing

As noted in the previous section, the BYM prior has a tendancy to oversmooth the counts at each regions. This can be of particular hinderance to models such as **spatial**, and the modified version above, as we specifically looking for regions which can not be explained by the spatial smoothing term. This leaves regions that would otherwise have been correctly identified as atypical, deemed typical due to the smooth progression of the spatial heterogeneity locally.

Of course, from an epidemiological perspective, the transmission of disease and the dispersed nature of some enviromental hazards would lead us to *expect* pathalogical regions to cluster. Ideally, we need our model that is able to smooth the counts while retaining the ability to identify clustered unusual areas.

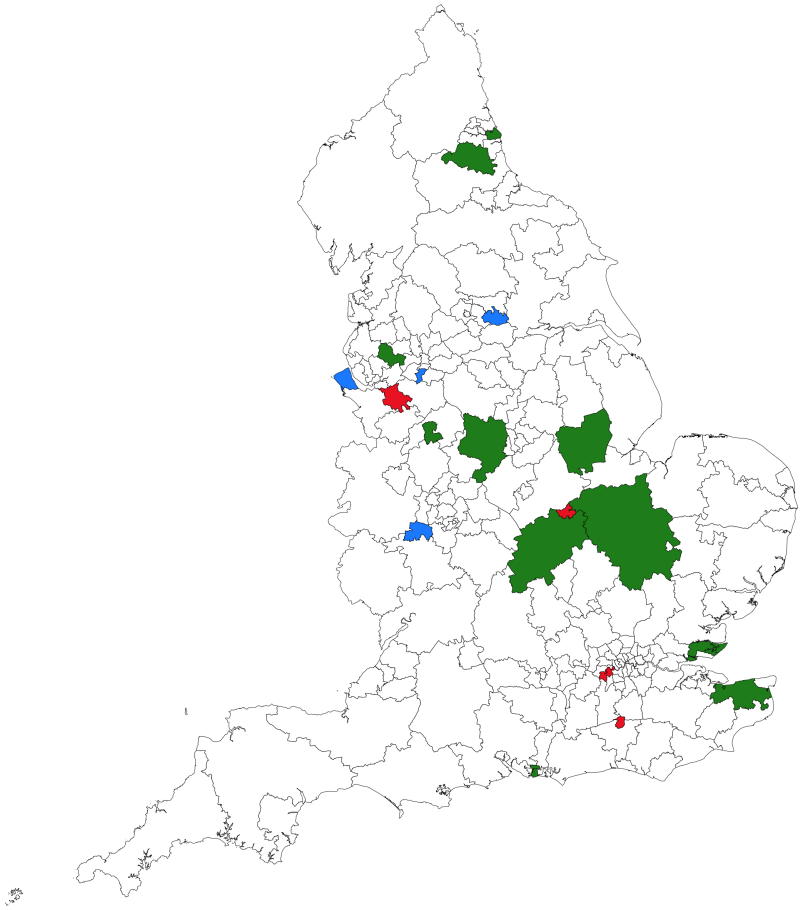


Figure 7.2: Map of unusual regions for the excess variance model. Correctly identified regions are **green**, false-negatives are **blue** and false-positives are **red**

7.4 Simulated Data

We test the severity of this issue of generating a new set of simulated data where the unusual areas are clustered together.

In order to generate the list of unusual areas we suggest a simple algorithm inspired by preferential attatchment processes in network theory. The method is as follows

1. Construct a list of the regions and a corresponding list of 'weights'.
2. Assign all regions i an initial weight of 1 and set $k = 1$.
3. Select a new region u_k from the list of potential regions with probability corresponding to the categorical distribution defined by
$$P(u_k = j) = \frac{w_j}{\sum_i w_i} \tag{7.7}$$
4. Mark the selected region as unusual and remove it (and its weight) from the list of potential regions.
5. Add w^* to each region i that was adjacent to the region removed
$$w_i = w_i + w^* \quad \forall i \in \partial u_k \tag{7.8}$$
where we call w^* the preference weight.
6. Increment k and return to step 3 until the desired number of unusual regions have been selected.

This procedure was applied to the shape data of the 210 CCGs with a preference weight of $w^* = 20$ to select 15 regions in total. The resulting unusual regions selected can be seen in fig fkjhk. We see that the selected regions form a series of clusters of varying sizes.

Using this list of unusual regions we proceeded to generate the data as follows. A random walk was simulated

$$\gamma_{1:15} \sim \mathrm{RW}_1(0.1^2)$$

and then normalised to have a mean of zero. The deviant temporal trend was created generated by adding seperate half-normal variables to the final six time points of the general trend. The resulting two trends can be seen in fig 543.

The spatial component was simulated from a BYM prior and was combined with the temporal trend, identically to as in Chapter 2, to create the rate variables for each region and time point. Multiplying by the same expected values as previously, we drew from a Poisson distribution to arrive at the final values. The code for this can be found in the appendix.

Figure 7.3: Simulated temporal trend for data with clustered unusual regions. The general temporal trend is blue and the red line shows the departure that the deviant trend exhibits.

(a) Truly unsual regions

(b) Model classification

Figure 7.4: Map of unusual regions for the excess variance model where there are clustered unusual regions. Correctly identified regions are **green**, false-negatives are **blue** and false-positives are **red**

Now applying the same procedure as described earlier in the chapter, we only 3 of the total 15 regions were correctly identified. For a null hypothesis where we select the atypical regions randomly, this number of identified regions would give a p-value of around 8%[1] – not very good! Examining the regions identified in fig 321, we that the regions we did correctly identify are non-contiguous. Futher confirming our hypothesis that this method will fail to identify unusual areas when they cluster together.

1. ¿ 1 - phyper(2, 15, 195, 15) = 0.07968

Chapter 8

Computational considerations

In this chapter we briefly note some of the computational considerations that one should be cogent of for some the methods used in this project.

8.1 Computational software

8.1.1 Auto differentiation

Symplectic Integration

- Leap-frog integration
- Metropolis correction

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)

8.2 Autodiff/black-box variational inference?

8.3 Reparameterization of the models

Removing conditional dependencies e.g.

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

8.4 Marginilizing over the discrte parameters

As *Stan* uses 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 confers some computational benefits in terms of the speed at which the chain converges due to the Rao-Blackwellization of the estimator (Casella et al. 1996).

Take for example a mixture of two Normal distributions

$$k \cdot \text{Normal}(\mu_1, \sigma_1^2) + (1-k) \cdot \text{Normal}(\mu_2, \sigma_2^2) \tag{8.1}$$

$$k \sim \text{Bernoulli}(\lambda)\,. \tag{8.2}$$

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

$$\begin{aligned}\text{log\_posterior} = \text{log\_posterior} + \log(&\lambda \cdot \text{Normal\_pdf}(x|\mu_1, \sigma_1^2) \\ &+ (1-\lambda) \cdot \text{Normal\_pdf}(x|\mu_2, \sigma_2^2))\end{aligned} \tag{8.3}$$

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

$$\begin{aligned}\log(p_X(x|\lambda, \mu_i, \sigma_i^2)) &= \log(\lambda \cdot \text{Normal\_pdf}(x|\mu_1, \sigma_1^2) \\ &\quad + (1-\lambda) \cdot \text{Normal\_pdf}(x|\mu_2, \sigma_2^2)) &(8.4)\\ &= \log(\exp(\log(\lambda \cdot \text{Normal\_pdf}(x|\mu_1, \sigma_1^2))) \\ &\quad + \exp(\log((1-\lambda) \cdot \text{Normal\_pdf}(x|\mu_2, \sigma_2^2)))) &(8.5)\\ &= \text{log\_sum\_exp}(\log(\lambda) + \log \text{Normal\_pdf}(x|\mu_1, \sigma_1^2), \\ &\quad \log(1-\lambda) + \log \text{Normal\_pdf}(x|\mu_2, \sigma_2^2)) &(8.6)\end{aligned}$$

For more on this see page 185 of Stan Development Team 2016.

Chapter 9

Comparisions and conclusions

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 Mixture model issues reduced due to setup of variance
Timing things

Bibliography

Abellan, Juan Jose, Sylvia Richardson, and Nicky Best. 2008. "Use of space-time models to investigate the stability of patterns of disease." *Environmental Health Perspectives* 116 (8): 1111.

Banerjee, Sudipto, Bradley P Carlin, and Alan E Gelfand. 2014. *Hierarchical modeling and analysis for spatial data.* Crc Press.

Banerjee, Sudipto, Melanie M Wall, and Bradley P Carlin. 2003. "Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota." *Biostatistics* 4 (1): 123–142.

Besag, J, J York, and A Mollie. n.d. "Bayesian image restoration with two applications in spatial statistics (with discussion) Ann Inst Stat Math. 1991; 43: 1–59. doi: 10.1007." *BF00116466.[Cross Ref].*

Best, Nicky, Sylvia Richardson, and Andrew Thomson. 2005. "A comparison of Bayesian spatial models for disease mapping." *Statistical methods in medical research* 14 (1): 35–59.

Bivand, Roger, Jan Hauke, and Tomasz Kossowski. 2013. "Computing the Jacobian in Gaussian spatial autoregressive models: An illustrated comparison of available methods." *Geographical Analysis* 45 (2): 150–179. `http://www.jstatsoft.org/v63/i18/`.

Bivand, Roger, Tim Keitt, and Barry Rowlingson. 2015. *rgdal: Bindings for the Geospatial Data Abstraction Library.* R package version 0.9-2. `http://CRAN.R-project.org/package=rgdal`.

Casella, George, and Christian P Robert. 1996. "Rao-Blackwellisation of sampling schemes." *Biometrika* 83 (1): 81–94.

Clayton, David, and John Kaldor. 1987. "Empirical Bayes estimates of age-standardized relative risks for use in disease mapping." *Biometrics:* 671–681.

Crainiceanu, Ciprian M, David Ruppert, and Matthew P Wand. 2007. "Bayesian analysis for penalized spline regression using Win BUGS."

Dassanayaka, Sesha K. 2014. "Spatial CUSUM Chart Based Method for Rapid Detection of Outbreaks." *Online journal of public health informatics* 7 (1).

Gandy, Axel, and Jan Terje Kvaloy. 2013. "Guaranteed conditional performance of control charts via bootstrap methods." *Scandinavian Journal of Statistics* 40 (4): 647–668.

Gelman, Andrew, et al. 2006. "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." *Bayesian analysis* 1 (3): 515–534.

Gelman, Andrew. 2013. "workaround for cut() function? – comments." `https://groups.google.com/forum/#!msg/stan-users/ruxPAhBPI4A/SrsbM-rese0J`.

Goicoa, T, MD Ugarte, J Etxeberria, and AF Militino. 2012. "Comparing CAR and P-spline models in spatial disease mapping." *Environmental and Ecological Statistics:* 1–27.

Jiang, Wei, Lian Jie Shu, and Kwok Leung Tsui. 2011. "Weighed CUSUM control charts for monitoring inhomogeneous poisson processes with varying sample sizes." *Journal of quality technology* 43 (4): 346–362.

Joseph, Max. 2016. *Exact sparse CAR models in Stan.* `http://mc-stan.org/documentation/case-studies/mbjoseph-CARStan.html`.

Kulldorff, Martin. 1997. "A spatial scan statistic." *Communications in Statistics-Theory and methods* 26 (6): 1481–1496.

Li, Guangquan, Nicky Best, Anna L Hansell, Ismail Ahmed, and Sylvia Richardson. 2012. "BaySTDetect: detecting unusual temporal patterns in small area data via Bayesian model choice." *Biostatistics:* kxs005.

Lunn, David, Chris Jackson, Nicky Best, Andrew Thomas, and David Spiegelhalter. 2012. *The BUGS book: A practical introduction to Bayesian analysis.* CRC press.

Mandt, Stephan, Matthew D Hoffman, and David M Blei. 2017. "Stochastic Gradient Descent as Approximate Bayesian Inference." *arXiv preprint arXiv:1704.04289.*

Mei, Yajun, Sung Won Han, and Kwok-Leung Tsui. 2011. "Early detection of a change in Poisson rate after accounting for population size effects." *Statistica Sinica:* 597–624.

Ng, Andrew Y, and Michael I Jordan. 2002. "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes." *Advances in neural information processing systems* 2:841–848.

Page, ES. 1954. "Continuous inspection schemes." *Biometrika* 41 (1/2): 100–115.

Plummer, Martyn. 2015. "Cuts in Bayesian graphical models." *Statistics and Computing* 25 (1): 37–43.

Raubertas, Richard F. 1989. "An analysis of disease surveillance data that uses the geographic locations of the reporting units." *Statistics in Medicine* 8 (3): 267–271.

Ritov, Y. 1990. "Decision Theoretic Optimality of the Cusum Procedure." *Ann. Statist.* 18, no. 3 (September): 1464–1469. doi:10.1214/aos/1176347761. http://dx.doi.org/10.1214/aos/1176347761.

Stan Development Team. 2016. *Stan Modeling Language Users Guide and Reference Manual.* Version 2.14.0. http://mc-stan.org/.

Vapnik, Vladimir Naumovich, and Vlamimir Vapnik. 1998. *Statistical learning theory.* Vol. 1. Wiley New York.