Detecting pathologies in spatiotemporal data

Michael Redman, CID: 00826863

December 2016

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 [Abellen et al], 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 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, like an outbreak.

Consequently, in this project we will be looking at data in both space and time. A variety of methods have been investigated in this context. 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 thier ability to detect abnormalities at the individual level, i.e. isolated to a particular region (cite Baystdetect).

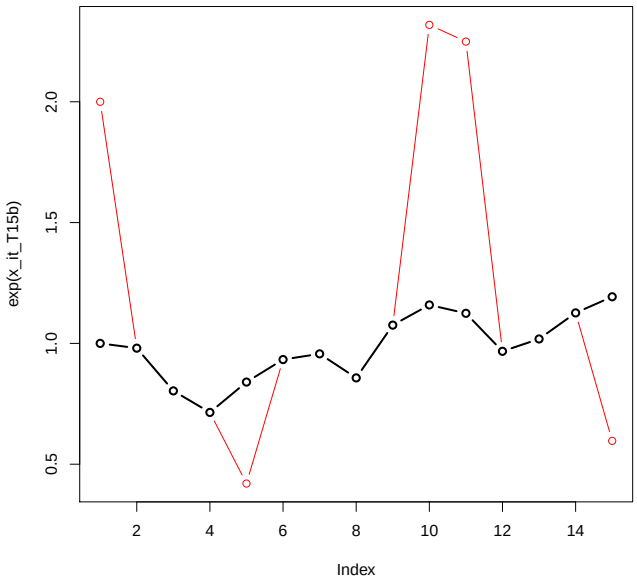
- Preventative medicine
- small/fine-grained areas
- Allows for greater spcificity while presenting new challenges
- Data sparcity
- Computational advances
- Enviromental exposures
- Great care/subtleties of the computational methods
- Spatial scan statistics
- Type of data we will be looking at

What I do:

1. Investigate the use of control chart based methods
2. Explore the use of Bayesian models
3. Look at the use of the cut function
4. Performance in terms of false-positive/negatives
5. Speed performance
6. Try a new model for identifying unseperable variance
7. Create new simulated data with preferential attachment to test issues with it
8. Hamiltonian monte carlo and issues with mixture models

2 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:

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

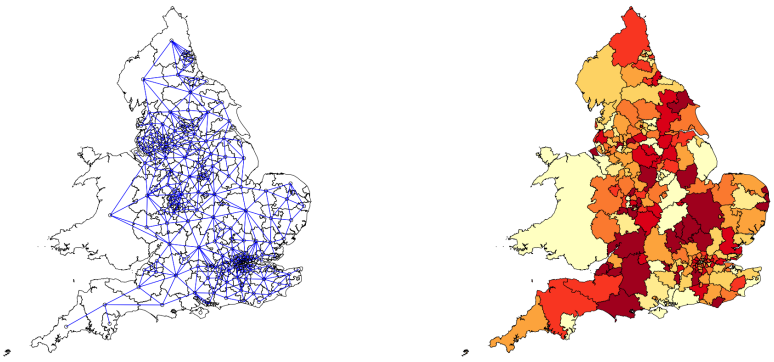
$$g^*(t) = g(t) - \log(2) \qquad t = 5, 15 \tag{2}$$

- 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.

2.1 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.



(a) Adjacency (b) Expected

Figure 1: Mapped data

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

3 CUSUM

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. 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.

- Mention that the process is sequential and so can be calculated as we go.

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 of a relavant test statistic. We define S_t as the value of the control chart at time t and progress its value as follows

$$S_0 = 0 \tag{3}$$

$$s_{t+1} = \max(0, S_t + K_t). \tag{4}$$

Where K_t is the value of some statistic calculated at time point t. We will use the log of the ratio between the likelihoods of the data, at the time point, under an "in-control" and an "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. The logarithms are taken to minimise floating-point arithmetic issues as the densities will often be very small. We then signal that the process is pathological if the value of S_t exceeds some threshold value h.

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 area 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 neglat 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

$$\tilde{Y}_{i,t} = \frac{Y_{i,t}}{E_i}, \qquad i = 1, \ldots, R,\ t = 1, \ldots, T. \tag{5}$$

We then construct the "in-control" rate for each region $I_{i,t}$ as described

$$I_{i,t} = E_i \cdot \frac{1}{R} \sum_{i=1}^{N} \tilde{Y}_{i,t}, \qquad i = 1, \ldots, R,\ t = 1, \ldots, T. \tag{6}$$

For simplicity we only consider the case of abnormally high count rates and ignore artificially low counts but 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

$$O_{i,t} = \lambda \cdot I_{i,t}, \qquad i = 1, \ldots, R,\ t = 1, \ldots, T. \tag{7}$$

with $\lambda > 1$.

Now for a given value of λ 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 n times and find the smallest value of h for which the the false-postive rate is below 1%.

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 λ 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 5.2%.

To conclude, while CUSUM correctly identified the unusual regions, we have a high false-positive rate as the "in-control" likelihood was a poor approximation to the true likelihood for a number of regions.

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

[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.

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

4.1 Hyperparameter priors

- Gelman 2006 for variance parameters
- Stan wiki for others

4.2 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"

4.3 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.

4.3.1 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

$$p(\theta) = \frac{1}{Z} e^{-E(\theta)}. \tag{8}$$

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:

$$H(p,q) = \underbrace{T(p|q)}_{\text{Kinetic energy}} + \underbrace{V(q)}_{\text{Potential energy}} \tag{9}$$

looking at the canonical ensemble of the join distribution

$$\begin{aligned} \pi(p,q) &\propto e^{-H(p,q)} && (10)\\ &= e^{-[T(p|q)+V(q)]} && (11)\\ &= e^{-T(p|q)} \cdot e^{-V(q)} && (12)\\ &\propto \pi(p|q) \cdot \pi(q) && (13) \end{aligned}$$

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

$$V(q) = -\log \pi(q) \tag{14}$$

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

$$\frac{dq}{dt} = \frac{\partial H}{\partial p} = \frac{\partial T}{\partial p} \tag{15}$$

$$\frac{dp}{dt} = -\frac{\partial H}{\partial q} = -\frac{\partial T}{\partial q} - \frac{\partial V}{\partial q} \tag{16}$$

- Finish off this bit more gracefully

4.3.2 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

4.4 General Model Form

- Use best et al to summarise basic model form

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 (source here) but in this project we will primarily look at using a conditionally autoregressive prior.

5.1 CAR models

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

$$v_i \mid v_j \; j \neq i \sim N(\alpha \cdot \bar{\mu_i}, \; \sigma_v^2/k_i) \tag{17}$$

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

$$\bar{\mu}_i = \sum_{j \in \partial i} \frac{\mu_j}{k_i} \tag{18}$$

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,

$$v \; \sim \; N(0, \; \sigma_v^2 \cdot [D(I_n - \alpha B)]^{-1}) \tag{19}$$

where

$$D = \text{diag}(k_i) \tag{20}$$

$$B = D^{-1}W \tag{21}$$

$$(W)_{ij} = \begin{cases} 1, & i \leftrightarrow j \\ 0, & \text{otherwise} \end{cases} \tag{22}$$

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 [Joseph, 2016]

- 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

5.2 BYM prior

- CIte Besag York Mollie paper
- Increases degrees of freedom for spatial componenet

5.3 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

$$\xi_{1:T} \sim \text{RW}(1) \tag{23}$$

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.

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

$$Y_{i,t} \sim \text{Poisson}(E_{i,t} \cdot \mu_{i,t}) \tag{24}$$

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

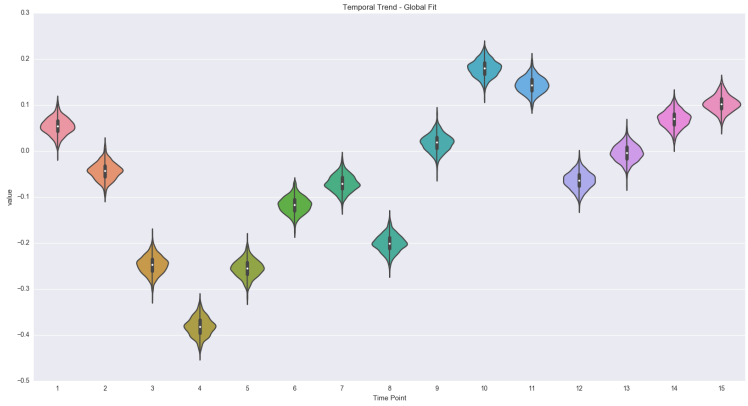
$$\log(\mu_{i,t}) = \begin{cases} \lambda_i + \gamma_t \ (+ \ \alpha_0) & \text{Model 1 for all } i,t \\ u_i + \xi_{i,t} & \text{Model 2 for all } i,t \end{cases} \tag{25}$$

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

$$\alpha_0 \sim \text{Flat}(\mathbb{R}) \tag{26}$$
$$v_{1:N} \sim \text{CAR}(W, \sigma_v) \tag{27}$$
$$\lambda_{1:N} \sim \text{Normal}(v, \sigma_\lambda) \tag{28}$$
$$\gamma_{1:T} \sim \text{RW}(1) \tag{29}$$

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



$$i \in 1, \ldots, N \begin{cases} u_i \sim \text{Normal}(0, 1000) \\ \xi_{i,1:T} \sim \text{RW}(1) \end{cases} \tag{30}$$

6.2 Combining the models

- BUGS can be used with a mixture component
- In this case the likelihood can be worked out by hand with the probability of the mixture component prior
- Describe likelihood code

6.2.1 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.

6.2.2 Classification and accuracy

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

6.3 Fully Bayesian model

- Can be fit fully in Stan
- Reference the use of Rao-Blackwellization
- Graph the temporal components failure

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
- Mixture model issues reduced due to setup of variance

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

$$\log(\mu_{i,t}) = \lambda_i + \gamma_t + \psi_{i,t} \tag{31}$$

where, as before,

$$v_{1:N} \sim \text{CAR}(W, \sigma_v^2) \tag{32}$$
$$\lambda_{1:N} \sim \text{Normal}(v, \sigma_\lambda^2) \tag{33}$$
$$\gamma_{1:T} \sim \text{RW}(1). \tag{34}$$

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 distributions, with the mixture component at the regional level,

$$\psi_{i,t} \sim z_i \cdot \text{Normal}(0, \tau_1^2) + (1 - z_i) \cdot \text{Normal}(0, \tau_2^2) \tag{35}$$

where

$$z_i \sim \text{Bernoulli}(q_i) \tag{36}$$

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:

$$\tau_1 \sim \text{Normal}(0, 0.01) \cdot I(0, \infty) \tag{37}$$

$$k \sim \text{Normal}(0, 100) \cdot I(0, \infty) \tag{38}$$

$$\tau_2 = \tau_2 + k \tag{39}$$

Here I is the indicator function.

7.2 Classification and accuracy

This model was fit using HMC in *Stan*

7.3 Over-smoothing

- Expected disease trends will cluster
- Model relies on finding unseperable variance but local high variance could be explained by CAR component
- Need a model which smooths regions but is able to still identify clustered unusual areas

7.4 Simulated Data

- Preferential attachment scheme
- Describe full data generation parameters and process
- Include plots of temporal and spatial components

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

¡¡¡¡¡¡¡ HEAD =======

8.1.3 Auto differentiation

8.1.4 Symplectic Integration

¿¿¿¿¿¿¿ origin/master

- Leap-frog integration
- Metropolis correction

¡¡¡¡¡¡¡ HEAD

8.1.5 Auto differentiation

8.1.6 Symplectic Integration

- Leap-frog integration
- Metropolis correction

8.1.7 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)

=======

8.1.8 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)

¿¿¿¿¿¿¿ origin/master

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 has significan computational benefits (expand on this).

Take for example a mixture of two Normal distributions

$$k \cdot \text{Normal}(\mu_1, \sigma_1^2) + (1-k) \cdot \text{Normal}(\mu_2, \sigma_2^2) \tag{40}$$

$$k \sim \text{Bernoulli}(\lambda)\,. \tag{41}$$

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

$$\begin{aligned}\text{log\_posterior} = \text{log\_posterior} &+ \log(\lambda \cdot \text{Normal\_pdf}(x|\mu_1, \sigma_1^2) \\ &+ (1-\lambda) \cdot \text{Normal\_pdf}(x|\mu_2, \sigma_2^2))\end{aligned} \tag{42}$$

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

$$\begin{aligned}\log(p_X(x|\lambda, \mu_i, \sigma_i^2)) &= \log(\lambda \cdot \text{Normal\_pdf}(x|\mu_1, \sigma_1^2) \\ &\quad + (1-\lambda) \cdot \text{Normal\_pdf}(x|\mu_2, \sigma_2^2)) &(43)\\ &= \log(\exp(\log(\lambda \cdot \text{Normal\_pdf}(x|\mu_1, \sigma_1^2))) \\ &\quad + \exp(\log((1-\lambda) \cdot \text{Normal\_pdf}(x|\mu_2, \sigma_2^2)))) &(44)\\ &= \text{log\_sum\_exp}(\log(\lambda) + \log \text{Normal\_pdf}(x|\mu_1, \sigma_1^2), \\ &\quad \log(1-\lambda) + \log \text{Normal\_pdf}(x|\mu_2, \sigma_2^2)) &(45)\end{aligned}$$

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

- Cite Stan manual
- Rao-Blackwellization?

8.5 Timing data?

¡¡¡¡¡¡¡ HEAD

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.

[Joseph, 2016] Joseph, M. (2016). Exact sparse car models in stan.

[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.

[Stan Development Team, 2016] Stan Development Team (2016). Stan modeling language users guide and reference manual. Version 2.14.0.

=======

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.

[Joseph, 2016] Joseph, M. (2016). Exact sparse car models in stan.

[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.

[Stan Development Team, 2016] Stan Development Team (2016). Stan modeling language users guide and reference manual. Version 2.14.0.

¿¿¿¿¿¿¿ origin/master