Examiner's Report

'Anomaly Detection in Patient Arrival, with Bayesian Hierarchical Models'

An MSc thesis in Statistics
submitted by Junhuang Xue
to the University of Auckland.

Thesis summary

This thesis addresses the problem of estimation of parameters at various levels of hierarchy of counts – such as counts of disease classified by a hierarchical coding system such as ICD9.

The goal of the thesis is to investigate how estimation can be improved by using the information in the hierarchy, rather than treating the counts at the lowest level as independent.

General Comments

The thesis is a good length, and contains few spelling errors, but multiple grammatical errors which were at times distracting. The style is a little too conversational (e.g. 'a little bird told me...', end of p19; 'it is kind or hard to tell ...', 1st para of discussion p55), but the writing is otherwise mostly fine.

However there are rather too many places where the description of a model is not matched by a suitable symbolic description which would clarify the ambiguities of the language. The formulae which are given often contain indices whose ranges are not stated, all leading to uncertainty in the reader about what is intended.

There are also some specific places where I think there have been some significant misapprehensions on the student's part – I list these in the chapter specific comments below.

There are key moments where it is very hard to work out what has actually been done, and my overall impression is of a student who doesn't fully understand the work he has carried out. The interpretations are mostly either absent, or seem to miss the point.

It's clear the student has strong programming ability, and has implemented simulations and model fitting for an array of non-trivial cases.

He has taken a real data set and transformed and modified it to make it suitable for analysis.

There are certain cases where he has correctly interpreted the findings of the simulation or model fitting exercise.

Specific Comments and corrections

Chapter 1 - Background

This chapter introduces the essential aspects of the problem: namely modelling a repated set of observations of counts in a hierarchy. The motivating setting is that of a hospital emergency department, which is subject to strong fluctuations in demand, including rare instances of overwhelming demand.

- Some space is devoted to anomaly detection in §1.4 - but this seems not to be the goal of the thesis.
- p7, eq (1.1) - the argument of $\Phi^{-1}(\cdot)$ is not restricted to within $[0,1]$ – I didn't find the source reference but did find one by the same authors Wong et al. from 2002, where the equation is

  $$\text{Threshold} = \mu + \sigma * \Phi^{-1}(1 - \tfrac{1}{2}\text{p-value})$$

  which I suspect should replace the incorrect (1.1).
- p7, eq (1.2) and surrounding text: the definition of threshold here seems somewhat circular: it looks like it is simply the capacity C of the system and the typical load μ is just the proportion k of the capacity that is typically occupied, i.e.

  $$k = \frac{\mu}{C}$$

  seems a better definition. Then the threshold is just the capacity, and an alarm should go off when the load exceeds capacity: a simple and logical criterion.
  In Chapter 2 this quantity k is referred to as a capacity, whereas it would be better referred to as the typical **load fraction**, and μ the typical **load**.
- §1.5 is an introduction to the basics of Bayesian statistics. It is somewhat incoherent, and doesn't define its notation: e.g. around eq (1.3) on p9 it should be stated that A here is typically a parameter of interest and B is a set of observations that are informative about A.
- There are a few typos in the section - 4th last line of p9, should be $\{\theta_1, \dots, \theta_p\}$, though p is not defined anywhere
- p10, the Poisson example in (1.4) could have been usefully carried onwards
- p10, eq (1.6), the first proportionality should be an equality, the variable of integration in the denominator should be θ not x (this is repeated in the text after the equation), and in the third line of (1.6) the prior is $f_\Theta(\theta)$ not $f_\Theta(x)$
- p10 following eq (1.6), $f_X(x)$ is the distribution of X, not its expectation, and c is a crucially only a constant *with respect to* θ, since it otherwise clearly depends on x
- §1.5.3, p11 – state what Gibbs sampling is, and what happens in an MCMC sampler when multiple parameters are being sampled. The concept of full conditional distribution should be introduced here.
- §1.5.4, p12 – line 6 after Fig 1.3: 0.51→0.051 and 0.49→0.049.

- §1.5.5, p13, first para – The CLT is not a motivating factor in the choice of a Normal distribution as a prior.
- §1.5.6, first para, last lines: "no independence": it would be better here to introduce the concept of conditional independence, and write – with the aid of suitable formulae – the dependence relationships that exist in a one-level model (called the IBM here) and a two-level model (called the HBM here). The expressions in (1.9) and (1.10) do not fully capture the essence here: α should not appear in (1.9), and the two models that should be contrasted are, for level 1 categories i and level 2 categories $j \in i$

  $$\begin{aligned} x_{ij}|\theta_{ij} &\overset{\text{ind}}{\sim} f(\cdot|\theta_{ij}) \\ \theta_{ij}|\alpha &\overset{\text{iid}}{\sim} f(\alpha) \\ \alpha &\sim f(\cdot) \end{aligned}$$

  where $i = 1,\dots,n$, and $j = 1,\dots,n_i$, and

  $$\begin{aligned} x_{ij}|\theta_{ij} &\overset{\text{ind}}{\sim} f(\cdot|\theta_{ij}) \\ \theta_{ij}|\phi_i &\overset{\text{iid}}{\sim} f(\theta_{ij}|\phi_i) \\ \phi_i|\alpha &\overset{\text{iid}}{\sim} f(\phi_i|\alpha) \\ \alpha &\sim f(\cdot) \end{aligned}$$

  The former model is the 'independence' model (no pooling), and the latter is the 'hierarchical' model (pooling within level 2).
- §1.5.7, first para – tidy notation - pD, Pd, or better p_D? The definition of the DIC is incomplete here, and doesn't define $\bar{D}$ and $\bar{\theta}$, and doesn't note how the DIC is used in the context of an MCMC posterior sample.
- p17, (1.16), index of sum for $y_{i,t}$ should be j

Chapter 2 - Simulations

This chapter contains three sets of simulations and model fitting.

- §2.2, p24 - a function `simdata()` has been written to generate two level hierarchical data (one two way split into A or B, and a further split of each into AA, AB on the A branch, and BA, BB on the B branch. The other details of the simulation are not given, aside from the construction of a fixed matrix ρ which specifies the probabilities of each of the four leaves. It isn't clear why these values have to add to 1000 (instead of being four non-negative real numbers which add to 1 – i.e. probabilities), particularly because they are scaled later by a set of multipliers λ which encode the actual sample sizes. (In the one instance of the matrix given, in Figure 2.2, the values add to 999 not 1000.)

  The details of the `simdata()` function aren't given (the Appendix only specifies the arguments to the function), and so it isn't clear exactly what it is doing. It appears just to implement the IBM

model. As far as I can see this is the only function being used to generate the data in the simulations. If so, then the comparisons of the IBM and HBM models doesn't make sense. The true model is apparently IBM, yet the student in many places in the chapter claims that the HBM model is better than, or at least as good as, the IBM model.
The logic of the simulation studies is therefore unclear.

- When the simulations are described, they are not accompanied by statements of (1) sample size, (2) values of ρ, or (3) number of replicates. I'm guessing that there are replicates, because multiple values of DIC have to be calculated in order for tables like Table 2.1 to be constructed. I just can't work out how this was done, and it wasn't explained.
- p25 last lines – 'binding' – this is an R language term, rather than describing the appropriate statistical/data structure operation. It would be best to include equivalent mathematical symbols that describe what is being done here.
- §2.3.1 – six models are defined which are fitted to the synthesised data. This was difficult to follow for several reasons.
  - The symbol $\mu_{i,t}$ is used to mean two different things, e.g. in Model 2 where it is the mean of the distribution of $y_{i,t}$ and later the mean of the distribution of $\lambda_{i,t}$. The distinction is kept correctly in the code in the Appendix, but the mathematical formulae are needlessly confusing.
  - The mean and variance of the prior for $\lambda_{i,t}$ in Model 2 (and elsewhere) are indexed by i,t, whereas in fact the do not vary over i or t
  - The prior for σ^2 in Model 2 is a Normal distribution which could go negative (even though it is unlikely to do so), and should have a formal restriction to be non-negative
  - The student seems not have realised that JAGS parameterises the variance of a Normal distribution by the **precision**, $\tau = \sigma^{-2}$, so that $N(1,0.1)$ has a variance of 10. This is evidenced in the code on p110 in Appendix A.2 where the prior for λ is specified, and the code suggests that the Normal distribution is parameterised (μ,σ) rather than either the standard (μ,σ^2) or the correct $(\mu,1/\sigma^2)$.
  - Models 2 and 3 are referred to as Normal(a,b) distributions in the text – however a and b are not the mean and variance of a Normal distribution, they are the separate means of two of the Normal priors.
  - The constants ρ_i are always indexed as $\rho_{i,t}$, but as far as I can see they do not vary with t and should be referred to as ρ_i.

  These errors suggest that the student has not fully understood the model he has implemented.
- p29, although they are standard distributions, the Laplace distribution and its parameterisation should be defined
- p31, Model 1 is described here, and repeatedly in the pages that follow, as having a non-informative prior. In fact it has a highly-informative prior: the prior is degenerate with $\lambda_{i,t} = 1$ with zero variance. On p38 the student notes that there is 'no mixing at all' in these parameters for Model 1 – but doesn't state the obvious: that the model fixes this parameter so it cannot vary.
- With it being unclear exactly how the data have been synthesised, it's not obvious how to interpret the fitting of these six models: e.g. I can't work out what we are supposed to learn from Table 2.1.

- The tables and figures that are given in pages 36-40 do not have captions which say which parameters are being plotted/summarised. I guess that 'Total' means the sum $S_t = \sum_i \mu_{i,t}$, or is it $\frac{1}{T}\sum_t S_t$ over T time periods?
- p37 – the discussion of what causes the lowest effective sample sizes (ESS): Model 5 has the lowest, and that would seem to me to be because the prior has the highest prior variance, so the widest posterior support, which may lead to slower mixing.
- The remainder of the chapter (Simulations 2 and 3) are just as difficult to follow as the discussion of Simulation 1. The claim that the HBM (which is I think Model 2) outperforms the IBM (which I think is Model 1) is made on several occasions, despite the fact that the tables of DIC values are equivocal, and despite the fact that the true model is actually the IBM model.
- The exceedance probabilities discussed on p46 should be defined in symbols. They are a nice way of showing sensitivity to the setting of the model.
- p49, the student correctly interprets the findings that high k (high prior load proportion) means high likelihood of exceedence. Again it would be less confusing if k were not referred to as a capacity here.
- p55 – last line 'indecency' isn't the right word here - not sure what was intended
- p56 the discussion of ESS shows a misunderstanding of what ESS signifiies: it is a property of the estimation/sampling process being MCMC Gibbs sampling. It doesn't say anything about the appropriateness of the model, or the 'disregarding of information' about the hierarchy by the IBM.
- p57, a diagram of the structure of the branching (not 'brunching') model would have been good here
- p60 – the definition of the multiple branching model on p57 suggests that the mean occupancy of each leaf should decrease as the branching number increases That isn't the case in Table 2.17 – is something wrong here?

Chapter 3 - MIMIC-III

This chapter uses a real dataset of diagnoses using the ICD9 code and its internal levels to define the hierarchy.

- p74. A visualisation of the data would have been good. How many distinct ICD codes appear in the dataset?
- §3.6, p75 – how and why is time series prediction being used to estimate prior values? I couldn't work it out. Are these the $\rho_{i,t}$ values?
- §3.6, 3rd para, line 5 – I think the IBM here should be HBM
- p76 the same reuse of $\mu_{i,t}$ occurs here as in Chapter 2.
- p76 it's not stated how the λ values at the various levels of the hierarchy aggregate. From the code in the Appendix it looks as if the nodes at all levels of the hierarchy might be being modelled separately? without the imposition of the additivity constraint?
- The selection of three specific ICD9 codes with different degrees of prevalance was a good idea. p77.

- There is a missing Figure on p78 (referred to in line 2)
- Comparing Table 3.5 and 3.6 (pages 82 and 83) – if the anomaly is added to just one year then I would expect the mean for AA to rise slightly - as it does in Table 3.6, rather than doubling as it does in Table 3.5. What is the interpretation of this?
- In contrast we don't see the same difference in Table 3.9, 3.10 – why not?
- p100, the student correctly identifies the reason for varying success in anomaly detection at different levels of the hierarchy as being due to the size of the group. Doubling the size of a small group makes very little difference to a total, and is therefore visible only at the end leaf level.

Chapter 4 - Conclusion

This chapter summarises the findings of earlier chapters in the thesis.

Appendix - Codes

The appendix contains some of the code used in the thesis. It was helpful in finding what the models actually were, when the main text wasn't clear.