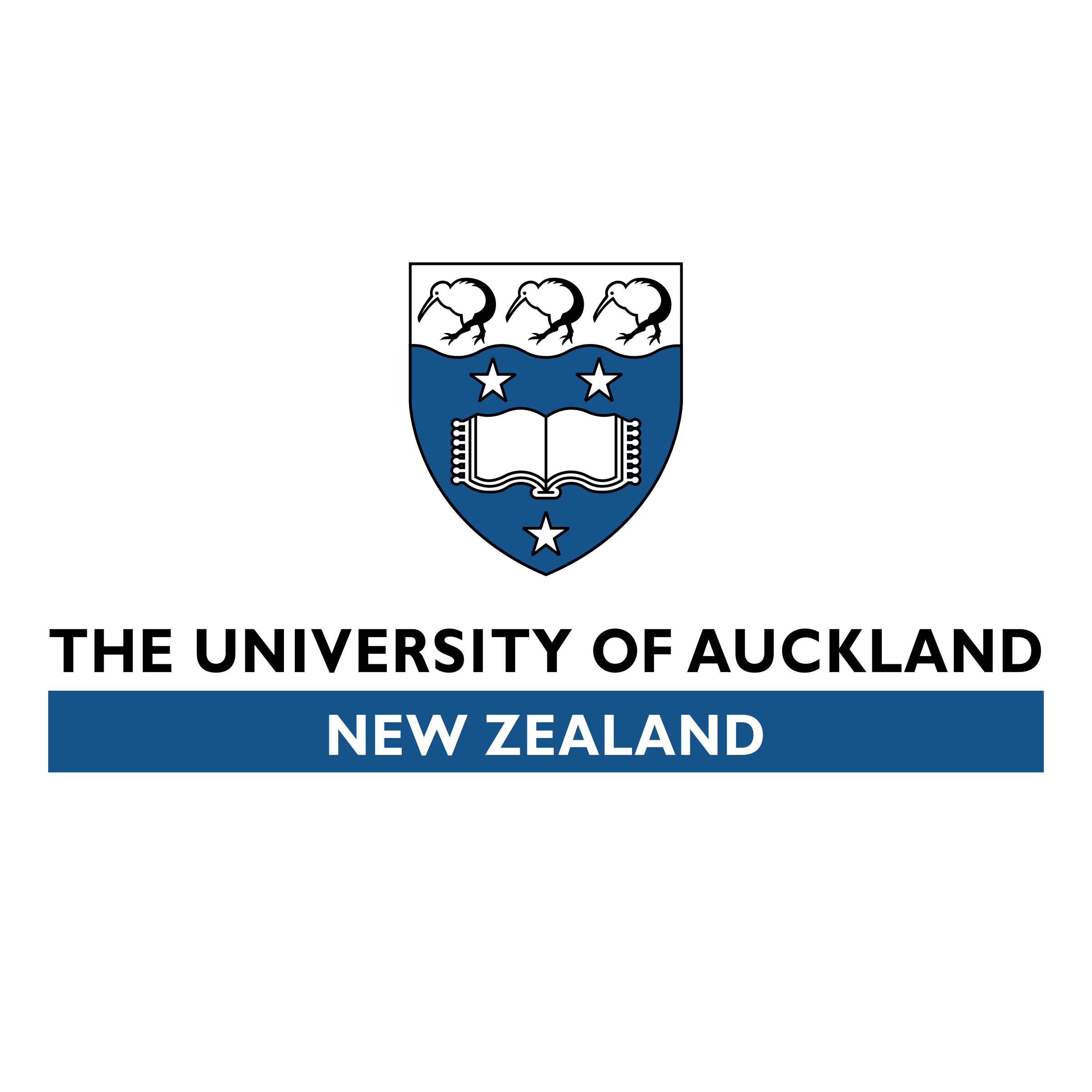
**Anomaly Detection in Patient Arrival, with Bayesian Hierarchical Models**



*A Thesis Submitted in Partial Fulfilment of the Requirement for the Degree of Masters of Science (Research Masters) in Statistics,*

*Department of Statistics, The University of Auckland, 2019*

Supervisor: Professor Thomas Lumley

by Junhuang Xue (薛君煌) July 15, 2019

**Abstract**

For datasets with a hierarchical structure, there may be important information that is contained within its hierarchical relationship, such as the aggregation constraint, that children of a category will sum to itself. With a hierarchical structure, there are diff t levels of hierarchy, modelling with only one level of hierarchy may miss vital information and reduce the power of our models. A hierarchical model that considers all levels could yield more promising predictions.

We compared posterior outputs and anomaly signals using Individual Bayesian Model (IBM) and Hierarchical Bayesian Model (HBB), with several simulation studies and found that HBM seems to out-perform. However, there is only a minimal diff in anomaly signals.

ICD-9 is a standard of disease diagnosis with a hierarchical structure. We also investigate the use of HBM as a possible approach for the detection of anomalies for diff t rarities of disease with ICD-9 categories, and found that for common disease with a large counts it is possible to detect anomalies at higher levels, however when the disease is rare of extremely rare it is recommended to look at lower levels for more reliable anomaly signals.

Hierarchical models could provide better predictions, but the process of building the model is not an easy task, and with complex hierarchical structures, it becomes a tedious task. Independent models may be inferior but are easier to build and could produce adequate results.

Use of HBM provides better results than IBM for the case of anomaly detection for hierarchical datasets, HBM tends to have better goodness of fi and the diff

in detection results are apparent. HBM is a feasible design that is well-suited for datasets that have attributes with large hierarchical structures and have an advantage over IBM in many areas.

**Declaration**

I declare that this thesis is partial fulfilment of the requirement for the University of Auckland degree of Masters of Science (Msc) in Statistics.

Except where specific reference is made to the work of others, the contents of this dissertation are original and follows academic integrity. This thesis is the result of my own work and shall includes nothing which is the outcome of work done in collaboration,unless it was specifically indicated in the text.

Signed: Date: / /

**Acknowledgments**

First I would like to express the up-most gratitude towards my supervisor Professor Thomas Lumley, Chair in Bio-statistics, Department of Statistics, University of Auckland New Zealand. Who I fi meet from STATS 773 Design and Analysis of Clinical Trials back in 2017. Without his supervision and constant guidance over the past year. It would not have been impossible to write my thesis.

I would also like to thank my family, for all the emotional and fi support they unconditionally provided through my never-ending studies, they had done the best there is for me to have the best chance to success in life. I just hope that my old men would stop thinking that I studied fortune telling (Psychology), and recognise that I actually studied a real, academic subject called Statistics.

And, many thanks to all the academic and professors from University of Auckland, around new Zealand and all round the world, who provided help and guidance over the past year. Knowledge does not stay in a box, it is everywhere and it produces friendly encounters.

I would like to extend my warmth to all my friend from psychology, statistics, biological science, education, math and computer science, who struggled through countless overnight assignments, fell asleep together during lecture and panic over exam marks with me.

Lastly, to whoever had the time to spent on my Thesis, thank you for your time, and if you are a postgraduate student who is currently doing a dissertation/Thesis, I wish you good luck.

— Jung.

**Contents**

### [Background](#_bookmark0) 1

* 1. [Introduction](#_bookmark1) 1
  2. [Emergency Department Congestion](#_bookmark3) 2
  3. [Motivation](#_bookmark4) 4
  4. [Anomaly Detection](#_bookmark5) 5
  5. [Hierarchical Bayesian Models](#_bookmark9) 8
     1. [The history of Bayesian statistics](#_bookmark10) 8
     2. [Bayesian inference](#_bookmark11) 9
     3. [Simulation-based estimation](#_bookmark15) 10
     4. [Bayesian against Frequentist](#_bookmark18) 12
     5. [Prior Selection](#_bookmark20) 13
     6. [Hierarchical Bayesian models](#_bookmark21) 14
     7. [DIC](#_bookmark25) 15
     8. [Effective Sample Size](#_bookmark26) 16
  6. [Hierarchical time-series](#_bookmark27) 16
     1. [Structure of Hierarchical time-series](#_bookmark28) 16
     2. [Forecasting approaches](#_bookmark30) 18
  7. [Outline of Thesis](#_bookmark34) 21
  8. [Notations and conventions](#_bookmark35) 22

### [Simulations](#_bookmark36) 23

* 1. [Introduction](#_bookmark37) 23
  2. [Simulation Methods](#_bookmark38) 23
  3. [Simulation 1: Prior](#_bookmark43) 27
     1. [Priors](#_bookmark44) 27
     2. [Simulation setups](#_bookmark46) 32
     3. [Results](#_bookmark47) 32
     4. [Discussion](#_bookmark61) 41
  4. [Simulation 2: Anomaly](#_bookmark62) 42
     1. [Simulation setups](#_bookmark63) 42
     2. [Results](#_bookmark64) 42
     3. [Discussion](#_bookmark78) 55
  5. [Simulation 3: Structure](#_bookmark79) 57
     1. [Simulation setups](#_bookmark80) 57
     2. [Results](#_bookmark81) 57
     3. [Discussion](#_bookmark95) 69

1. [MIMIC-III](#_bookmark96) 71
   1. [Introduction](#_bookmark97) 71
   2. [Source of Data](#_bookmark98) 71

[3.3 ICD-9](#_bookmark99) 72

* 1. [Study Population](#_bookmark102) 74
  2. [Study Variables](#_bookmark103) 74
  3. [Model](#_bookmark104) 75
  4. [Anomaly](#_bookmark105) 77
  5. [Statistical Analyses](#_bookmark106) 78
  6. [Results](#_bookmark107) 79
  7. [Discussion](#_bookmark133) 99

1. [Conclusion](#_bookmark134) 101
   1. [Introduction](#_bookmark135) 101
   2. [Summary of fi](#_bookmark136) 101
   3. [extensions on methodology](#_bookmark137) 102
   4. [Implications](#_bookmark138) 102
   5. [Future Directions](#_bookmark139) 103
   6. [Concluding remarks](#_bookmark140) 104
2. [Codes](#_bookmark141) 105
   1. [Simulations](#_bookmark142) 105
      1. [rand.day.time](#_bookmark143) 106
      2. [rand.day](#_bookmark144) 106
      3. [simdata](#_bookmark145) 106
      4. [cumdata](#_bookmark146) 107
      5. [tabulatedata](#_bookmark147) 107
      6. [ggcumplot](#_bookmark148) 107
      7. [ggdailyplot](#_bookmark149) 107
      8. [adddaily.anomaly](#_bookmark150) 108
   2. [JAGS models](#_bookmark151) 109
   3. [Hierarchical Time-series Predictions](#_bookmark155) 112
   4. [Independent and Hierarchical Models](#_bookmark157) 112

**List of Figures**

* 1. [A two level hierarchical tree diagram](#_bookmark2) 1
  2. [Anomaly Visualisation from](#_bookmark6) [Rah](#_bookmark227)[tz](#_bookmark6) [(2015)](#_bookmark227) 7
  3. [Visualised comparison of frequentist and Bayesian results, from](#_bookmark19) [W](#_bookmark239)[olf-](#_bookmark19) [gang](#_bookmark239) [(2016)](#_bookmark239) 12
  4. [Modelling structure between regular and Hierarchical Bayesian Models,](#_bookmark22) [from](#_bookmark22) [W](#_bookmark239)[olfgang](#_bookmark22) [(2016)](#_bookmark239) 14
  5. [A 2 level hierarchical structure](#_bookmark29) 17
  6. [Top-down, Bottom up and Optimum reconciliation approach](#_bookmark33) 19
  7. [Entries of the raw dataset](#_bookmark39) 24
  8. [Conversion of a simple hierarchical structure to a matrix](#_bookmark40) 25
  9. [Entries of the daily count dataset](#_bookmark41) 26
  10. [Construction of the proportion vector](#_bookmark42) 26
  11. [Histo-density plot of the prior probability distribution of proposed](#_bookmark45) [prior density models. The red dashed line indicates the center point](#_bookmark45)

[of these distributions, which is 1.](#_bookmark45) 31

* 1. [Posterior distributions of diff t models for Total, with added anoma-](#_bookmark50)

[lies, and calculated with independent Bayes model](#_bookmark50) 34

* 1. [Posterior distributions of diff t models for A , with added anomalies,](#_bookmark52)

[and calculated with independent Bayes model](#_bookmark52) 35

* 1. [Posterior distributions of diff t models for AA, with added anoma-](#_bookmark54)

[lies, and calculated with independent Bayes model](#_bookmark54) 36

* 1. [Traces plot for category total of model 1 to model 6](#_bookmark55) 38
  2. [ACF plots for category total of model 1 to model 6](#_bookmark56) 38
  3. [Traces plot for category A of model 1 to model 6](#_bookmark57) 39
  4. [ACFs plot for category A of model 1 to model 6](#_bookmark58) 39
  5. [Traces plot for category AA of model 1 to model 6](#_bookmark59) 40
  6. [ACFs plot for category AA of model 1 to model 6](#_bookmark60) 40
  7. [Heat-map of anomaly alarms for diff t size of anomaly under diff](#_bookmark72)

[ent k values, of IBM and HBM, at category total.](#_bookmark72) 47

* 1. [Heat-map of anomaly alarms for diff t size of anomaly under diff](#_bookmark73) [ent k values, of IBM and HBM zoomed in at k values of 0.6 to 1, at](#_bookmark73)

[category total.](#_bookmark73) 48

* 1. [Heat-map of anomaly alarms for diff t size of anomaly under diff](#_bookmark74)

[ent k values, of IBM and HBM, at category A.](#_bookmark74) 50

* 1. [Heat-map of anomaly alarms for diff t size of anomaly under diff](#_bookmark75) [ent k values, of IBM and HBM zoomed in at k values of 0.6 to 1, at](#_bookmark75)

[category A.](#_bookmark75) 51

* 1. [Heat-map of anomaly alarms for diff t size of anomaly under diff](#_bookmark76)

[ent k values, of IBM and HBM, at category AA.](#_bookmark76) 53

* 1. [Heat-map of anomaly alarms for diff t size of anomaly under diff](#_bookmark77) [ent k values, of IBM and HBM zoomed in at k values of 0.6 to 1, at](#_bookmark77)

[category AA.](#_bookmark77) 54

* 1. [Heat-map of anomaly signals for diff t diff t hierarchical struc-](#_bookmark89) [ture, of IBM and HBM at category total.](#_bookmark89) 61
  2. [Heat-map of anomaly signals for diff t diff t hierarchical struc-](#_bookmark90) [ture, of IBM and HBM zoomed in at k values of 0.6 to 1, at category](#_bookmark90) [total.](#_bookmark90) 62
  3. [Heat-map of anomaly signals for diff t diff t hierarchical struc-](#_bookmark91) [ture, of IBM and HBM, at category A.](#_bookmark91) 64
  4. [Heat-map of anomaly signals for diff t diff t hierarchical struc-](#_bookmark92) [ture, of IBM and HBM zoomed in at k values of 0.6 to 1, at category](#_bookmark92)

[A.](#_bookmark92) 65

* 1. [Heat-map of anomaly signals for diff t diff t hierarchical struc-](#_bookmark93) [ture, of IBM and HBM, at category AA.](#_bookmark93) 67
  2. [Heat-map of anomaly signals for diff t diff t hierarchical struc-](#_bookmark94) [ture, of IBM and HBM zoomed in at k values of 0.6 to 1, at category](#_bookmark94)

[AA.](#_bookmark94) 68

* 1. [Visualisation of ICD-10 Code Structure, from](#_bookmark100) [Andrus](#_bookmark163) [and Bohnett](#_bookmark100) [(2013).](#_bookmark163) 72
  2. [Posterior distributions for ICD 410,415 and 452 and its ancestors, with](#_bookmark110)

[no added anomalies, and calculated with Independent Bayes Model](#_bookmark110) . 80

* 1. [Posterior distributions for ICD 410,415 and 452 and its ancestors, with](#_bookmark112)

[no added anomalies, and calculated with Hierarchical Bayes model](#_bookmark112) . . 81

* 1. [Posterior distributions for ICD 410 and its ancestors, with added](#_bookmark114) [anomalies, and calculated with independent Bayes model](#_bookmark114) 82
  2. [Posterior distributions for ICD 410 and its ancestors, with added](#_bookmark116) [anomalies, and calculated with Hierarchiacl Bayes model](#_bookmark116) 83
  3. [Posterior distributions for ICD 415 and its ancestors, with added](#_bookmark118) [anomalies, and calculated with independent Bayes model](#_bookmark118) 84
  4. [Posterior distributions for ICD 415 and its ancestors, with added](#_bookmark120) [anomalies, and calculated with Hierarchical Bayes model](#_bookmark120) 85
  5. [Posterior distributions for ICD 452 and its ancestors, with added](#_bookmark122) [anomalies, and calculated with independent Bayes model](#_bookmark122) 86
  6. [Posterior distributions for ICD 452 and its ancestors, with added](#_bookmark124) [anomalies, and calculated with Hierarchical Bayes model](#_bookmark124) 87
  7. [Heat-map of anomaly signals for ICD 410,415 and 452 and its ancestors,](#_bookmark125) [with no added anomalies, of IBM and HBM.](#_bookmark125) 88
  8. [Heat-map of anomaly signals for ICD 410,415 and 452 and their](#_bookmark126) [ancestors, with no added anomalies, of IBM and HBM zoomed in at k](#_bookmark126) [values of 0.9 to 1.](#_bookmark126) 89
  9. [Heat-map of anomaly signals for ICD 410 and its ancestors, with no](#_bookmark127) [added anomalies, of IBM and HBM.](#_bookmark127) 91
  10. [Heat-map of anomaly signals for ICD 410 and its ancestors, with no](#_bookmark128) [added anomalies, of IBM and HBM zoomed in at k values of 0.9 to 1.](#_bookmark128) 92
  11. [Heat-map of anomaly signals for ICD 415 and its ancestors, with no](#_bookmark129) [added anomalies, of IBM and HBM.](#_bookmark129) 94
  12. [Heat-map of anomaly signals for ICD 415 and its ancestors, with no](#_bookmark130) [added anomalies, of IBM and HBM zoomed in at k values of 0.9 to 1.](#_bookmark130) 95
  13. [Heat-map of anomaly signals for ICD 452 and its ancestors, with no](#_bookmark131) [added anomalies, of IBM and HBM.](#_bookmark131) 97
  14. [Heat-map of anomaly signals for ICD 452 and its ancestors, with no](#_bookmark132) [added anomalies, of IBM and HBM zoomed in at k values of 0.9 to 1.](#_bookmark132) 98

**List of Tables**

* 1. [DIC (with 1,000,000 iteration) summary distribution cross comparisons](#_bookmark48)

[for model 1 to model 6](#_bookmark48) 32

* 1. [Posterior distributions of diff t models for Total, with added anoma-](#_bookmark49)

[lies, and calculated with independent Bayes model](#_bookmark49) 34

* 1. [Posterior distributions of diff t models for A , with added anomalies,](#_bookmark51)

[and calculated with independent Bayes model](#_bookmark51) 35

* 1. [Posterior distributions of diff t models for AA, with added anoma-](#_bookmark53)

[lies, and calculated with independent Bayes model](#_bookmark53) 36

* 1. [DIC (with 1,000,000 iteration) comparisons between Independent Bayes](#_bookmark65) [model and Hierarchical Bayes model (HBM -IBM) for diff t anoma-](#_bookmark65)

[lies sizes](#_bookmark65) 42

* 1. [Posterior distributions of diff t models for Total, with diff t](#_bookmark66) [increments of added anomalies, and calculated with independent Bayes](#_bookmark66) [model](#_bookmark66) 43
  2. [Posterior distributions of diff t models for Total, with diff t](#_bookmark67) [increments of added anomalies, and calculated with Hierarchical Bayes](#_bookmark67) [model](#_bookmark67) 43
  3. [Posterior distributions of diff t models for A , with diff t incre-](#_bookmark68) [ments of added anomalies, and calculated with Independent Bayes](#_bookmark68) [model](#_bookmark68) 44
  4. [Posterior distributions of diff t models for A , with diff t in-](#_bookmark69) [crements of added anomalies, and calculated with Hierarchical Bayes](#_bookmark69) [model](#_bookmark69) 44
  5. [Posterior distributions of diff t models for AA, with diff t in-](#_bookmark70) [crements of added anomalies, and calculated with Independent Bayes](#_bookmark70) [model](#_bookmark70) 45
  6. [Posterior distributions of diff t models for AA, with diff t in-](#_bookmark71) [crements of added anomalies, and calculated with Hierarchical Bayes](#_bookmark71) [model](#_bookmark71) 45
  7. [DIC comparisons for brunching number at A](#_bookmark82) 57
  8. [Posterior distributions of diff t models for Total, with diff t](#_bookmark83) [brunch number at A, and calculated with Independent Bayes model](#_bookmark83) . 58
  9. [Posterior distributions of diff t models for Total, with diff t](#_bookmark84) [brunch number at A, and calculated with Hierarchical Bayes model](#_bookmark84) . 58
  10. [Posterior distributions of diff t models for A , with diff t brunch](#_bookmark85) [number at A, and calculated with independent Bayes model](#_bookmark85) 59
  11. [Posterior distributions of diff t modelsfor A , with diff t brunch](#_bookmark86) [number at A, and calculated with Hierarchical Bayes model](#_bookmark86) 59
  12. [Posterior distributions of diff t models for AA, with diff t brunch](#_bookmark87) [number at A, and calculated with independent Bayes model](#_bookmark87) 60
  13. [Posterior distributions of diff t modelsfor AA, with diff t brunch](#_bookmark88) [number at A, and calculated with Hierarchiacl Bayes model](#_bookmark88) 60
  14. [ICD-9 Diagnosis Codes](#_bookmark101) 73
  15. [DIC comparisons for Independent Bayesian Model and Hierarchical](#_bookmark108) [Bayesian model for no anomaly, and anomaly added for common, rare](#_bookmark108)

[and extremely rare diseases](#_bookmark108) 79

* 1. [Posterior distributions for ICD 410,415 and 452 and its ancestors, with](#_bookmark109)

[no added anomalies, and calculated with Independent Bayes Model](#_bookmark109) . 80

* 1. [Posterior distributions for ICD 410,415 and 452 and its ancestors, with](#_bookmark111)

[no added anomalies, and calculated with Hierarchical Bayes model](#_bookmark111) . . 81

* 1. [Posterior distributions for ICD 410 and its ancestors, with added](#_bookmark113) [anomalies, and calculated with independent Bayes model](#_bookmark113) 82
  2. [Posterior distributions for ICD 410 and its ancestors, with added](#_bookmark115) [anomalies, and calculated with Hierarchiacl Bayes model](#_bookmark115) 83
  3. [Posterior distributions for ICD 415 and its ancestors, with added](#_bookmark117) [anomalies, and calculated with independent Bayes model](#_bookmark117) 84
  4. [Posterior distributions for ICD 415 and its ancestors, with added](#_bookmark119) [anomalies, and calculated with Hierarchical Bayes model](#_bookmark119) 85
  5. [Posterior distributions for ICD 452 and its ancestors, with added](#_bookmark121) [anomalies, and calculated with independent Bayes model](#_bookmark121) 86
  6. [Posterior distributions for ICD 452 and its ancestors, with added](#_bookmark123) [anomalies, and calculated with Hierarchical Bayes model](#_bookmark123) 87
  7. [Code used to specify Independent Bayes Models for Chapter 2.3 a](#_bookmark152) . . 109
  8. [Code used to specify BHMs for Chapter 3 b](#_bookmark153) 110
  9. [Code used to specify BHMs for Chapter 3 c](#_bookmark154) 111
  10. [Example of R code used to make time-series prediction used in Chapter 3](#_bookmark156)112
  11. [JAGS code used for the Independent Bayesian Model used in Chapter 3](#_bookmark158)112
  12. [JAGS code used for the Hierarchical Bayesian Model used in Chapter 3](#_bookmark159)113

**Chapter 1 Background**

* 1. **Introduction**

Large datasets often have a hierarchical structure (as seen in fi [1.1](#_bookmark2)), in other words, data may organised into a tree-like structure, where broader categories with higher hierarchy may disaggregate into smaller categories with lower hierarchy. Vice versa, smaller categories with lower hierarchy may aggregate into broader categories with higher hierarchy. In a clinical trial, Adverse Events (AEs) caused by treatment drugs can be aggregate into body systems that are associated with AEs, thus creating a 2-level hierarchical structure. An example of a study considering such hierarchical structure is [Berry and Berry](#_bookmark166) [(2004),](#_bookmark166) the study considers the probabilities of treatment drug cause an AEs and found signifi t diff es in results between solo ( with each category independently modelled) and Hierarchical Bayesian Model (HBM) at the bottom level of the hierarchy, and suggested that the HBM could be a solution to the multiplicity problem found in large datasets. Disease diagnoses also have a complex, hierarchical structure, where diff t categories of diagnoses aggregated into broader categories. Standard such as International Statistical Classification of Diseases and Related Health Problems [(ICD)(Organization,](#_bookmark217) [2019)](#_bookmark217) are used to code for and organise disease diagnosis into a hierarchical structure. Information about ICD

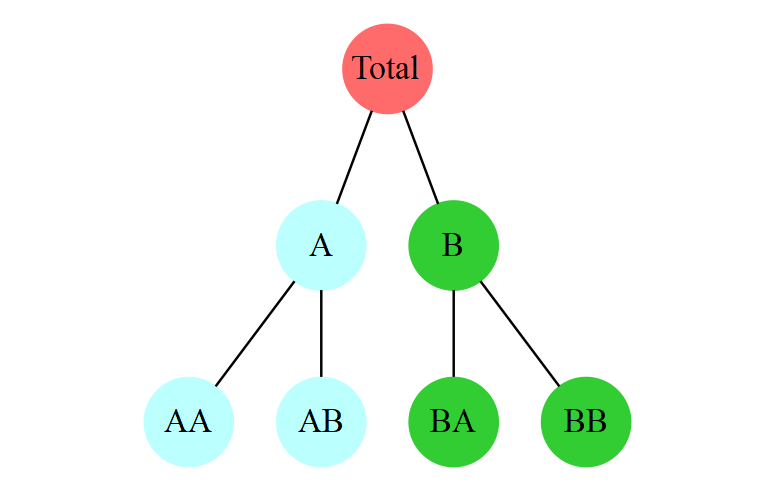


Figure 1.1: A two level hierarchical tree diagram

diagnoses often included in hospital archives for various managerial purposes. Note that ICD diagnosis is not the equivalent to original diagnosis by doctors, descriptions of diagnoses can vary doctor to doctor, it is non-consistent and unorganised at fi this diagnosis is then classified into standard ICD diagnosis which is standard, organised and has hierarchical properties.

We are interested in whether modelling the full hierarchy gives a better under- standing of anomalies compares to modelling each category independently one level at a time. When modelling too broad a category, individual variations could be missed; when model too narrow a category, there may be too little data with little significance. Vital information are likely be lost if we only consider one level at a time if the information doe not situated at the level considered, so why not consider every possible level? A hierarchical model retains information from all level of the hierarchy and may provide better predictions compare to standard models that do not implicate the hierarchical structure of the data. The purpose of this research thesis is to investigate the use of Hierarchical Bayesian Models (HBM) as a possible approach for the detection of anomalies from hierarchical datasets. In this chapter, we describe examples of where these issues arise, and layout the technical background for our methods.

* 1. **Emergency Department Congestion**

While the world population continues to grow and age at an unprecedented rate, public healthcare system all over the world are expected to experience an increasing number of situations where a massive infl of patients overwhelms the current healthcare capability and in which congestion occur [(Morley et al.,](#_bookmark214) [2018;](#_bookmark214) [Wong et al.,](#_bookmark240) [2010).](#_bookmark240) Other similar terms used to describe this situation include bottleneck and overcrowding. Congestion is shown to have negative infl es on both patient and hospital staff, studies [(Chambers et al.,](#_bookmark173) [2016;](#_bookmark173) [Khare et al.,](#_bookmark202) [2009;](#_bookmark202) [Salway et al.,](#_bookmark229) [2017)](#_bookmark229) suggest that congestion is positively associated with adverse outcomes such as patient waiting time, short term patient mortality, medical errors, and professional burnout in hospital staff; therefore we would want to avoid or reduce the possibility of occurrence of congestions. Congestion seriously challenges the capability of public healthcare system worldwide and are becoming increasingly burdensome in many locations, and efforts from healthcare systems are struggling to kept pace with increasing demands. The overall consensus is that the demand from the ageing population is one of the major contributing factors to the current crisis, and the factor of ageing is unlikely to improve by time [(Bayley et al.,](#_bookmark165) [2005;](#_bookmark165) [Cornwall and Davey,](#_bookmark178) [2004;](#_bookmark178) [Wong et al.,](#_bookmark240) [2010).](#_bookmark240) For instance, the national population projections by Stats New Zealand (2016) predicts a constant increasing trend in older (age 65 +) populations and show no

sign of slowing down, so the problem of congestion in New Zealand in another hand, will most likely to correspond to this and worsen.

Shortage of staff for public healthcare services has been a prominent issue ever since a decade of decrease in the proportion of healthcare funding and defi for the public health system in New Zealand [(Akoorie,](#_bookmark161) [2018;](#_bookmark161) [The World Bank,](#_bookmark236) [2019).](#_bookmark236) Due to limitations with medical material and human resources, hospital staff often had to operate at full capacity and are therefore under tremendous stress. Studies have shown that under-staffi is one of the key factors that contribute to the high prevalence of professional burnout in New Zealand’s public hospital senior medical workforce [(Chambers et al.,](#_bookmark173) [2016).](#_bookmark173) Described as “erosion of the soul” by Doctor Len Gabbard, in an article by [HemOnc Today](#_bookmark192) (2008), professional burnout is a term used in industry research that refers to a sense of emotional exhaustion, negative attitudes and sense of incompetence that often lead to reduced work effectiveness (P[aterson](#_bookmark220) [and Adams,](#_bookmark220) [2011](#_bookmark220)). Therefore burnout has negative impacts and to reduce this; we would need to optimise the way the hospital operates.

Due to the increasing patient infl and understaffi some Emergency De- partment (ED) in the country struggled to cope with increasing patient arrivals; this is reflected from unacceptable delays before patients are admitted to hospital, transferred or sent home. Patient delay is a direct result of triage strategy [(Ministry](#_bookmark210) [of Health,](#_bookmark210) [2011)](#_bookmark210) utilised by ED if congestion of patients occurs. When there is not enough staff to cope with patient demand, the patient at a higher triage is given a priority for treatment (higher the triage level, the more critical patient is believed to be). However, a critical fl w of the current strategy is that if congestion does not get resolved, and lots of patients require immediate treatment, a patient that is evaluated to be less urgent often gets delay after delay. In theory, a patient can be delayed for infi y if he is unfortunate enough, and the chance of further complication developing during the waiting period will only rise at every second of the clock ticking by.

A tragic case happened in New Zealand, back in 2015, where a woman in Wellington died after 12 hours of delay and ignorance in emergency medical service [(Southland](#_bookmark231) [times,](#_bookmark231) [2015).](#_bookmark231) The news reported that *“The emergency department was stretched by an influx of patient which lead to congestion and overcrowding”*, so ED staff had decided that the patient was not in an immediately life-threatening situation and decided to apply the triage strategy and therefore delayed medical service. What supposed to be a 30-minute delay had turned into more than an hour, and over the last 12 hour of the patient’s life, ED staff failed to provide adequate monitor and care, which was believed to have contributed to the patient’s untimely demise. It breaks your heart when you hear such case happening and you cannot stop wondering, what

if the patient was treated in time?

* 1. **Motivation**

Located at the start of the healthcare pathway [(Ministry of Health,](#_bookmark212) [2016),](#_bookmark212) the Emergency Department (ED) and other forms of emergency health care are feeling tremendous pressure from increasing healthcare demands. Compare to primary healthcare, emergency healthcare is a 24/7 operating service characterised with unscheduled visits, and mostly deals with critical conditions that need immediate attention [(Ministry of Health,](#_bookmark209) [2009,](#_bookmark209) [2015),](#_bookmark211) therefore ED frequently encounters an infl of unanticipated emergencies such as major traffic accident, mass food poisoning and disease outbreaks. Unfortunately, due to limited and often diminishing resources, public healthcare facilities often had to operate at near full capacity to reduce operation costs [(Salway et al.,](#_bookmark229) [2017),](#_bookmark229) and this leaves a very small buff zone to deal with abnormally high arrivals, so it is quite safe to suggest that ED staffs are underprepared for congestion.

The congestion crisis in the New Zealand healthcare service stood well recognised by the academic community, and there is a substantial amount of recent and ongoing research that aims at providing alternative operational strategies. For instance, [Adams et al.](#_bookmark160) [(2018](#_bookmark160)) is currently working on simulations of Auckland city hospital’s Emergency Department with Pod model, where doctors form small cross-functional and multidisciplinary teams. Experimental trials had taken place at the Auckland city hospital during December 2018 - January 2019 to evaluate the effectiveness of Pods strategy compare to the precedent triage strategy. Another interesting study is a study by [Cleland et al.](#_bookmark175) [(2018)](#_bookmark175) that looked at developing algorithms to optimise staff rostering, compares to precedent; his algorithm found many new ways staff rostering can be improved and thus could improve the operational effectiveness of hospital. As we can see, we do have people approaching the problem in various ways.

For this thesis, we want to approach the problem of congestion in public hospitals through the perspective of Healthcare Logistics. As Prof.Dr. Stephan Nickle from Karlsruhe Institute of Technology (KIT) Quoted during the 2018 Operations Research Society of New Zealand (ORSNA) conference [(Nickel,](#_bookmark215) [2018),](#_bookmark215) *“Healthcare Logistics is all about the robustness and smoothness of the operation”*, what we want is stability and predictability of the operation, even if on paper we have a high efficiency, it is no good if this efficiency is often compromised. ED congestion often occurs during the periods of abnormal arrivals when the max capacity is overwhelmed; therefore, a success prediction/anticipation of such abnormal arrivals could provide valuable insight for operational planning and management. Public hospital was supposedly

prepared, medical facilities continually being built, medical supplies prepared in advance, and staff were recruited base on predictions from historical mean arrivals. It would make sense to the thought that the public health system was well prepared, right? However, why are we still seeing congestions? Well, what happened is that the public health system had well prepared for normality, but has overlooked instances of abnormality. Lots of operational and logistics planning and based on mean values, however, means does not contain information about abnormality. To the inconvenience of everybody involved, congestion occurs during abnormal arrival events.

If we could make sense of and make informed predictions about abnormal patient arrivals, it would provide helpfulness to the planning and management of the public healthcare system. Doctors can transfer, and medical supplies can be in preparation as soon as there is enough evidence suggesting an abnormal event is or will be happening. With better preparation and faster response to abnormal patient arrivals, the problem of congestion could be alleviated, and benefit both patient and healthcare staff. For patients, a shorter delay waiting period will increase the quality and timeliness of the medical service provided and reduce the risk of further complications during the wait. Moreover, for doctors, more preparedness should reduce professional burnout and increase efficiency (C[hambers et al.,](#_bookmark173) [2016;](#_bookmark173) [Khare et al.,](#_bookmark202) [2009).](#_bookmark202) For this thesis, a strong emphasis is placed on modelling anomaly detection for patient arrivals, in the hopes of that extension on the understanding of anomalous arrival events could be made, and provide information for theoretical and real-life implications.

* 1. **Anomaly Detection**

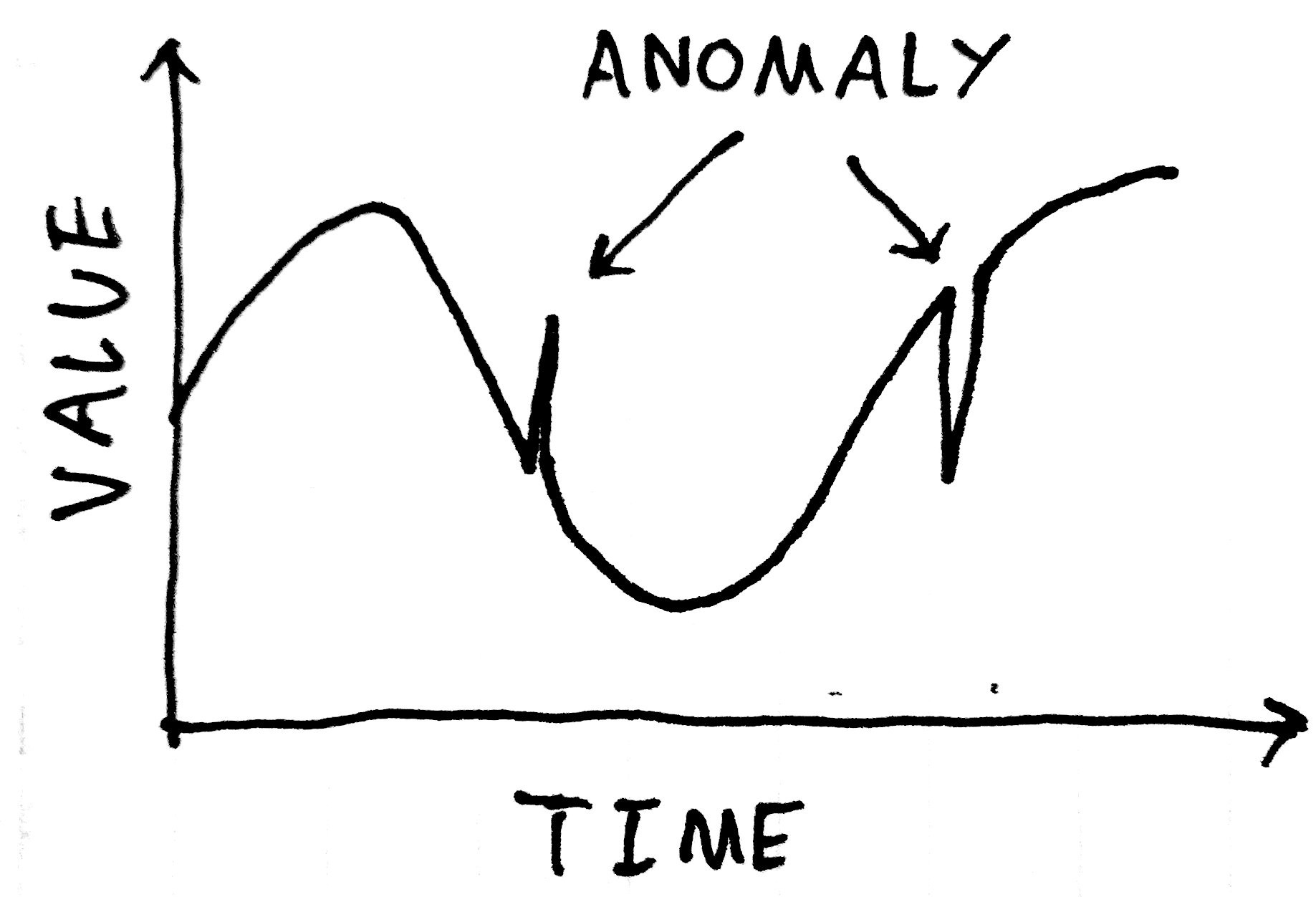
Anomaly detection refers to the fi of studies of fi patterns in data that does not conform to the expectation, or that is out of normality. Studies of anomaly can trace back to the 19th century; an example is that in 1887, F. Y. Edgeworth a lecturer from King’s College, London wrote about what he termed as discordant observations, observations which *“present the appearance of differencing in respect of their law of frequency from other observation with which they are combined”* [(Edgeworth,](#_bookmark182) [1887).](#_bookmark182) The term anomaly is quite similar to the term outlier, which is defi by “A data point that diff significantly from other observations” [(Grubbs,](#_bookmark190) [1969),](#_bookmark190) and these two terms often used interchangeably. However, the term anomaly has a stronger contextual connection to the application side of the statistics, the wording itself literally meant, an event that is not of normal circumstance.

Anomaly detection found extensive application in fi such as cyber-security, fi and healthcare. The reason why anomaly is catching so much research interest is that it often strongly associate with and directly translates into the occurrence

of harmful situations and events in dynamic environments [(Chandola et al.,](#_bookmark174) [2009;](#_bookmark174) [La Rosa,](#_bookmark204) [2018).](#_bookmark204) The classic example of anomaly detection often described by Data scientists is the application of intrusion detection systems that are triggered by anomalies that are believed to be indicators of cyber-attacks. An indicator such as abnormally high website visits could be an indication of an HTTP POST attack, where the website processing capacity is overwhelmed, or maybe illegal data scraping, where private information is processed and exported from a website (P[atcha and](#_bookmark219) [Park,](#_bookmark219) [2007;](#_bookmark219) [Target Internet,](#_bookmark235) [2018).](#_bookmark235) Due to growing concerns of potential bioterrorism attack following the anthrax letters attacks to various government officials during 2001, the US had been invested in the fi of Biosurveillance in preparation for a potential bioterrorism attack (G[rundmann,](#_bookmark191) [2014).](#_bookmark191) Therefore a substantial amount of research regarding population-level anomaly detection is available. For instance, back in 2003, [(Wong et al.,](#_bookmark241) [2003)](#_bookmark241) has presented the WSARE 3.0 algorithm, which stands for ”What’s Strange About Recent Events, version 3” for the detection of disease outbreaks in the US. Also, in *Intelligence and Security Informatics: Biosurveillance*, a text published in 2007, [Bauer et al.](#_bookmark164) [(2007)](#_bookmark164) suggest the need of rapid data processing for anomaly detection to meet the increasing demand to disease surveillance at a population level. Their idea foreshadowed the recent rise in popularity of online streaming data systems that collect data from multiple sources continuously and provide statistical results right with each update.

For the case of patient arrivals, researchers often take an interest in time series data. That is, a continuing set of data that is recorded over specific time intervals [(Das,](#_bookmark179) [1994).](#_bookmark179) If information about the time of each patient visit is available, it can be counted to give a time series data over time-periods, such as the number of patient arrivals per day. From the perspective of time series data, there are several important types of anomalies to be considered. The fi and most straightforward type is point anomalies, where a single time-period is anomalous. For example, traffic of a website may spike during a single night, when the website is under cyber-attack. The second and more complex type is period anomalies, where a continuing series of time-periods is anomalous. For example, the average daily spending during the Christmas period (let us say, the days between 1 December to 31 December) is expected to be much higher when compared to the average usual daily spending in other time frames. The last type, Contextual anomalies, where the occurrence of the anomalous time-period is believed to be context-specific. For example, the Admission rates for Emergency department patients may spike on any given day, when there is a mass injury event such as major car crashes [(Chandola et al.,](#_bookmark174) [2009).](#_bookmark174) In this thesis, we want to focus on just the point anomaly, because other forms of anomalies are very context-specific and hard to perform in simulated settings.

There are many ways to decide whether a point or a pattern is abnormal and deviate from the expected; the most straightforward and often used approach is to set up value as thresholds, and then anomalies can be detected simply by raising alarms for observing values beyond the threshold (figure [1.2).](#_bookmark6)



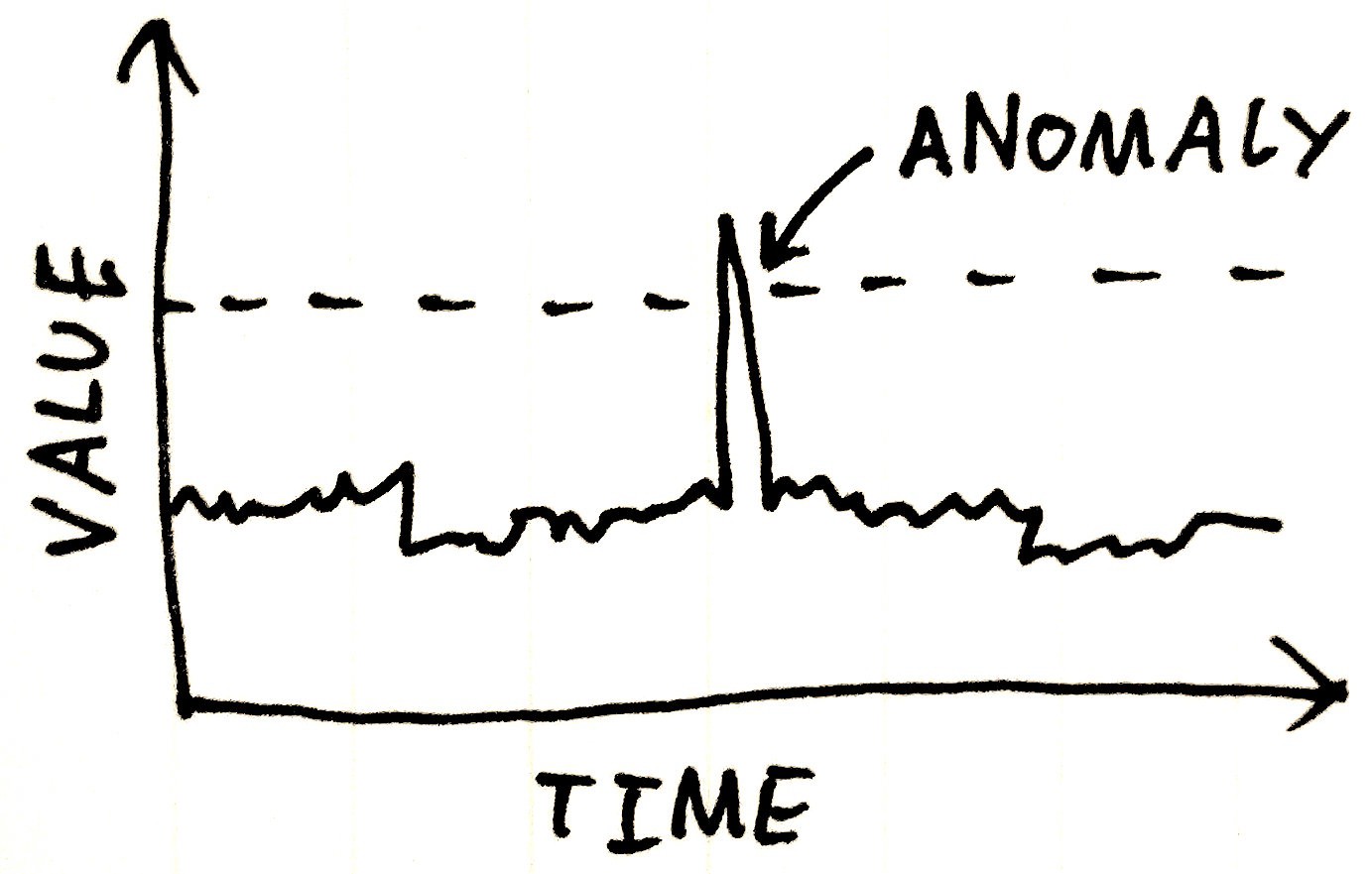
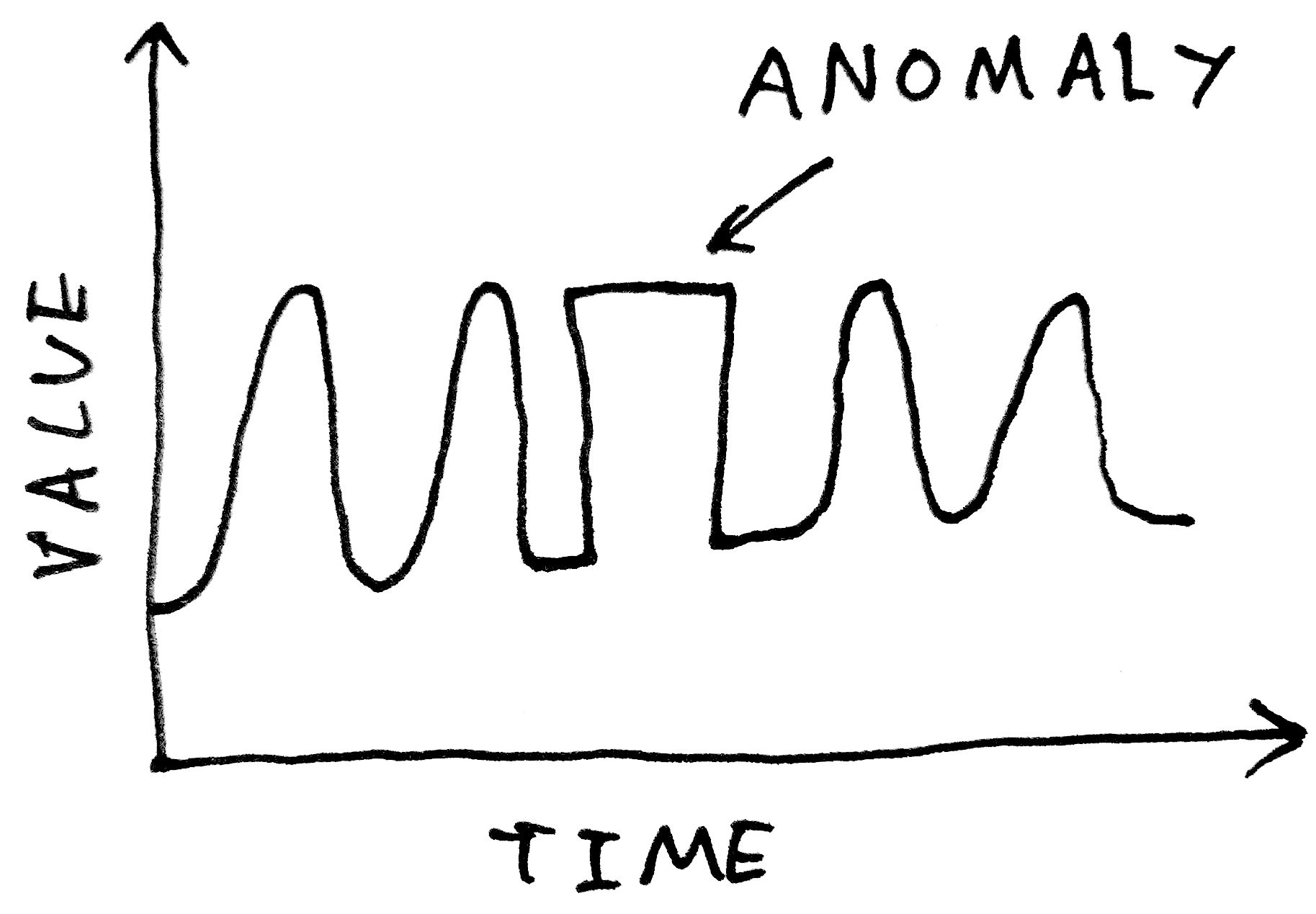
 

Figure 1.2: Anomaly Visualisation from [Rahtz](#_bookmark227) [(2015)](#_bookmark227)

[Wong et al.](#_bookmark241) [(2003)](#_bookmark241) considered the threshold approach and proposed the standard algorithm. Their anomaly detector considers the mean *µ* and variance *σ*of the total number of records on each day, and used algorithm shown in equation [1.1](#_bookmark7) to calculate a threshold, For the algorithm Φ*−*1 is the inverse to the cumulative distribution function of a standard normal with an user-defi p-value. If the aggregate daily counts of health care data exceed this threshold during the evaluation period, an alarm can raise.

*threshold* = *µ* + *σ ∗* Φ*−*1(

1

p-value

) (1.1)

We propose a simpler algorithm that approaches the threshold set up from a more practical point of view, rather than a theoretical statistical value such as Φ*−*1 used by [Wong et al.](#_bookmark241) [(2003).](#_bookmark241) We want to create a value called capacity (*k*), this number should reflect the average proportion of treatment capacity a department is expected to operate at during specific period. In theory, this number could generate from the number of hospital beds occupied, the number of doctors that are currently treating a patient and not available for another patient in a short notice, and any other available measures of hospital operational capacity. If we assume that the healthcare system is expected to be operating at a capacity of 0.8, it would mean that on average, 80% of total possible treatment resource during a specific period is utilised. For our algorithm shown in equation [1.2,](#_bookmark8) the threshold calculation comes from the product of average expected patient arrival and a threshold multiplier, *k−*1.

*Threshold* = *µk−*1 (1.2)

So if a hospital department on average treats 100 patient per day, and this department is considered to be operate at 80% capacity, the threshold would be 100 *×* 0*.*8*−*1 = 100 *×* 1*.*25 = 125. The idea is that when we reached the threshold, it

would mean that the healthcare capacity of the period had reached 100%, and there is no more capacity for more patients, and in theory, this is when congestion would for defi , occur. Instead of only considering operating cost, an ideal operating percentage should be relatively high to reduce operating cost, and have gives a low chance of overcapacity due to the anomalous event. Following up on the idea by Stephan Nickle, mentioned in the previous section, we propose that increasing the operational capacity without considering anomalous events could jeopardise the smoothness of operation, and cause even more loss of operating efficiency due to increased probability congestions.

To create an alarm system that raises alarms for hospital congestion, we could apply a threshold on to patient arrivals, with every increase in the number of patients arrived, it would become more likely that a threshold can exceed. However, the catch is that we should also incorporate external information such as the number of bed available instead of only considering with numerical probability. How could we possibly know if we have reached the full hospital capacity if we only consider the number of beds used, without knowledge of the number of total beds available? With only information about observed arrivals, we could work out the likelihood of observing specific arrival numbers, but this outcome does not make any practical sense by itself. Information is one of the reasons why the standard algorithm proposed by [Wong et al.](#_bookmark241) [(2003)](#_bookmark241) although performed excellently, still seem weak for practical applications. The standard algorithm considers the threshold as the im-probability from the data structure and does not consider what determines the threshold, the ability or the capacity for hospitals to treat patients. A particular statistical inference approach does incorporate external information and solves this problem, as people might have guessed already, Bayes.

* 1. **Hierarchical Bayesian Models**
     1. **The history of Bayesian statistics**

The term Bayesian refers to a range of statistical techniques that approach statistical inference. The term Bayesian makes tributes to the 18th-century mathematician, the late Reverend Thomas Bayes (1702-1751) [(Stigler,](#_bookmark233) [2002).](#_bookmark233) Bayes theory was published after Bayes’s death and was pioneered and popularised by the much-acclaimed Mathematician Pierre-Simon Laplace (1749-1827) in his 1814 ”Philosophical essay on probabilities” [(McGrayne,](#_bookmark207) [2011).](#_bookmark207) The theorem was non-popular at fi as statistical solutions were often impossible to obtain, apart from conjugate distributions [ref pg 493 of stats 732 note], which refers to solutions in the same algebraic form of prior with alteration in parameters. The development of Markov chain Monte Carlo

(MCMC) techniques in 1940 and generalisation of personal computing device provided the solution to the fatal fl w, application of Bayesian methods was seen since 1990 and is now becoming ever popular and poses a severe challenge to the classical frequentist approach to statistical inference [(Allenby et al.,](#_bookmark162) [2005;](#_bookmark162) [McGrayne,](#_bookmark207) [2011).](#_bookmark207)

* + 1. **Bayesian inference**

*How do you tell if a statistician is a true Bayesian? Ask them what time it is.*

*If they tell you the time, they’re a frequentist;*

*A Bayesian will ask “What time do you think it is?”.*

— Geoff [Jones](#_bookmark200) [(2018)](#_bookmark200)

The Bayesian approach to statistical inference is a form of evidential probability, where repeatable observations serve as evidence in the determination of probability characteristics of certain events (P[aulos,](#_bookmark221) [2011;](#_bookmark221) [Wheeler and Williamson,](#_bookmark237) [2011).](#_bookmark237) The principle is, given information of past event as prior probability and taken into account current evidence as likelihood, Bayesian are able to update and deduce the posterior probability, an evidence based proposition [(Laplace,](#_bookmark205) [1998).](#_bookmark205) The essence of Bayesian inference lies within the Bayes’ theorem (alternatively Bayes’ law, Bayes’ rule)(Brew[er,](#_bookmark170) [2017),](#_bookmark170) it can summarised mathematically as the following equation:

likelihood

prior

P,. r(""*B*'- *A*"") ,.

""'- ""

Pr(*A|B*) =

*|* Pr(*A*)

(1.3)

'-pos""te,.rior""

Pr(*B*)

'- "",. ""

marginal likelihood

Given two events *A* and *B*, Pr(*A|B*) is the conditional likelihood of event A occurring given that B is true, which represents the posterior. Pr(*B|A*) is the conditional likelihood of event B occurring given that A is true, and this represents the likelihood. Pr(*A*) and Pr(*B*) are the probabilities of observing A and B independently of each other; this is the marginal likelihoods. Pr(*A*) represents the prior, and Pr(*B*) is a measure of evidence, without conditioning on prior.

In probability statistics, the characteristic of a statistical population or a statistical model can be quantified with a number of parameters, often noted as *θ* the parameter vector, where *θ* = *{θi...θp}*. For example, the parameter *λ* characterised a Poisson distribution, where *λ* can be interpreted as the average number of events per interval. The probability distribution function of Poisson can be given an equation with the parameter *λ* and a vector of observed data as:

*P* (*x*) =

*e−λλx x*!

*.* (1.4)

If we consider Θ as our parameter and *X* as our observations, in terms of the generalised distribution function, Bayes theorem can express as

*fX|*Θ(*x, θ*)

*f*Θ*|X* (*θ|x*) =

*fX* (*x*)

(1.5)

= *fX|*Θ(*x|θ*)*f*Θ(*θ*)

*fX* (*x*)

For equation [1.5](#_bookmark13) *f*Θ*|X* (*θ|x*) is the posterior density, *fX|*Θ(*x, θ*) is the joint density of *x* and Θ, which can be transformed into conditional density *fX|*Θ(*x|θ*) and marginal density *f*Θ(*θ*), lastly we have the marginal density of *fX* (*x*).

*fX|*Θ(*x|θ*)*f*Θ(*θ*)

*f*Θ*|X* (*θ|x*) =

*fX* (*x*)

*fX|*Θ(*x|θ*)*f*Θ(*θ*)

*∝* { *fX|*Θ(*x|θ*)*f*Θ

*L*(*θ*)*f*Θ(*x*)

(*θ*)*dx*

(1.6)

*∝ c*

Pr(*θ|x*) *∝* Pr(*x|θ*) Pr(*θ*)

### Posterior *∝* Likelihood *×* Prior

If we consider the case of continuous distributions in equation [1.6,](#_bookmark14) *fX* (*x*) is the expectation of X, which can be expressed as integral { *fX|*Θ(*x|θ*)*f*Θ(*θ*)*dx* and we

know it is a constant, *c*. *fX|*Θ(*x|θ*) also correspond to the Likelihood function *L*(*θ*) as a function of *θ*. After these transformation we see the proportional property of *fX|*Θ(*x|θ*) ,*L*(*θ*) and *f*Θ(*x*). We can derive the conclusion that posterior probability distribution of the parameter of interest is proportional to the product of prior probability distribution and observed evidence, and the posterior is an adequate estimation that reflects the probability characteristics of certain event or distributions [(Wong,](#_bookmark242) [2019).](#_bookmark242) Overall, this is the theorem behind Bayesian inference, given prior and likelihood, we can deduce the posterior.

* + 1. **Simulation-based estimation**

As we had mentioned in the previous section, in Bayesian statistics, the posterior distribution can be manually calculated to obtain conjugate solutions. The basic idea is that a standard distribution can work as the approximation of prior distribution, and with certain likelihood functions a posterior distribution with a distribution from the same family can be formed, and this made it possible to derive solutions as going from

prior to posterior distribution is simply a matter of fi out how the parameter of the distribution had changed. [(Wong,](#_bookmark242) [2019).](#_bookmark242) For example, a *Beta*(*α, β*) prior can be multiplied with a *Binomial*(*n, x*) likelihood to form a *Beta*(*α* + *x, β* + *n − x*) posterior distribution with some mathematical manipulations, shown here in equation [1.7.](#_bookmark16)

Γ(*α* + *β*) Γ(*α*)Γ(*β*)

*θα−*1(1 *− θ*)*β−*1 *×*

*n*

*x*

*θx*(1 *− θ*)*n−x*

(1.7)

=Constant *× θα*+*x−*1(1 *− θ*)*β*+*n−x−*1

However, for cases with non-conjugate prior distribution, it is challenging to perform mathematical calculations; therefore, alternative methods are often used to calculate the characteristics of the posterior distributions. A way to calculate the posterior distribution is to approach with a simulation-based estimation method, for which multiple generations of repeated samples can be used to converge into a reasonably estimation [(Congdon,](#_bookmark176) [2007).](#_bookmark176) Markov chain Monte Carlo (MCMC) methods are often utilised by Bayesian to obtain repeated samples from a probability distribution, the idea of a Markov chain is that the probability characteristics of each draw depend only on the state in the previous event (F[ewster,](#_bookmark183) [2014).](#_bookmark183) We could represent the Markov property in mathematical notation as following equation [1.8,](#_bookmark17) where a sample from time *t* + 1 only depends on a sample from time *t* and all previous observation does not matter.

Pr(*Xt*+1*|Xt, ..., X*0) = Pr(*Xt*+1*|Xt*) (1.8)

The Metropolis-Hastings algorithm is a family of MCMC methods useful for sampling from probability distributions from which direct sampling was not possible. [Hastings, 1970] And the Gibbs sampler is a particular case of the Metropolis-Hastings algorithm often used in Bayesian statistics. Just Another Gibbs Sampler (JAGS) is a program designated for the analysis of Bayesian models using MCMC simulation [(Plummer et al.,](#_bookmark223) [2003).](#_bookmark223) JAGS can work within the R language and environment, with the use of rjags [(Plummer,](#_bookmark222) [2018)](#_bookmark222) and or R2jags [(Su and Yajima,](#_bookmark234) [2015)](#_bookmark234) package from the Comprehensive R Archive Network (CRAN), using the statistical software, R (R [Core Team,](#_bookmark226) [2013).](#_bookmark226) Details about download and installation of JAGS and R is in official websites, and examples of a JAGS script used to simulate results used for this thesis is presented in the Appendix.

* + 1. **Bayesian against Frequentist**

There are two significant approaches to statistical inference, the frequentist approach, and the Bayesian approach. The frequentist school place emphasises on the sampling distribution of the observed data, Frequentist treats observations as a random ex- periment with fi parameters across random samples and aims to calculates the im-probability of obtaining an observed result. By contrast, the Bayesian school base their inference on Bayes formula (equation [1.](#_bookmark12)3), Bayesian treat observations to be fi and assume randomness for model parameters, and aims to update prior into posterior. Thus a significant diff between frequentist and Bayesian is that they give a diff t interpretation of results, as shown in fi [1.3,](#_bookmark19) frequentist often give a p-value where Bayesian often present posterior probability. Each approach has its strength and weaknesses. However, the Bayesian approach seems to be a more suitable option for our analysis in many ways.

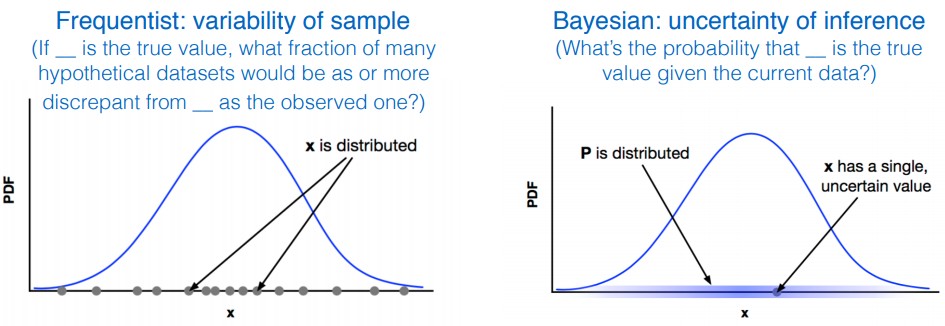


Figure 1.3: Visualised comparison of frequentist and Bayesian results, from [Wolfgang](#_bookmark239) [(2016)](#_bookmark239)

A significant problem of the frequentist approach is that its interpretation of result often encourage people to fi on numbers such as p-value and ignores uncertainty. Study such as [Open Science Collaboration and others](#_bookmark216) [(2015)](#_bookmark216) have shown that psychology experiments with frequentist results often have weaker evidence than evidence suggested by p-value from studies, and a large proportion of the studies are not repeatable. A p-value of 0.51 and a p-value 0.49 are very similar probability-wise; however, people could deduce significance and non-signifi just because 0.049 is lower than the significance level, and 0.051 is not if they fi on the significance level of 0,05. Results from Bayesian analyses are less prone to such issue, as it results are in posterior probability distributions, even though a bit harder to interpret, it contains much more information within the posterior probability distribution compares to the frequentist p-values. Secondly, Bayesian inference allows the inclusion of external information from sources other than just the observation; this is a form of evidential probability and allows for sequential learning ([B˚a˚ath,](#_bookmark169) [2015;](#_bookmark169) [Jia,](#_bookmark198) [2018).](#_bookmark198) For time-series data, the way Bayesian includes information from both the past and present make much more sense than considering only the present observation

because datasets with sequential measurements such as time-series data often observe autocorrelation, where information from the past strongly associates with present observations. Also, as mentioned in the previous section, incorporating external information about hospital capacity makes much more sense than the frequentist approach as the primary determining factor for hospital congestion are the capacity it holds. We could observe a very improbable arrival, but if this number still falls under the capacity, there is no congestion. Lastly, Bayesian has an advantage over frequentist for large datasets with complex structures. As models get more complex and more nested, the likelihood for over-fitting, over-parametrisation and multiplicity often increase for simple frequentist models. Bayesian allows utilisations of fi defi for highly custom models that also penalises over-parametrisation ([B˚a˚ath,](#_bookmark169) [2015;](#_bookmark169) [Bolker et al.,](#_bookmark168) [2009;](#_bookmark168) [Kruschke and Vanpaemel,](#_bookmark203) [2015).](#_bookmark203)

* + 1. **Prior Selection**

The essence of Bayesian inference lies within the Bayes theorem, that is, given prior and likelihood, we can deduce the posterior. A prior probability distribution refers to a probability distribution that expresses ones’ prior beliefs about an observation. For example, the prior could be the probability distribution representing the expected patient arrivals at an Emergency Department (ED). There are several acceptable methods of obtaining the prior. Firstly, prior can be determined from past information, such as historical hospital records. Secondly a prior can be a purely subjective opinion from an expert in the fi an ED staff making claims about average patient arrivals also seem reasonable, and lastly statistical principles can be applied to synthesis a prior, a normal distribution is often used for large samples because The central limit theorem states that as the sample size gets larger the distribution of sample means will approximate to a normal distribution. Also, people often choose to use a conjugate prior to simplifying the calculations of the posterior distribution.

Priors can be classified by the amount of information it contains, in terms of informative, weakly informative and non-informative priors. Informative priors defi specific information about a variable. Historical information about patient arrivals can be modelled to give predictions, and optimum solutions can be used as the prior. A weakly informative prior contains partial information about a variable, parameters are more vague and rounded but still regularises the parameter estimates reasonably well. A possible standard to distinguish informative and weakly informative prior is that for informative priors, people with diff nt backgrounds are likely to have very diff t idea about the prior, and for weakly-informative prior, people from diff t background should agree with the prior. Non-informative priors in another hand, contains no information, such as the use of a uniform or fl line distribution

that contain no variation. Informative priors are the proper way to introduce the available informations to the model, and will provide solutions to computational issues and improve efficiency if it was correctly assigned, non informative prior in other hand challenges the notion of prior, but are often thought of as the reference models,and serve as starting point of more informative prior distributions. The weakly informative prior stands as a compromise between two sides [(Carlin and Louis,](#_bookmark171) [2008;](#_bookmark171) [Golchi,](#_bookmark188) [2016).](#_bookmark188)

Parameters of prior distributions are also the hyper-prior because it gives rise to the parameter being estimated. For example, if parameter *λ* of a Poisson distribution is used to model the parameter *µ* of a normal distribution, it is then said that *λ* is a hyper-parameter. Hyper-parameters themselves may have hyper-hyper-prior distributions expressing beliefs about their values. Moreover, a Bayesian model with more one level of priors is called a hierarchical Bayesian model.

* + 1. **Hierarchical Bayesian models**

Information about the structure of hierarchical data can be processed and utilised with the use of Hierarchical Bayesian Models (HBM). Parameters within the HBMs contain a nested hierarchical structure, which meant that parameters of a higher hierarchy control the parameters of the lower hierarchy. The left plot in fi [1.4](#_bookmark22) described the relationship between parameters and observations of an Independent Bayesian Model (IBM), for the case of IBM, observations are directly generated from parameters from a single level probability distributions and thus considered independent. The right plot in fi [1.4](#_bookmark22) described the relationship between parameters and observations of an HBM, for the case of HBM, population parameters and observations control individual parameters is generated from a multi-level probability distributions [(Gelman and](#_bookmark186) [Hill,](#_bookmark186) [2007;](#_bookmark186) [Wolfgang,](#_bookmark239) [2016](#_bookmark239)), therefore we have this pooling of information, and there is no independence.

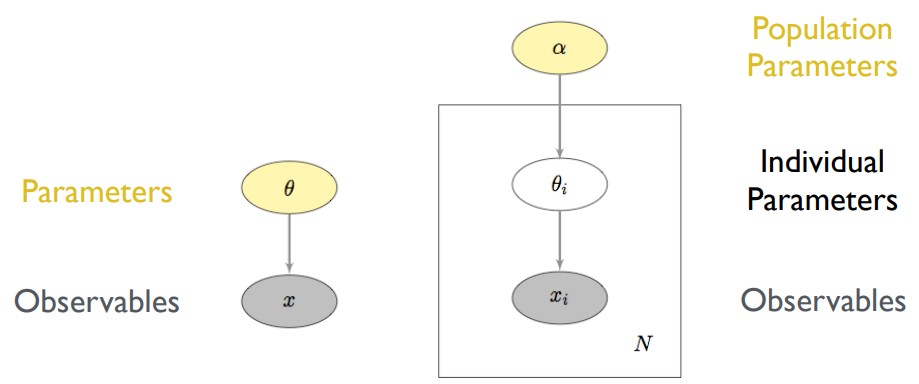


Figure 1.4: Modelling structure between regular and Hierarchical Bayesian Models, from [Wolfgang](#_bookmark239) [(2016)](#_bookmark239)

In Bayesian terms, a simple IBM follows equation [1.9,](#_bookmark23) considering only the individual parameter *θ* and observations *x*. For an HBM extra population parameter, *α* can be incorporated and written with the formula shown in equation [1.10.](#_bookmark24)

Posterior

*p*,.(""*θ*'-*x*"") *∝*

*|*

Likelihood

,. ""'- ""

*|*

*p*(*x θ*)

Prior

,. ""'- ""

*|*

*p*(*θ α*)

(1.9)

Posterior

Likelihood

Prior

,.*p*(*α*""*,*'-*θ|x*"") *∝* ,.*p*(*x|θ, α*""'-)*p*(*θ|α*"") *p*,.""(*α*'-"") (1.10)

The way HBMs integrate group information with a pooling process is a partial- pooling model that compromise between complete-pooling and no-pooling models. With complete pooling, each unit is assumed to have the same parameter, leaving no independence and therefore contain zero population variance. With no pooling, each unit is assumed to have diff t and independent parameters, and this complete independence corresponds to infi population variance. Both complete-pooling and no-pooling models have significant applications in preliminary statistical analyses but unlikely to accommodate real-life data very well; as it is unlikely and does not make sense for scenarios of no independence and complete independence to happen in real-life. For HBMs, individual estimates tend to shrink toward the population mean, and this effectively lowers the overall Root Mean Square Error (RMSE). RMSE is a popular goodness-of-fi value used to measure the diff between real and predicted values during cross-validation analyses. Therefore, there is no surprise that studies have shown that hierarchical models tend to give a better prediction compare to no-pooling and complete-pooling models in all levels [(Gelman,](#_bookmark184) [2006;](#_bookmark184) [Gelman and](#_bookmark186) [Hill,](#_bookmark186) [2007;](#_bookmark186) [Park et al.,](#_bookmark218) [2004).](#_bookmark218)

* + 1. **DIC**

In Bayesian statistics, deviance is often used to evaluates the goodness-of-fi statistic for a model. [Spiegelhalter et al.](#_bookmark232) [(2002)](#_bookmark232) has proposed the idea of using deviance information criterion (DIC) as a tool for goodness-of-fi comparison between Bayesian

models. DIC is defi as the expected deviance *D*¯ parameters in the model *pD*.

plus the effective number of

which is also equivalent to

*DIC* = *D*¯ + *Pd* (1.11)

*DIC* = *D*(*θ*¯) + 2*Pd* (1.12)

DIC can be thought of as an approximation to the deviance, or the out-of-sample predictive error, therefore models with lower DIC are preferable. The smaller the

DIC, the more accurate the model is for predictions. Model DICs can be generated by using dic.samples function from the rjags package within the R environment, The defi of *pD* used by dic.samples function is the one proposed by Plummer (2002).

* + 1. **Effective Sample Size**

The idea effective sample size (ESS) is to provide a unit that converts dependent samples into units of independent samples. In other words, the ESS is an estimate of the sample size required to achieve the same level of precision if that sample was a simple random sample. For the case of highly correlated samples, 1,000 samples from a Markov chain can be only equivalent to 100 independent samples. In the case of a weakly correlated sample, 1000 samples from a Markov chain can be only equivalent to 300 independent samples (C[ook,](#_bookmark177) [2017;](#_bookmark177) [Kass et al.,](#_bookmark201) [1998).](#_bookmark201) The equation for ESS is:

ESS =

1 + 2

*n*

*k*=1 *ρ*(*k*)

*∞*

(1.13)

Where n is the number of samples and *ρ*(*k*) is the correlation at lag *k*.The MCMC process causes the draws to be correlated, as shown in the Markov chain algorithm (equation [1.8)](#_bookmark17) the sample from time *t* + 1 depends on the sample from time *t*. This means that the effective sample size is generally lower than the number of draws and is used for calculations. Any Markov chain between 0 ESS and completely ESS is reasonable.

* 1. **Hierarchical time-series**
     1. **Structure of Hierarchical time-series**

A time series refers to a series of data points that are ordered in time and or contain information about when the numbers were recorded [(Das,](#_bookmark179) [1994;](#_bookmark179) [Hyndman and](#_bookmark196) [Athanasopoulos,](#_bookmark196) [2018).](#_bookmark196) Time series analysis is an intriguing part of epidemiology and clinical research because time often has a strong association with the development of diseases and often reflect essential population-level information. Time series in epidemiology and clinical research often can have a hierarchical structure, where many time-series aggregate together to form broader categories of time series by specific population attributes. Finer International Statistical Classification of Diseases and Related Health Problems (ICD)(O[rganization,](#_bookmark217) [2019)](#_bookmark217) categories at a lower level of hierarchy diagnoses can be aggregated into broader categories at a higher hierarchy level.

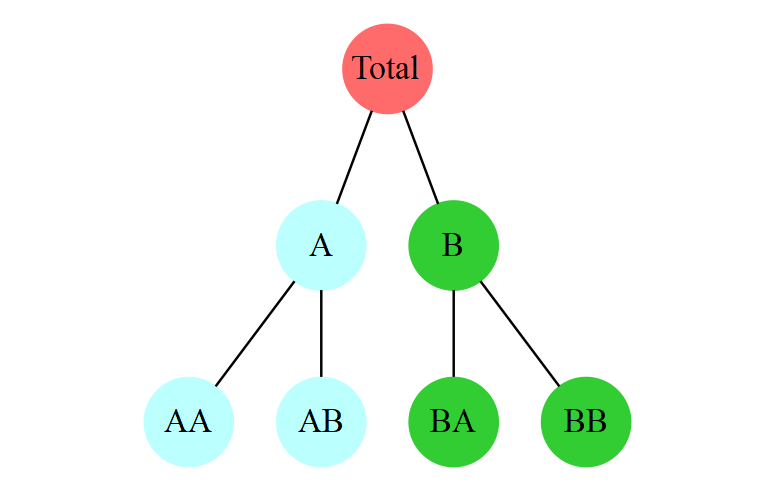


Figure 1.5: A 2 level hierarchical structure

A hierarchical time series (some time called grouped time series) can be represented with a hierarchical tree diagram. Figure [1.5](#_bookmark29) shows a simple 2 level hierarchical structure. At the very top of the hierarchy, it is the total; it can be calculated by summing of all categories, it represents the sampling population, and can be thought of as the start (level 0) of the hierarchy. The total can be disaggregated into subcategories at level 1 of the hierarchy, and subcategories can be disaggregated into sub-subcategories at level 2 of the hierarchy, and so on. Categories a hierarchy higher can be thought as the parent, and categories a hierarchy lower can be thought as the children of another hierarchy. The observations of the total series given time *t* is denoted by *ytotal,t* for *t* = 1*,* 2*,* 3*. . . , n* with *, t* standing for given time *t* and *n* standing for the length of the time period. Additional subscripts should be added for the subsequent levels, for example, *yA,t* denotes the *t*th observation of series A at level 1 of the hierarchy, *yAA,t* denotes the *t*th observation of series AA at level 2 of the hierarchy, and so on.

Hierarchical relationship descripted in fi [1.5](#_bookmark29) can be presented with following equations:

*ytotal,t* = *yA,t* + *yB,t* (1.14)

*yA,t* = *yAA,t* + *yAB,t yB,t* = *yBA,t* + *yBB,t* (1.15) The relation can then be generalised into: ,

*ni nij*

*ytotal,t* = '\" *yi,t yi,t* = '\" *yij,t ...*

*i*=1

*⇒ yparent,t* = '\" *ychildren,t*

*i*=1

(1.16)

Moreover, this can be thought of as aggregation constraints where observed values

from bottom level hierarchies must sum to the observed value of the parent of these values.

Hierarchical time series can also be translated into a *n × m* matrix *S*, when *n* equals to the number of observation of each time series, and when m equal to the total number of categories when considering categories from all levels [(Hyndman,](#_bookmark194) [2015;](#_bookmark194) [Wickramasuriya et al.,](#_bookmark238) [2019).](#_bookmark238) For the hierarchical structure in Figure [1.5,](#_bookmark29) we can write with a matrix formula:

*Yt S bt*

*yT otal,t* 1 1

|  |  |
| --- | --- |
| 1 | 1 |
| 1 | 0 |
| 0 | 1 |
| 0 | 0 |

 *yA,t*  1

 *yB,t*  0







  

0 





1

 

*yAA,t* 



(1.17)

  = 

 *yAB,t* 

 *yAA,t*  1

0  

  

 *yBA,t* 

 *yAB,t* 

 *yBA,t* 

0 1 0 0

0 0 1 0

*yBB,t*

 

*yBB,t*

 

0 0 0 1

* + 1. **Forecasting approaches**

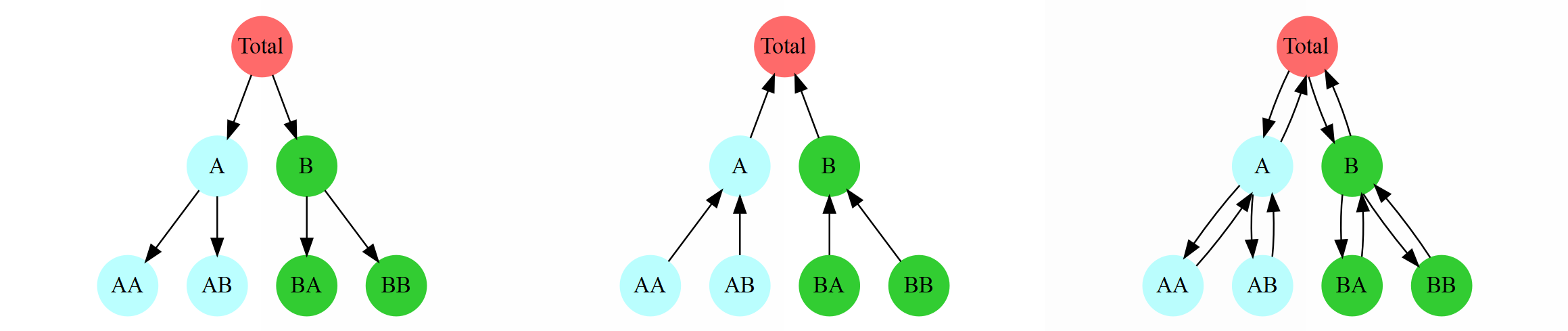
A classical method for hierarchical forecasting is the **bottom-up** approach. In this approach, independent forecasts of each time-series series at the leaf(very bottom) level of the hierarchy are aggregate to workout the forecasts for the upper levels of the hierarchy until reaching the total. The main advantage of the approach is that no information is lost during the whole process; however, modelling and predictions of lower-level time-series often very challenging due to an abundant presence of noise [(Hyndman et al.,](#_bookmark197) [2014).](#_bookmark197) A level one level on top another level can be considered as the parent as it is where the lower level come from, vice versa a level one level below another level can be considered the children. We can think of the hierarchical relationship, as shown in equation [1.18,](#_bookmark31) where the estimated values from the parent levels are the aggregations of estimates from children levels.

'\" *ychildren,t ⇒ yparent,t* (1.18)

Another method for hierarchical forecasting is the **top-down** approach. In this approach, the forecast of each time-series series at the top level(total) is disaggregated down the hierarchy, with information of the hierarchical structure from historical data. This approach has reduced noise and is easier to model, but some information is lost during the process [(Hyndman et al.,](#_bookmark197) [2014).](#_bookmark197) We can think of the relationship as shown in equation [1.19,](#_bookmark32) where the estimated values from the children level are comes

from the dis-aggregations of estimates from the parent levels, given an estimation of a hierarchical structure *H*.

*yparent,t|H ⇒* '\" *ychildren,t* (1.19)



* + - 1. Top down (b) Bottom up (c) Optimum reconciliation

Figure 1.6: Top-down, Bottom up and Optimum reconciliation approach

[Hyndman et al.](#_bookmark195) [(2011)](#_bookmark195) recommend a new method of **optimal reconciliation** for the forecast of hierarchical time series data for its feature and performance. The optimal reconciliation works by independently forecasting each time-series at all levels of the hierarchy, and then reconcile each forecast with regression models by minimising the forecasting error (Figure [1.6).](#_bookmark33) The most significant advantage the optimal reconciliation approach is that all of the information available within each level of the hierarchical structure are in consideration during the forecasting process. The idea is that important information may lye within any particular level within the whole hierarchical structure, and can be hard to detect from other levels if we only consider one level at the time [(Hyndman and Athanasopoulos,](#_bookmark196) [2018).](#_bookmark196) A weakness of the optimal reconciliation is that even if the observations considered are natural numbers; it could still produce negative predictions that are the optimum solution to the prediction but does not make sense in real-life. In many instances, time-series predictions need a set of non-negative reconciled forecasts. If the demand for a particular commodity for the time-period *t* = 2020 is *−*5, it does not have any sense at all.

Time-series forecast with all approached mentioned above can be made by using the forecast() function with an hts object (created using ts and hts function)in under the hts package ([Hyndman et al.,](#_bookmark193) [2018)](#_bookmark193) in the R environment. Optimal non-negative reconciled forecasts are currently not available for the hts package within the R environment, but a little bird told me hts package with non-negative reconciled forecasts will be updated at the end of July 2019. For another note, optimum reconciliation forecasts used for the next session does have negative values,

but the values are mostly extremely small so they are treated as 0 values.

* 1. **Outline of Thesis**

The organisation of the thesis is described in the follows:

* + - **Chapter** [**1**](#_bookmark0)starts with the introduction of examples of where these issues in hierarchical structure arise; these include background and concepts of the top- ics, including emergency department congestion and anomaly detection. Technical background about statistical methods and techniques necessary for our methods, these include topics of and within Bayesian inference and hierarchical time-series predictions. Lastly, we included outlines for the thesis and notations and conventions for this thesis.
    - **Chapter** [**2**](#_bookmark36)presents methods, results and discussion of our simulation studies of anomaly detection in several settings. The fi simulation looks on how diff t priors would affect the goodness of fi and posterior results of Independent Bayesian Models(IBM). The second simulation compares scenarios with diff t increments of anomalies added. The third simulation compares scenarios with diff t diff t brunching at Category A, with addition of a constant anomaly.
    - **Chapter** [**3**](#_bookmark96)presents methods, results and discussion of our studies of anomaly detection with MIMIC-III, a publicly-available database with health-related data. More specifically, for four scenarios which include original observations and observa- tions with anomalies added on categories of common, rare and extremely rare disease diagnoses.
    - **Chapter** [**4**](#_bookmark134)provides a fi l conclusion and overview of studies presented in this thesis.
* **The Appendix** includes:

**–** Descriptions of custom R function created, and R code used to implement diff t prior models, Independent Bayesian Model and Hierarchical Bayesian Model, used in Chapter [2](#_bookmark36) and Chapter [3.](#_bookmark96)

* 1. **Notations and conventions**

The following notation and conventions are used throughout this thesis to aid readability:

**Software** Software is denoted by mono-space font with boldface.

For example: This plot is created using R codes.

**R packages** Package names are given in mono-space font .

For example: This function is from the GOpro package.

**R functions** Functions are denoted with mono-space font with two trailing parentheses. Arguments are similarly displayed in monospace with format ”value” or ”argument = value”.

For example: This plot is created using ggplot() function.

For example: We used dnorm(1, sd = 0.1) as our hyper-prior

**Quotes** Direct quotes should be denoted using Italic font

For example: ”*Shall I compare thee to a summer’s day* ”

**Textbooks** Text book and journal titles should be denoted using Italic font

For example: ”*High Performance Computing for Disease Surveillance*”

**Key words** Some key words are denoted in bold font For example: We have **significant** results.

**Chapter 2 Simulations**

* 1. **Introduction**

In this chapter, we carried-out simulations with synthetic patient arrival datasets and compares the posterior distribution results and prediction rate between Independent Bayes Model (IBM) and Hierarchical Bayes Model (HBM) in scenarios of with diff t anomalies and diff t hierarchy structures. Sections 2.1 details functions used to synthesis simulation datasets. Sections 2.2 evaluate the use of diff t priors for Bayesian Models to used during the simulations. Section 2.3 explores the impact of the size of the anomaly on anomaly detection at diff t levels of hierarchy. Lastly, section 2.4 considers how the complexity of the hierarchical structure affects anomaly detection at different levels of hierarchy.

* 1. **Simulation Methods**

All datasets used for the simulation studies are synthetically generated. Synthetic datasets refer to data information artifi manufactured rather than generated from real-world events. Synthetic data takes preference over the alternative option, perturbed data, a dataset generated by adding alteration and noise to real-world data [(Drechsler et al.,](#_bookmark181) [2008;](#_bookmark181) [Drechsler,](#_bookmark180) [2011).](#_bookmark180) Advantages of synthetic data include (1) Disclosure protection, making sure no sensitive information about individuals from the public can not be leaked or extracted. (2) Reduce the effects of possible background noises presented in real-world data. (3) Allows control over the hierarchical structure of datasets, note that real-world hierarchical data are often extremely complex and can have lots of levels, and complexity that comes with it, increasing the diffi y to control our simulations, and making interpretations diffi

Several custom R-functions has been created to synthesis datasets used in our

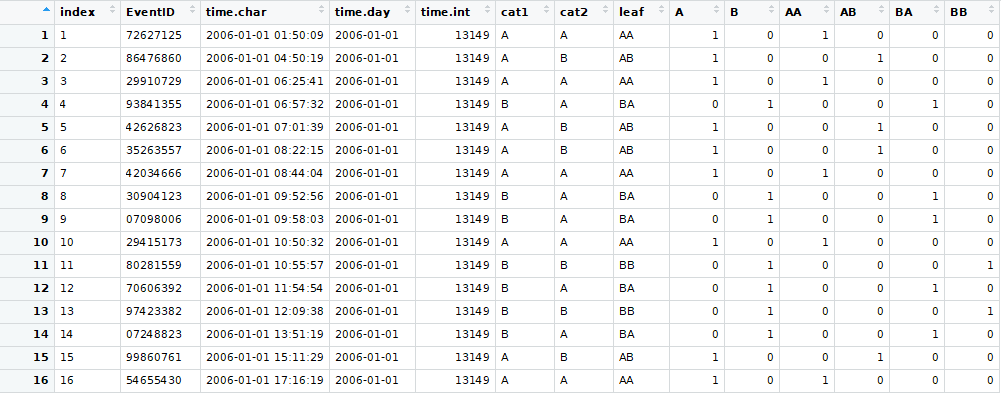


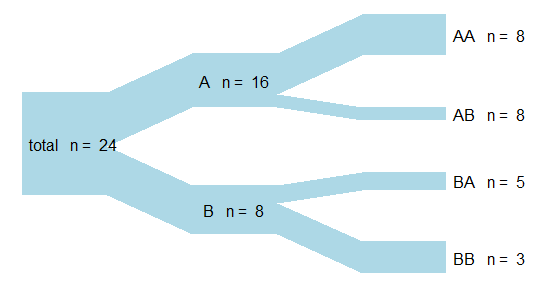
Figure 2.1: Entries of the raw dataset

studies, Appendix [A.1](#_bookmark142) provides descriptions of the custom functions. There are several reasons for creating custom simulation synthesis functions. Firstly, a large number of simulated datasets with variable settings were expected to be required to perform our analyses; this required a large number of custom function arguments that existing packages may not contain. Secondly, synthetic datasets were simulated to mimic a health dataset in a purely theoretical setting. Thirdly, an emphasis was placed on the hierarchical nature of particular variable of the dataset (for example, ICD codes and geographic location), custom structure hierarchical structures can be used to generate brunching, and nodes during the simulation process. Lastly, for exploration research, it was unclear what functionalities were needed, existing statistical data simulation packages may become insuffi t at a later stage.

simdata and adddaily.anomaly are the two major function used during the data synthesis process. simdata function were used to synthesise a raw dataset (figure [2.1)](#_bookmark39) that simulate a hospital record that you would typically found in hospital digital archives, where each row corresponds to a single entry of hospital event with a unique hospital event identifier. The function automatically generates Time, dates and various information about the hierarchical structure. The default number of simulation is 1,000,000, and default period is a 12 year-long period between 1 January 2006 and 31 December 2018, and a matrix that contains information of the hierarchical structure are manipulated and used to produce a hierarchical structure with different characteristics.

The matrix used in the simdata function contains the theoretical value of each leaf of a two-level hierarchical structure in a proportion out of 1000. As shown with the example in fi [2.2,](#_bookmark40) if the theoretical value at of the leaves (refers to categories at the very bottom level hierarchy) are 8, 8, 5, and 3, you would represent

8 5



*⇒* 8 3

333 208

*⇒* 333 125

Figure 2.2: Conversion of a simple hierarchical structure to a matrix

this information with a matrix with numbers 333,333,208 and 125. Each column corresponds to a single group; therefore, column sums of the matrix correspond to the count of a level 1 categories, and the sum of all numbers in the matrix equal to the total count. The parent levels are simply the sum of children levels, and the complete hierarchical structure can be generated automatically just from numerical information of leaves. Note that the number of column of the matrix corresponds to the maximum number of leaves within a level 1 category, for level 1 categories with less than maximum amount of leaves, 0 is used to represent no leaf. The default value of the matrix is 250, 250, 250 and 250, which represent a two-level hierarchical of equal proportions, with 2 level-1 categories, each with 2 level-1 categories of equal proportions. This way of storing hierarchical information is simple and efficient but is not suitable for hierarchical models that contain 3 or more levels, and maybe a little confusing for other people at fi future improvements of the function could be allowing usage of a user-generated hierarchical time series (hts) object from the hts package instead of the matrix, to allow generations of hierarchical dataset with 3 or more levels.

adddaily.anomaly were used to addanomaly function to add anomaly to daily count data, a hierarchical time-series dataset generated by tabulating the raw dataset using tabulatedata function. The default setting is a point anomaly of 1 day, on 2015-11-15. The amount of anomaly added are in terms of the proportion of the total count, and amount of anomaly assigned to each leaf can be varied.

For point anomalies the amount of anomaly assigned to each leaf is transcribed with binding of 3 vector: (1) a vector of amount of anomaly to be added to total,

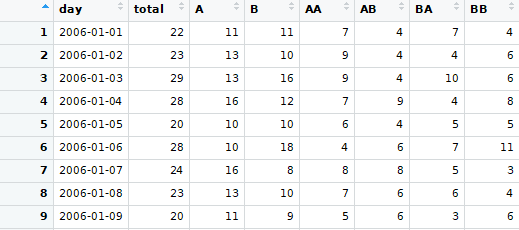


Figure 2.3: Entries of the daily count dataset

(2) a vector of amount of anomaly to be added to level 1 categories and (3) a vector of amount of anomaly to be added to level 2 categories, all in values of proportions. For example, for a hierarchical time series with the same hierarchical structure as the example given in fi [2.2,](#_bookmark40) if we want to add all of our anomalies to AA, the proportion vector of anomaly for total is *{*1*}* and this will be the case for any hierarchical structure because the proportion of all categories will sum to 1. The proportion vector of anomaly for level 1 categories is *{*1*,* 0*}* and the percentage vector for level 2 categories is *{*1*,* 0*,* 0*,* 0*}*, so the proportion vector that is required for our function to indicate our setup, is *{*1*,* 1*,* 0*,* 1*,* 0*,* 0*,* 0*}*.

0

|  |  |  |  |
| --- | --- | --- | --- |
| *Ptotal* | *Plv*1 | *Plv*2 *P* |  |
| 1 + | 1 0 + 1 | 0 0 0 *⇒* 1 1 0 1 0 | 0 |

Figure 2.4: Construction of the proportion vector

Details about setups used for the data synthesis process can be found in the start of each section of our simulation tests, and the R codes used can be found in 91-createdata-test.R, 92-createdata-anoamly.R, 93-createdata-proportion.R and 94-createdata-count.R found in GitHub Repository: [https://github.com/jungxue/](https://github.com/jungxue/research-masters-Jung) [research-masters-Jung.](https://github.com/jungxue/research-masters-Jung)

* 1. **Simulation 1: Prior**
     1. **Priors**

For Bayesian inference to make sense, a prior is essential and often required. However, there are diff t ways in which prior can be generated. For simulations, no external information is available; therefore, the informative prior approach is out of the question. However, it is possible for us to assign relatively simple weak-informative and non-informative priors by generating probability distributions with vaguely defi parameters. Hence, we proposed six diff t priors that express diff t hypothesised information about the prior and test how they affect our posterior distribution calculations.

A simple Poisson process can be used to model for a random, and mutually independent arrivals rate, with the likelihood model:

*y,t ∼ Poisson*(*µ,t*) *log*(*µ,t*) = *log*(*ρ,t*)

For the Poisson process, observations *yi,t* (*y* of *i*th category given time *t*) is hypothesised to follow the distribution of the Poisson process, with parameter *µi,t* stand the expected number of arrivals per time-interval. For our simulations, this time interval is set up to be a 12 year-long period between 1 January 2006 and 31 December 2018. Distributions of priors can then assigned to *µi, t*, this provides essential information to the parameter and will have a significant impact in our calculations. Samples of the JAG codes used to build our models is presented in appendix [A.2,](#_bookmark151) and complete JAGS codes and R codes used to obtain our graphical results are available on GitHub repository: [https://github.com/jungxue/research-](https://github.com/jungxue/research-masters-Jung) [masters-Jung.](https://github.com/jungxue/research-masters-Jung)

### Model 1 Non-informative Prior

Model 1 is an Independent Bayesian Model with no priors that models each categories independently on all hierarchical levels. This model provides no information for the estimation of distribution parameter *µi,t*, and therefore the prior can be defi as a non-informative prior (although this term often refers to fl prior), here in this thesis we call it the null prior. The likelihood model for Poisson model with null prior is:

*yi,t ∼ Poisson*(*µi,t*) *log*(*µi,t*) = *log*(*ρi,t*)

### Model 2 Normal(1,0.3) Prior

Model 2 is an Independent Bayesian Model with *ρ ∼ Normal*(1*,* 0*.*3) prior that models each categories independently on all hierarchical levels. This model does contain information for the estimation of distribution parameter *µi,t*, and therefore the prior can be defi as a weak-informative prior. *I*[*λ>*0] indicates truncation of distribution at 0; in other words, the prior does not contain*λi,t* with values less than 0 and should not contain any negative values. The likelihood model for a Poisson process with a *Normal*(1*,* 0*.*3) prior is:

*yi,t ∼ Poisson*(*µi,t*) *log*(*µi,t*) = *log*(*ρi,tλi,t*)

Priors for model parameters:

*λi,t ∼ Normal*(*µi,t, σi,t*)*I*[*λ>*0]

Hyper-priors for model parameters:

*µi,t ∼ Normal*(1*,* 0*.*1)

*i,t ∼ Normal*(0*.*3*,* 0*.*1)

*σ*2

### Model 3 Normal(1,0.1) Prior

Model 3 is an Independent Bayesian Model with *ρ ∼ Normal*(1*,* 0*.*1) priors that models each categories independently on all hierarchical levels. This model does contain information for the estimation of distribution parameter *µi,t*, and therefore the prior can be defi as a weak-informative prior. *I*[*λ>*0] indicates truncation of distribution at 0; in other words, the prior does not contain*λi,t* with values less than 0 and should not contain any negative values. The likelihood for a Poisson process with a *Normal*(1*,* 0*.*1) prior model is:

*yi,t ∼ Poisson*(*µi,t*) *log*(*µi,t*) = *log*(*ρi,tλi,t*)

Priors for model parameters:

*λi,t ∼ Normal*(*µi,t, σi,t*)*I*[*λ>*0]

Hyper-priors for model parameters:

*µi,t ∼ Normal*(1*,* 0*.*1)

*i,t ∼ Normal*(0*.*1*,* 0*.*1)

*σ*2

### Model 4 Gamma(4,3) Prior

Model 4 is an Independent Bayesian Model with *ρ ∼ Gamma*(4*,* 3) priors that models each categories independently on all hierarchical levels. This model does contain information for the estimation of distribution parameter *µi,t*, and therefore the prior can be defi as a weak-informative prior. *I*[*λ>*0] indicates truncation of distribution at 0; in other words, the prior does not contain*λi,t* with values less than 0 and should not contain any negative values. The likelihood model for Poisson process with *Gamma*(4*,* 3) prior is:

*yi,t ∼ Poisson*(*µi,t*) *log*(*µi,t*) = *log*(*ρi,tλi,t*)

Priors for model parameters:

*λi,t ∼ Gamma*(*αi,t, βi,t*)*I*[*λ>*0]

Hyper-priors for model parameters:

*αi,t ∼ Normal*(4*,* 0*.*1)

*βi,t ∼ Normal*(3*,* 0*.*1)

### Model 5 Laplace(1,1) Prior

Model 5 is an Independent Bayesian Model with *ρ ∼ Laplace*(1*,* 1) priors that models each categories independently on all hierarchical levels. This model does contain information for the estimation of distribution parameter *µi,t*, and therefore the prior can be defi as a weak-informative prior. *I*[*λ>*0] indicates truncation of distribution at 0; in other words, the prior does not contain*λi,t* with values less than 0 and should not contain any negative values. The likelihood model for Poisson process with *Laplace*(1*,* 1) prior is:

*yi,t ∼ Poisson*(*µi,t*) *log*(*µi,t*) = *log*(*ρi,tλi,t*)

Priors for model parameters:

*λi,t ∼ Laplace*(*αi,t, βi,t*)*T*[*λ>*0]

Hyper-priors for model parameters:

*αi,t ∼ Normal*(4*,* 0*.*1)

*βi,t ∼ Normal*(3*,* 0*.*1)

### Model 6 Mixture Prior

Lastly, Model 6 is an Independent Bayesian Model with a mixture prior that models each category independently on all hierarchical levels. The prior is thought to be a mixture of no information and weak information, in other words, we are proposing that for most of the time there is no variation, and for some of the time, there is information. The idea comes from [Berry and Berry](#_bookmark166) [(2004),](#_bookmark166) where they proposed a scenario where the majority of the difference between treatment and control is zero, and there are just some variations. However note that their prior is an one- sided distribution that centre at 0, and our prior is two-sided and centre at 1. This model does contain information for the estimation of distribution parameter *µi,t*, and therefore, the prior can be defi as a weak-informative prior. *I*[*λ>*0] indicates truncation of distribution at 0; in other words, the prior does not contain*λi,t* with values less than 0 and should not contain any negative values. The likelihood model for a Poisson model process with mixture prior is:

*yi,t ∼ Poisson*(*µi,t*) *log*(*µi,t*) = *log*(*ρi,tλi,t*)

Priors for model parameters:

*λi,t ∼ Normal*(*spikei,t* + (1 *− spikei,t*)*slabi,t,* 0*.*1)*T*[*λ>*0] Hyper-priors for model parameters:

*spikei,t ∼ Binomial*(0*.*9*,* 1)

*slabi,t ∼ Normal*(*µi,t, σi,t*) Hyper-hyper-priors for model parameters:

*µi,t ∼ Normal*(1*,* 0*.*1)

*i,t ∼ Normal*(0*.*1*,* 0*.*1)

*σ*2

10.0

model 1: Null prior

0.8

model 2: Normal(1,0.3) prior

7.5

density

0.6

5.0

density

0.4

2.5

0.2

0.0

1.25

1.00

0 1 2 3 4 5

y1

model 3: Normal(1,0.1) prior

0.0

0.6

0 1 2 3 4 5

y2

model 4: Gamma(4,3) prior

0.75

density

0.4

density

0.50

0.25

0.2

0.00

0 1 2 3 4 5

y3

0.0

0 1 2 3 4 5

y4

model 5: Laplace(1,1) prior

model 6: 0.5 Mixture prior

3

0.4

density

density

2

0.2

1

0.0

0 1 2 3 4 5

y5

0

0 1 2 3 4 5

y6

Figure 2.5: Histo-density plot of the prior probability distribution of proposed prior density models. The red dashed line indicates the center point of these distributions, which is 1.

Figure [2.5](#_bookmark45) show the distribution of *λi,t* in histograms. Because we assign a prior distribution to our *µi,t* by directly multiplying by *λi,t*, the distribution of *λi,t* is thought as the equivalent to the distribution of our prior. As presented in histograms, our priors have diff t distributions that suggest a diff t belief. Model 1 used a non-informative fl prior, that contains no information, and hence no distribution is observed, and only have 1 bar at 0. Model 2 used a *Normal*(1*,* 0*.*3) prior that seems reasonable; however note that, because its left tail was truncated at 0, the distribution has lost its symmetrical property. Model 3 used a *Normal*(1*,* 0*.*1) that has a smaller variance, hence the distribution is a little bit taller, indicating more observation tend

to be close to the mean. Model 4 used a *Gamma*(4*,* 3) prior; it skews towards the right-hand side. The distribution translates into that we are observing more positive anomalies compare to negative anomalies. Model 5 used a *Laplace*(4*,* 3) prior, it also skews to the right, but with a heavier tail and a much higher distribution at extreme positive values, this translates into information at we are consistently observing a large anomaly. Lastly, we have model 6 that used a mixture prior; there is a very strong spike at 1 and some variation around it.

* + 1. **Simulation setups**

In section 2.3.2, we present the results of the posterior distribution calculations for the synthesised data, with anomalies added on category AA on level 2 (leaf) of the hierarchy. The default setting for the data synthesis process is applied, with the addition of 25% anomaly. An important note, the 25% anomaly here meant an addition of 25% of the count of category total on top of the hierarchy, not addition of 25% on category AA.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **2.3.3** | **Results** |  | | | | | |
|  | Comparisons | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|  | M1:M2 | -2.126 | -1.574 | -0.599 | -0.868 | -0.158 | 0.113 |
|  | M1:M3 | -4.456 | -3.295 | -2.069 | -2.032 | -0.650 | 0.189 |
|  | M1:M4 | -2.992 | -1.729 | -1.174 | -1.303 | -0.800 | 0.106 |
|  | M1:M5 | -2.645 | -1.745 | -0.764 | -1.172 | -0.590 | -0.124 |
|  | M1:M6 | -3.383 | -2.200 | -1.645 | -1.683 | -1.047 | -0.258 |
|  | M2:M3 | -4.555 | -2.288 | -1.884 | -1.164 | 0.713 | 1.440 |
|  | M2:M4 | -2.393 | -1.009 | -0.552 | -0.434 | -0.155 | 2.232 |
|  | M2:M5 | -2.382 | -0.494 | -0.071 | -0.304 | 0.267 | 0.782 |
|  | M2:M6 | -3.331 | -1.571 | -0.626 | -0.815 | -0.122 | 1.639 |
|  | M3:M4 | -2.450 | -0.285 | 0.895 | 0.730 | 1.787 | 3.658 |
|  | M3:M5 | -1.020 | -0.578 | -0.168 | 0.860 | 2.345 | 3.677 |
|  | M3:M6 | -1.995 | -1.322 | -0.412 | 0.349 | 1.649 | 4.198 |
|  | M4:M5 | -1.884 | -0.905 | 0.501 | 0.131 | 0.913 | 2.282 |
|  | M4:M6 | -2.858 | -1.537 | -0.829 | -0.380 | 1.029 | 2.039 |
|  | M5:M6 | -3.259 | -0.929 | -0.672 | -0.511 | 0.378 | 1.454 |

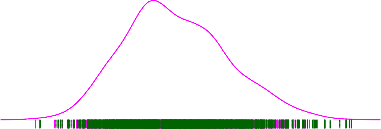
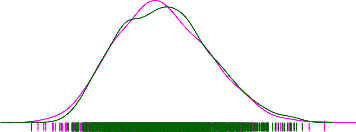
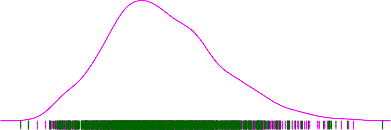
Table 2.1: DIC (with 1,000,000 iteration) summary distribution cross comparisons for model 1 to model 6

Table [2.1](#_bookmark48) gives the comparison of of DIC distributions. A comparison of M1: M2 indicated comparisons between model 1 and model 2, with DIC distribution of model 2 taking away the DIC distribution of model 1. An important note to take before the interpretation of results in this table is the weak signals (a negative value signals a model is better than another) of the DIC estimations. The initial DIC results with 1000 iterations give weak signals, showing signs of slow convergence, this meant that the DIC values for each model diff by a large margin each time DIC ran for diff t jags model, sometimes we get negative mean and median values, and sometimes we get positive mean and median values, making it tough to decide which model is the best. Weak signals are still present when the number of iteration increase to 1,000,000. With repeated trials, Model 2 to model 6 consistently gives a lower DIC value compares to model 1, showing strong signals at 3rd Quartile and Max value. The observation provides strong evidence that Model 1 with non-informative prior gives the highest and worse DIC estimations out of all models. The DIC comparisons for model 2 to model 6 have weak signals on mean and median values, even with high iterations, and it is almost impossible to tell the diff this suggests that there is no or close to no diff between the DIC values for model 2 to model 6.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | sd | 2.5% | 50% | 97.5% | Rhat | n.eff |
| model 1 | 21.0615 | 0.0000 | 21.0615 | 21.0615 | 21.0615 |  | 0 |
| model 2 | 24.4976 | 4.7805 | 15.7987 | 24.2398 | 34.1927 | 1 | 3119 |
| model 3 | 24.3373 | 4.9322 | 15.7804 | 23.9781 | 34.8648 | 1 | 3009 |
| model 4 | 24.1116 | 4.8248 | 15.5574 | 23.7915 | 34.4035 | 1 | 3587 |
| model 5 | 24.4041 | 4.9549 | 15.9001 | 24.1139 | 34.7204 | 1 | 3152 |
| model 6 | 24.8840 | 4.9712 | 16.4075 | 24.5082 | 35.4618 | 1 | 2895 |

Table 2.2: Posterior distributions of diff t models for Total, with added anomalies, and calculated with independent Bayes model

10 20 30 40 50



Density

0.00 0.02 0.04 0.06 0.08

0.00 0.02 0.04 0.06 0.08

model3

model5

0.00 0.02 0.04 0.06 0.08

model4

model6

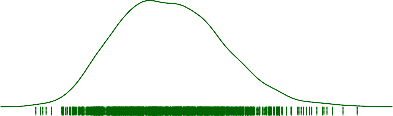
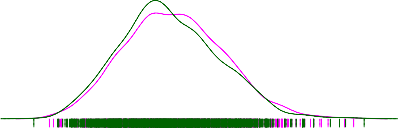
10 20 30 40 50

10 20 30 40

0.00 0.02 0.04 0.06 0.08

model1

10 20 30 40 50



0.00 0.02 0.04 0.06 0.08

0.00 0.02 0.04 0.06 0.08

model2

10 15 20 25 30 35

10 20 30 40 50

Figure 2.6: Posterior distributions of diff t models for Total, with added anomalies, and calculated with independent Bayes model

Table [2.2](#_bookmark49) and fi [2.6](#_bookmark50) gives the posterior distribution of the category total, for each of the model. Model 1 uses non-informative prior, and it gives the most particular posterior out of all models at level 0 (total), the mean value of the posterior distribution is around 4.5 unit lower than other models, there is no standard deviation, and the number of effective samples is 0. Our observation suggests that non-informative priors produce a significantly diff t result compares to weakly informative priors. Posterior distributions for model 2 to model 6 are very similar and very hard to tell any significant diff other than there seems to be a large diff between the number of effective sample size.

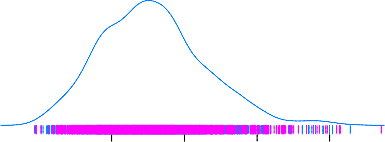
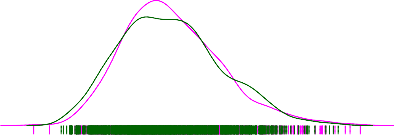
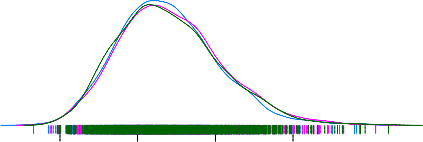
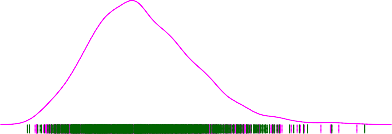
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | sd | 2.5% | 50% | 97.5% | Rhat | n.eff |
| model 1 | 10.5463 | 0.0000 | 10.5463 | 10.5463 | 10.5463 |  | 0 |
| model 2 | 12.3481 | 3.5241 | 6.4608 | 12.0826 | 19.7636 | 1 | 2833 |
| model 3 | 12.2433 | 3.5058 | 6.3331 | 11.9460 | 20.1100 | 1 | 2955 |
| model 4 | 12.0331 | 3.2668 | 6.4515 | 11.7501 | 19.1090 | 1 | 3322 |
| model 5 | 12.3475 | 3.5362 | 6.4799 | 12.0049 | 19.9832 | 1 | 2503 |
| model 6 | 12.9588 | 3.6536 | 6.8353 | 12.6056 | 21.2272 | 1 | 3000 |

Table 2.3: Posterior distributions of diff t models for A , with added anomalies, and calculated with independent Bayes model

0.12

0.12

0 5 10 15 20 25 30



Density

0.04 0.08

0.12

0.00

0.04 0.08

model3

model5

0.05 0.10

model4

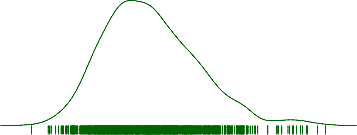
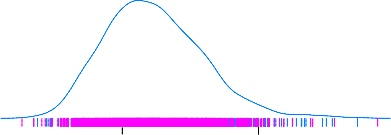
0.00

0.04 0.08

model6

5 10 15 20 25 30

5 10 15 20 25 30



0.00 0.05 0.10 0.15

0.00

model1

0.12

0.00

0 5 10 15 20 25

5 10 15

0.00 0.04 0.08

model2

0 10 20 30

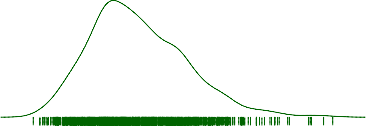
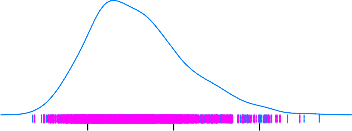
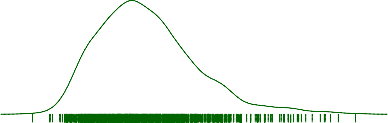
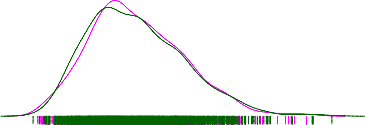
Figure 2.7: Posterior distributions of diff t models for A , with added anomalies, and calculated with independent Bayes model

Table [2.3](#_bookmark51) and fi [2.7](#_bookmark52) gives the posterior distribution of the category A, for each of the model. Model 1 uses non-informative prior, and it gives the most particular posterior out of all models at level 1 (A), the mean value of the posterior distribution is around 2 unit lower than other models, there is no standard deviation, and the number of effective samples is 0. Our observation suggests that non-informative priors produce a significantly diff t result compares to weakly informative priors. Posterior distributions for model 2 to model 6 are very similar and very hard to tell any significant difference, other than there seems to be a large diff between the number of effective sample size.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | sd | 2.5% | 50% | 97.5% | Rhat | n.eff |
| model 1 | 5.2477 | 0.0000 | 5.2477 | 5.2477 | 5.2477 |  | 0 |
| model 2 | 7.4081 | 2.7444 | 2.8634 | 7.1136 | 13.3165 | 1 | 2273 |
| model 3 | 7.3234 | 2.7423 | 2.8910 | 7.0234 | 13.3408 | 1 | 2372 |
| model 4 | 6.9192 | 2.4111 | 3.0404 | 6.6744 | 12.3585 | 1 | 3000 |
| model 5 | 7.6640 | 2.8689 | 3.0550 | 7.2466 | 14.0954 | 1 | 1886 |
| model 6 | 7.8524 | 2.7397 | 3.4434 | 7.5279 | 14.0425 | 1 | 2856 |

Table 2.4: Posterior distributions of diff t models for AA, with added anomalies, and calculated with independent Bayes model

0 5 10 15 20 25



Density

0.00 0.05 0.10 0.15

0.00 0.05 0.10 0.15

model3

model5

0.00 0.05 0.10 0.15

model4

model6

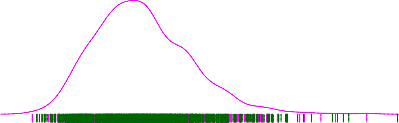
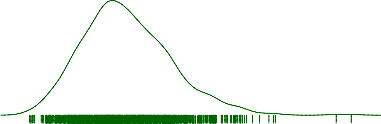
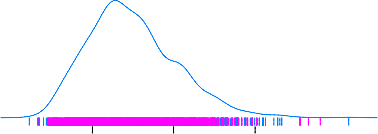
0 5 10 15 20 25

0 5 10 15 20 25

0.0 0.1 0.2 0.3

model1

0 5 10 15 20



0.00 0.05 0.10 0.15

0.00 0.05 0.10 0.15

model2

2 4 6 8

0 5 10 15 20 25

Figure 2.8: Posterior distributions of diff t models for AA, with added anomalies, and calculated with independent Bayes model

Table [2.4](#_bookmark53) and fi [2.8](#_bookmark54) gives the posterior distribution of the category AA, for each of the model. Model 1 uses non-informative prior, and it gives the most particular posterior out of all models at level 2 (AA), the mean value of the posterior distribution is around 2 unit lower than other models, there is no standard deviation, and the number of effective samples is 0. Our observation suggests that non-informative priors produce a significantly diff t result compares to weakly informative priors. Posterior distributions for model 2 to model 6 are very similar and very hard to tell any significant diff rence, other than there seems to be a large diff between the number of effective sample size.

Considering posterior distribution results of all three hierarchical levels, it is apparent that Model 1 gives a significantly diff rent posterior compares to all other models, suggesting that the usage of non-informative prior will generate significantly diff t results compares to weakly informative priors at all levels of Hierarchy. Model 2 to model 6 gives similar posterior distribution, with similar numeric values in tables and similar shapes in density plots in all levels, this suggests that there no or close to no diff between the posterior distributions of Model 2 to model 6 and using slightly diff t weakly informative prior does not make a significant impact on our posterior distribution results in all levels of hierarchy. The mean and standard deviation of posteriors for model 2 to model 6 decreases as the level of hierarchy increase (going down the hierarchy), with category total having the largest mean and standard deviation, category A having smaller mean and standard deviation, and category AA having the smallest mean and deviation. Our observation makes sense as the aggregation constraints of hierarchical datasets dictates that the mean and sample size of categories at higher hierarchy will always be greater than mean and sample size of categories at the lower hierarchy. The number of effective size for model 2 to 6 seem to decrease as the level of hierarchy increase, which made sense because due to aggregation constraints, observed values at a lower hierarchy tend to be smaller and less variable, and this will produce a stronger autocorrelation during the MCMC process.

2000 3000 4000 5000

20.6 21.0 21.410 20 30 40 10 20 30 40 10 20 30 40 10 20 30 40 5100 20 30 40

model1

model2

model3

model4

Iteration number

Figure 2.9: Traces plot for category total of model 1 to model 6

**Autocorrelation plot for category Total of all models**

0 5 10 15 20 25 30

0.10

●

●

●

●

●

●

●

● ●

● ●

●

●

● ●

●

●

● ●

●

● ● ● ●

●● ●

●

●

● ●

● ●

● ●● ●

●

●

● ● ●

● ● ● ● ●

● ●

●

●

● ●

● ● ● ●

● ●

● ●

● ● ●

● ●

●

●

●

● ●

● ● ●

● ●

●

●

●

●

●

●● ● ●

● ●

●

●

●

●

● ●● ● ● ● ● ● ● ●

● ● ● ●●● ●● ●

●

● ●

● ● ● ● ● ● ●

●

●

●● ● ● ● ●

●

●

●

●

●

●

●

●

●

●

●

●

●●

●

●

●

● ● ● ●

●

●

●

●

●

● ● ●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●●

●

●

● ●

●

● ● ●

● ● ●

● ●

●

● ●

● ●

●

●

●

●

●

● ●

● ● ● ● ● ●

● ●

●

●

●

● ●

●

● ●

● ●

●

● ● ● ●

●

● ● ● ● ●●

●

● ● ● ●

●

● ● ● ●

● ● ● ●

● ●●

● ●● ● ● ●● ● ●

● ● ●●

●

● ●

●

●

●

● ● ●

● ●

●

●

●

●

● ● ●

● ●

●●

● ● ● ● ●● ● ● ● ● ●

●

●

● ● ●

●●

● ● ● ● ●

● ● ●

●

●

●

●

●

● ● ●

●

●

● ●

●

●

● ● ●

●

●

● ●

●

●

●

●

●

●

●

●

●

●

●

●

●

● ●

● ● ●

● ● ●

● ●

●

●

● ●●

●

● ●

●

●

● ●

●

●

● ● ●

● ●

●●

●

●● ● ● ● ● ● ● ● ●

● ● ●

● ● ●

● ●

● ●

●

● ●●

● ● ● ● ●● ●

●● ●

●

●

● ●

●

●

●

● ● ●

●

●

0.05

model5

model6

0.00

−0.05

−0.10

0.10

Autocorrelation

0.05

model3

model4

0.00

−0.05

−0.10

0.10

0.05

model1

model2

0.00

−0.05

−0.10

0 5 10 15 20 25 30

Lag

model5

model6

Figure 2.10: ACF plots for category total of model 1 to model 6

Trace plots in fi [2.9](#_bookmark55) allows a visual inspection of convergence at the top level of the hierarchy. The trace plot for model 1 is completely fl suggesting that there is no mixing at all, and trace plots for model 2 to 6 vary around the mean and do not show any trend, this suggest that our samples have mixed well and chains have converged. The Auto Correlation Function (ACF) plots presented in fi [2.10](#_bookmark56) gives no ACF for model 1; this indicates that model 1 gives no autocorrelation. ACF plots for model 2 to model 6 shows no signs of signifi t autocorrelation.

2000 3000 4000 5000

10.2 10.6 11.05 10 20

5 10 20

model3

5 10 20

model4

model1

model2

Iteration number

Figure 2.11: Traces plot for category A of model 1 to model 6

**Autocorrelation plot for category A of all models**

0 5 10 15 20 25 30

0.10

0.05

model5

0.00

−0.05

−0.10

Autocorrelation

model3

0.10

0.05

model1

0.00

−0.05

−0.10

0.10

●

●

●

●

●

●

●

●

● ●

●

●

●

● ● ● ● ● ●● ● ●

● ● ● ● ● ● ●

● ● ●

● ● ● ● ●

●

●

●

●

● ● ● ●

●●

● ●

● ● ●

●

● ● ●

● ●

● ● ●

●

●

●

● ●

●

● ● ●

●●

●

● ● ●

●

● ●

●

● ● ●

● ● ●

●

●●

●

●

● ●

●

●

●

● ●

●

●

● ●● ●

● ●

●

● ●

●

● ●

●

●● ● ● ●●

●

●

●

●

●

●

●

● ●

●

● ●●●

●

● ●

● ●

● ● ●

●● ●

● ●

● ●

●

●

●

● ● ●

●

● ● ●●

●

●

● ●

●

●

●

●

●

●

●●

●

●

●

●

● ● ●

●

● ●

●

● ● ● ●

●

●

● ●

● ●

● ●

● ●

● ●●

● ●

● ● ● ●

●

●

●

● ●

● ● ●

●

●

●

● ●

● ●

●

●

● ● ● ●

●

●

●

●

●●

●

●

●

● ●

●

●

●

● ● ●

● ● ● ●

● ●

●

●● ●

●

●

● ●

●

●

●

● ● ● ●

●

●● ●

●

●● ● ●● ●

● ●

●● ●● ●

● ● ●

●

● ●

● ● ●

●

●

●

●

●

●

●

●

●

●

● ●

●

●

●

● ●

● ● ●

●

●

●

●

●

●

●

●

●

● ●

●

●

●

●

●

●

● ●

●

●

●

●

●

●

●

● ● ●

● ●

● ● ● ●

● ●

● ●

● ● ●●

● ● ● ● ●

●

●●●

●

●

●

● ●

●● ● ● ●

●

● ● ●

●

● ●

● ●

●●● ●

●

●

● ● ●

●

●

●

●

●

● ●●

●

● ●

●

●

●● ● ● ● ●

●

●

●

● ● ●

●

●

● ●

●

●

model6

0.05

0.00

−0.05

−0.10

model2

0 5 10 15 20 25 30

model4

Lag

5 10 20

model5

5 10 20

model6

Figure 2.12: ACFs plot for category A of model 1 to model 6

Trace plots in fi [2.11](#_bookmark57) allows a visual inspection of convergence at the second level of the hierarchy. The trace plot for model 1 is completely fl suggesting that there is no mixing at all, and trace plots for model 2 to 6 vary around the mean and do not show any trend, this suggest that our samples have mixed well and chains have converged. The Auto Correlation Function (ACF) plots presented in fi [2.12](#_bookmark58) gives no ACF for model 1. Our observation indicates that model 1 gives no autocorrelation. ACF plots for model 2 to model 6 shows no signs of significant autocorrelation.

2000 3000 4000 5000

4.8 5.2 5.6

5 10 20

5 101520

5 10 15 20 5 101520

model1

model2

model3

model4

Iteration number

Figure 2.13: Traces plot for category AA of model 1 to model 6

**Autocorrelation plot for category AA of all models**

0 5 10 15 20 25 30

0.2

●

●

●

●

●

●

●

●

● ●

●● ●

●

●

●

● ●

●

●

●

● ●

● ●

● ● ● ●

● ● ● ● ● ● ●

●

● ●● ●

● ●

●

●

●

● ● ● ●

● ●

● ●

●

● ●● ●●

●

● ● ● ●

●

●

●

●

●

● ●

●● ●

●●

● ●●

● ●

● ●

●

● ●

●

● ●● ● ● ● ● ● ● ● ● ●

● ● ●

● ● ● ● ●

● ● ●

●

● ●● ● ●

●

●● ●● ● ●

● ●

●

● ●

●

●●● ●

● ● ●

● ● ●●● ● ●

●

●

● ● ● ● ● ●

●

●

● ● ●● ● ● ● ● ● ● ●

●

● ● ● ●

● ● ●

●

● ● ● ●

● ● ● ● ●

● ● ● ● ● ● ● ● ● ●

●

●

● ● ●

● ●

● ●● ● ● ●● ●● ●

●

●

●

● ●

●

● ●

●

● ● ● ● ● ●

● ● ●

●

●

●

●

● ●

●

● ●

● ● ●

●

●

●

●

●

●

●

●●

●

● ● ●

●

● ● ● ● ●● ● ●

●● ● ● ● ● ●

● ●

●

●

●

● ● ● ●

● ●

●

●

●

●

● ● ● ● ●

● ●

●

● ●●

● ● ● ● ● ●

● ● ●● ●●

●

● ●● ●

●

●● ●

● ● ● ● ● ● ●

● ●● ● ● ● ● ●

●

● ● ●

●

● ●● ●

●

●

● ● ● ●

●

● ● ●

● ● ● ●

● ● ● ● ●

●

● ● ●

● ●●

●●

●●

● ● ● ● ● ●

●

● ● ●●

● ●

●

●

● ● ● ●

● ●

● ● ●

●

●

●

● ●

● ●

●

● ●

●

● ● ● ●

●

0.0

model4

model5

model6

−0.2

Autocorrelation

0.2

0.0

model1

model2

model3

−0.2

0 5 10 15 20 25 30

Lag

0 5 10 15 20 25 30

5 101520

model5

model6

Figure 2.14: ACFs plot for category AA of model 1 to model 6

Trace plots in fi [2.13](#_bookmark59) allows a visual inspection of convergence at the second level of the hierarchy. The trace plot for model 1 is completely fl suggesting that there is no mixing at all, and trace plots for model 2 to 6 vary around the mean and do not show any trend, this suggest that our samples have mixed well and chains have converged. The Auto Correlation Function (ACF) plots presented in fi [2.14](#_bookmark60)

gives no ACF for model 1; this indicates that model 1 gives no autocorrelation. ACF plots for model 2 to model 6 shows a weak sign of autocorrelation.

Considering trace plots and ACF plots results of all three hierarchical levels, we observe no strong correlation at all levels of hierarchy. However, there seems to be a slight increase in autocorrelation at lower levels of hierarchy. The is indicate that categories at lower hierarchy mixes slower compare to categories at the higher hierarchy.

**2.3.4 Discussion**

As expected, models that used weakly-informed priors had outperformed the model with non-informative prior, in terms of model goodness-of-fi this highlighted the importance of assigning a prior distribution for Bayesian calculations and we should at least provide some distribution to our prior, even if it is just a fl uniform prior (which is also considered as a non-informative prior). Different hypothesised weakly-informed priors performed similarly, and we were unable to detect a significant diff between our priors from the DIC comparisons, and we did not observe significant diff in posterior distributions. If there is no evidence suggesting a prior is better than another prior, how do we decide which prior is preferable? The reasonable idea is that we should use the simplest prior, so complex priors such as the mixture prior and Laplace prior is out of the question. Especially for the case of Hierarchical Bayesian Model (HBM), we would want to build our HBM with the simplest prior as possible. Otherwise, our model would be extremely complex and hard to build. Hence, a Normal(1,0.3) prior is used for the rest of our simulations. A model with normal distributions is simple and suitable for observations with large samples as indicated by Central limit theorem (CLT) theorem. The essential idea is that weakly informative prior should contain enough information to regularize, this means that the prior should rule out unreasonable parameter values but is not so strong as to rule out values that might make sense. The might be other considerations as well, for example the Stan Development Team favour the lighter tails of a weakly informative prior with a Gaussian shape as the most robust default prior for it’s accurate and fast statistical computation and ease of optimising the scale [(Betancourt,](#_bookmark167) [2018),](#_bookmark167) and [Gelman et al.](#_bookmark185) [(2006)](#_bookmark185) suggested the usage of half-Cauchy prior distribution for observations with small values.

Another interesting fi from this simulation is that uncertainty in Bayesian estimations will cause uncertainty in DIC results, and it is tough to confidently decide which model is the best if there is no substantial diff between the performance of models. Increasing the iteration number help a bit, but does not eliminate the problem and uncertainty still exists. A possible alternative to DIC is (Watanabe-Akaike information criteria), but it relies on a data partition that would cause difficulties with structured models [(Gelman et al.,](#_bookmark187) [2014).](#_bookmark187) Cross-validation could

also work, but for Bayesian stats, it will be computationally expensive. Our results agree with conclusions from [Gelman et al.](#_bookmark187) [(2014)](#_bookmark187), that predictive error comparisons (such as DIC) does work for highly dissimilar models, but have a weak performance for similar models, and sometimes it is challenging to draw significant conclusions.

* 1. **Simulation 2: Anomaly**
     1. **Simulation setups**

In section 2.4, we present the results of the posterior distribution calculations for the synthesised data, with diff t increments of anomalies added on category AA on level 2 (leaf) of the hierarchy. The default setting for the data synthesis process is applied for each scenario, with the addition of 0%, 10%, 25%, 50%, 100%, 250% and 500% anomaly. An important note, the 25% anomaly here meant an addition of 25% of the count of category total on top of the hierarchy, not addition of 25% on category AA.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **2.4.2** | **Results** |  | | | | | |
|  | Anomaly |
|  | (% of total) | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|  | 0 | -0.466 | 0.188 | 0.437 | 0.545 | 0.955 | 1.556 |
|  | 10 | -1.910 | -0.693 | 0.876 | 1.144 | 1.894 | 6.640 |
|  | 25 | -6.274 | -3.079 | -1.966 | -1.984 | -0.658 | 1.825 |
|  | 50 | -4.352 | -1.185 | -0.038 | -0.385 | 0.823 | 2.420 |
|  | 100 | -5.074 | -1.096 | 0.323 | -0.383 | 1.194 | 1.873 |
|  | 250 | -1.813 | -1.657 | -0.917 | -0.181 | 0.546 | 3.681 |
|  | 500 | -4.416 | -1.678 | -0.613 | -0.875 | 0.430 | 1.401 |

Table 2.5: DIC (with 1,000,000 iteration) comparisons between Independent Bayes model and Hierarchical Bayes model (HBM -IBM) for diff t anomalies sizes

Table [2.5](#_bookmark65) gives the comparison of DIC distributions of Independent Bayesian Model (IBM) and Hierarchical Bayesian Model (HBM) for scenarios with diff t increments of anomaly added. For all scenarios, the mean and median value is mostly negative; the negative signals suggest that HBM seem to perform better than IBM, most of the time. Uncertainty in Bayesian estimations still exist and have impacts on the reliability of our DIC comparison, although we mostly see negative values for median and mean of the DIC difference distribution, the signal still shifts every time

our JAGS model is re-run. Our observation suggests that although there might be a diff between IBM and HBM, it might not be a very big diff

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | sd | 2.5% | 50% | 97.5% | Rhat | n.eff |
| Anomaly0 | 19.1481 | 4.2840 | 11.7152 | 18.8136 | 28.3845 | 1 | 3000 |
| Anomaly10 | 21.4084 | 4.7272 | 13.1528 | 21.0041 | 31.6607 | 1 | 3000 |
| Anomaly25 | 24.4356 | 4.9374 | 16.0687 | 24.0399 | 34.7248 | 1 | 2975 |
| Anomaly50 | 30.2932 | 5.3613 | 20.8844 | 29.8973 | 41.9325 | 1 | 2912 |
| Anomaly100 | 40.4338 | 6.1531 | 29.3330 | 40.0987 | 53.8314 | 1 | 2889 |
| Anomaly250 | 71.9683 | 8.4372 | 56.5531 | 71.6763 | 89.4346 | 1 | 3000 |
| Anomaly500 | 122.4365 | 10.7971 | 101.9952 | 122.2718 | 144.5457 | 1 | 3132 |

Table 2.6: Posterior distributions of diff t models for Total, with diff t incre- ments of added anomalies, and calculated with independent Bayes model

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | sd | 2.5% | 50% | 97.5% | Rhat | n.eff |
| Anomaly0 | 19.3114 | 4.4361 | 11.5960 | 18.9555 | 28.9686 | 1 | 2934 |
| Anomaly10 | 21.3883 | 4.5971 | 13.3033 | 21.1152 | 31.2900 | 1 | 2799 |
| Anomaly25 | 24.3800 | 4.9083 | 15.8281 | 24.1057 | 34.9855 | 1 | 3000 |
| Anomaly50 | 30.5664 | 5.4857 | 20.7071 | 30.2800 | 42.3242 | 1 | 2713 |
| Anomaly100 | 40.5794 | 6.2480 | 29.1623 | 40.2459 | 53.0399 | 1 | 2704 |
| Anomaly250 | 72.0177 | 8.2580 | 57.1126 | 71.8129 | 88.7105 | 1 | 2723 |
| Anomaly500 | 121.9641 | 10.3188 | 102.7605 | 121.6646 | 142.4118 | 1 | 2173 |

Table 2.7: Posterior distributions of diff t models for Total, with diff t incre- ments of added anomalies, and calculated with Hierarchical Bayes model

Table [2.6](#_bookmark66) and table [2.7](#_bookmark67) gives the posterior distribution of the category total, with diff t increments of anomalies added, for IBM and HBM. There are diff between the mean and standard deviations of posteriors of IBM and HBM, but it is very small and very hard to tell. However, for HBM The number of Effective Sample Size (ESS) for HBM posteriors seem to be overall less than IBM posteriors, this suggests that values observed in total category on top of the hierarchy level for HBM are overall less independent than BM. As the size of the anomaly increases, the number of ESS tend to decrease.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | sd | 2.5% | 50% | 97.5% | Rhat | n.eff |
| Anomaly0 | 7.1758 | 2.7278 | 3.0586 | 6.7391 | 13.6749 | 1 | 2844 |
| Anomaly10 | 9.3215 | 3.1367 | 4.1286 | 8.9552 | 16.3998 | 1 | 2630 |
| Anomaly25 | 12.3410 | 3.5655 | 6.4488 | 11.9857 | 20.3604 | 1 | 3038 |
| Anomaly50 | 18.4421 | 4.3162 | 10.8431 | 18.1170 | 27.7122 | 1 | 2839 |
| Anomaly100 | 28.3032 | 5.3432 | 19.1588 | 27.8652 | 40.0187 | 1 | 3008 |
| Anomaly250 | 58.6906 | 7.3628 | 45.1079 | 58.6335 | 73.7809 | 1 | 2118 |
| Anomaly500 | 104.7585 | 9.5672 | 87.1657 | 104.5580 | 123.4846 | 1 | 2235 |

Table 2.8: Posterior distributions of diff t models for A , with diff t increments of added anomalies, and calculated with Independent Bayes model

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | sd | 2.5% | 50% | 97.5% | Rhat | n.eff |
| Anomaly0 | 7.1740 | 2.7263 | 2.8494 | 6.8467 | 13.2835 | 1 | 2559 |
| Anomaly10 | 9.1731 | 3.0292 | 4.1612 | 8.8307 | 15.8837 | 1 | 2961 |
| Anomaly25 | 12.4274 | 3.6049 | 6.3125 | 12.0656 | 20.2620 | 1 | 2703 |
| Anomaly50 | 18.5048 | 4.3162 | 10.9271 | 18.2437 | 28.0090 | 1 | 2580 |
| Anomaly100 | 28.6123 | 5.2341 | 19.2090 | 28.1992 | 39.9182 | 1 | 2370 |
| Anomaly250 | 59.6236 | 7.2375 | 46.4370 | 59.4535 | 74.8256 | 1 | 1940 |
| Anomaly500 | 107.9043 | 9.7689 | 89.7498 | 107.4554 | 129.0404 | 1 | 1338 |

Table 2.9: Posterior distributions of diff t models for A , with diff t increments of added anomalies, and calculated with Hierarchical Bayes model

Table [2.8](#_bookmark68) and table [2.9](#_bookmark69) gives the posterior distribution of the category A, with diff t increments of anomalies added, for IBM and HBM. There are also diff between the mean and standard deviations of posteriors of IBM and HBM, but it is also very small and very hard to tell. The number of ESS for HBM posteriors seem to be overall less than IBM posteriors. Our observation suggests that values observed in A category on second of the hierarchy level for HBM are overall less independent than BM. As the size of the anomaly increases, the number of ESS tend to decrease.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | sd | 2.5% | 50% | 97.5% | Rhat | n.eff |
| Anomaly0 | 2.3062 | 1.4859 | 0.4009 | 2.0034 | 6.0143 | 1 | 2401 |
| Anomaly10 | 4.2976 | 2.1384 | 1.2137 | 3.8992 | 9.5420 | 1 | 2461 |
| Anomaly25 | 7.4243 | 2.8041 | 2.8709 | 7.0744 | 13.8003 | 1 | 2214 |
| Anomaly50 | 13.3455 | 3.5920 | 7.3160 | 13.0175 | 21.0238 | 1 | 2428 |
| Anomaly100 | 22.4374 | 4.5194 | 14.6914 | 22.0898 | 31.9623 | 1 | 2029 |
| Anomaly250 | 48.8502 | 6.1544 | 37.2238 | 48.5964 | 61.2842 | 1 | 1568 |
| Anomaly500 | 100.5322 | 11.0472 | 79.0267 | 100.3259 | 122.1376 | 1 | 1115 |

Table 2.10: Posterior distributions of diff t models for AA, with diff t increments of added anomalies, and calculated with Independent Bayes model

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | sd | 2.5% | 50% | 97.5% | Rhat | n.eff |
| Anomaly0 | 2.1538 | 1.4960 | 0.3505 | 1.7838 | 6.1010 | 1 | 2073 |
| Anomaly10 | 4.3301 | 2.1448 | 1.1920 | 3.9651 | 9.5508 | 1 | 2046 |
| Anomaly25 | 7.6847 | 2.8131 | 3.0945 | 7.3379 | 13.8861 | 1 | 2149 |
| Anomaly50 | 13.6061 | 3.6369 | 7.2589 | 13.3337 | 21.4732 | 1 | 2054 |
| Anomaly100 | 23.0253 | 4.5067 | 14.9764 | 22.7288 | 32.6668 | 1 | 1679 |
| Anomaly250 | 51.9695 | 6.3876 | 40.2516 | 51.8076 | 65.5484 | 1 | 951 |
| Anomaly500 | 96.0807 | 9.1704 | 79.5825 | 95.8225 | 115.7111 | 1 | 951 |

Table 2.11: Posterior distributions of diff t models for AA, with diff t increments of added anomalies, and calculated with Hierarchical Bayes model

Table [2.10](#_bookmark70) and table [2.11](#_bookmark71) gives the posterior distribution of the category AA, with diff t increments of anomalies added, for IBM and HBM. There are also diff between the mean and standard deviations of posteriors of IBM and HBM, but it is also very small and very hard to tell. The number of ESS for HBM posteriors seem to be overall less than IBM posteriors. Our observation suggests that values observed in the AA category on the bottom of the hierarchy level for HBM are overall less independent than IBM. As the size of the anomaly increases, the number of ESS tend to decrease.

Overall, we make a similar observation on posterior distributions for categories on diff t levels of the hierarchy, for scenarios with diff t increments of anomalies added. Increase of size of anomalies will result in higher mean and standard deviation, and a smaller ESS overall, for categories on all levels of the hierarchy. Differences in posterior distributions of HBM and IBM seem to exist, but it is tiny and is likely to be masked by the noise from uncertainties in Bayesian estimates. However, the variability in diff seem to be greater for categories at a lower level of the hierarchy, i.e. there is a diff of 0.47, 2.74, and 4.49 in IBM and HBM means of 500% anomaly. Our observation suggests that the HBM model have a greater diff in performance, for observations in lower levels of hierarchy. The ESS also has a significant overall decrease for our posterior results in lower levels of hierarchy, suggesting that IBM greatly reduce the independence of our results.

### Notes on heat-maps

As mentioned in section [1.4,](#_bookmark5) a hypothetical value called capacity (*k*) is used to set up the threshold for what we believe to be values that reflect anomalous events. The basic idea is that *k* is a measure of hospital capacity, once the capacity is reached, congestion will defi occur. The threshold is simply the inverse of *k*, so a *k* of 0.8 would mean that the threshold is set to be 1.25 times the expected mean arrival *µ*. A larger *k* stands for higher capacity and would result in a smaller multiplication factor for *µ* for the threshold value. Vice versa, a smaller *k* stands for lower capacity and would result in a larger multiplication factor for *µ* for the threshold value. The idea makes sense because when hospitals operate at high capacity, it will not take much to overfill their capacity. However, when a hospital is operating at a lower capacity, there is a large buff and it would take a sizeable abnormal arrival to fi the capacity.

A threshold is used to calculate the area of the posterior distribution that falls under the distribution, that exceeds the threshold. The Threshold gives a single measure of posterior probability *θ* that is over the threshold, higher the *θ* more of the posterior probability will fall over the threshold, and smaller the *θ* less posterior probability will fall under the threshold. A *θ* close to 0.50 would suggest that our median value of posterior probability is falling right on the threshold. Larger the *θ* more confident we have for the occurrence of an anomaly. *θ* from our models and scenarios are then tabulated with diff t increments of *k* and presented in A 10 level colour-coded heat map. The colour of the heat map is in the classic red and blue spectrum, with red indicating a larger *θ* and a higher probability of observing an anomaly, and blue indicating a smaller *θ* and a lower probability of observing an anomaly.

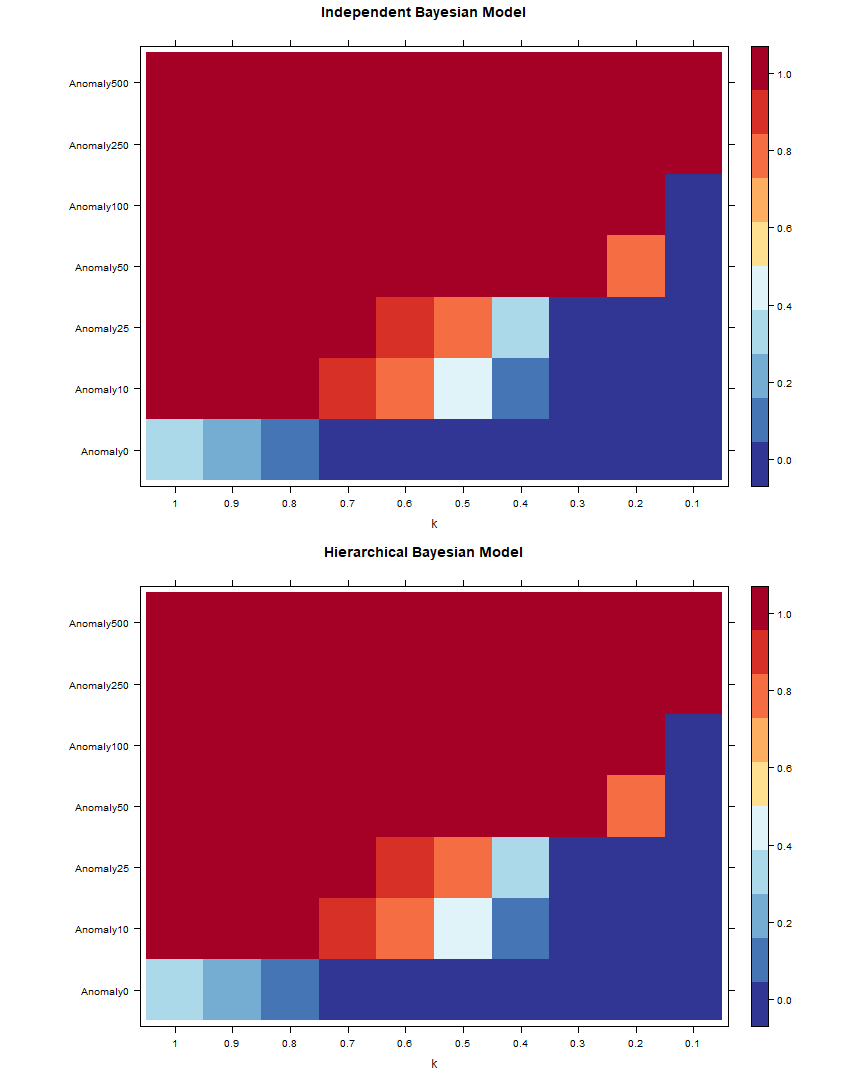


Figure 2.15: Heat-map of anomaly alarms for diff t size of anomaly under diff t k values, of IBM and HBM, at category total.

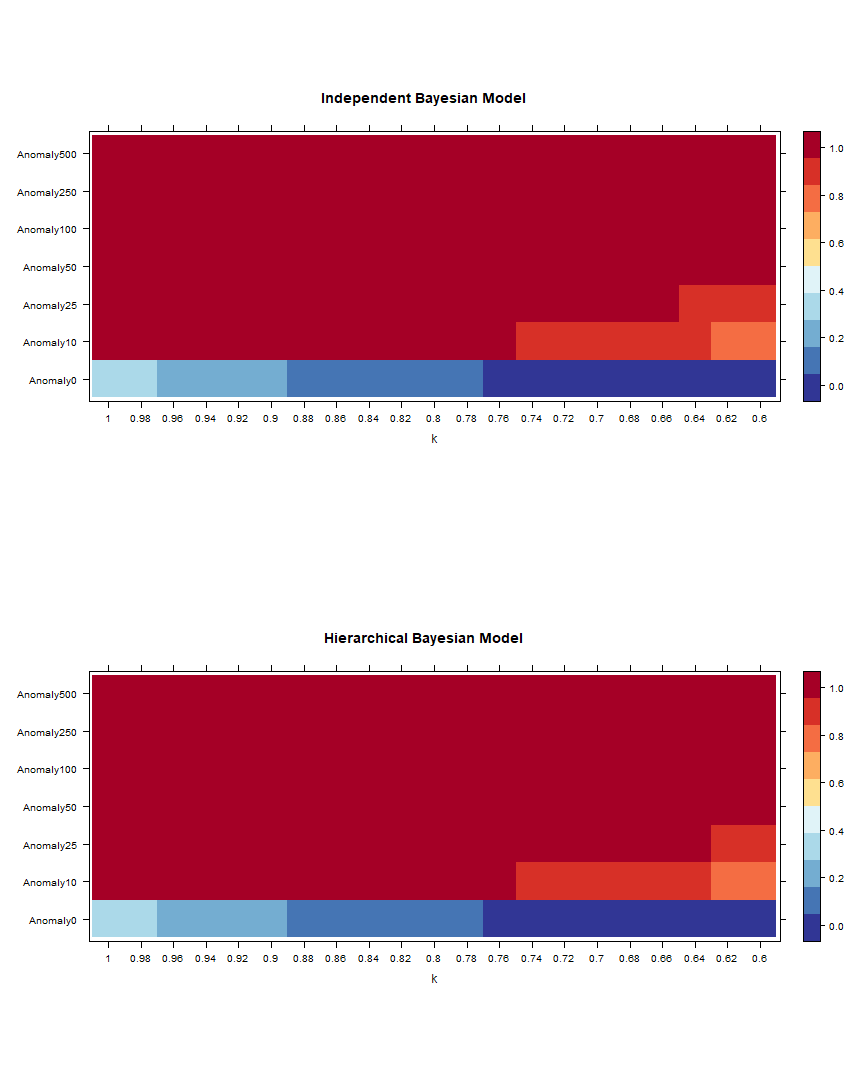


Figure 2.16: Heat-map of anomaly alarms for diff t size of anomaly under diff t k values, of IBM and HBM zoomed in at k values of 0.6 to 1, at category total.

The heat-map in fi [2.15](#_bookmark72) presents the anomaly signals of category total for diff t increments of anomaly added, with IBM and HBM, against diff t increments of *k*. Strong anomaly signals (in red) tend to appear toward the top left corner of our heap-map, this indicated that signals of anomaly tend to increase in strength with greater anomaly and higher capacity *k*. Zooming in at looking at *k* values from 0.6 to 1 in fi [2.16,](#_bookmark73) we observe a minimal diff in anomaly detection rate between IBM and HBM for category total. Our observation suggests that the diff between anomaly detection rate is small or close to none for IBM and HBM, on the top level of the hierarchy, for all sizes of the anomaly, in a simulated setting with each category generated independently.

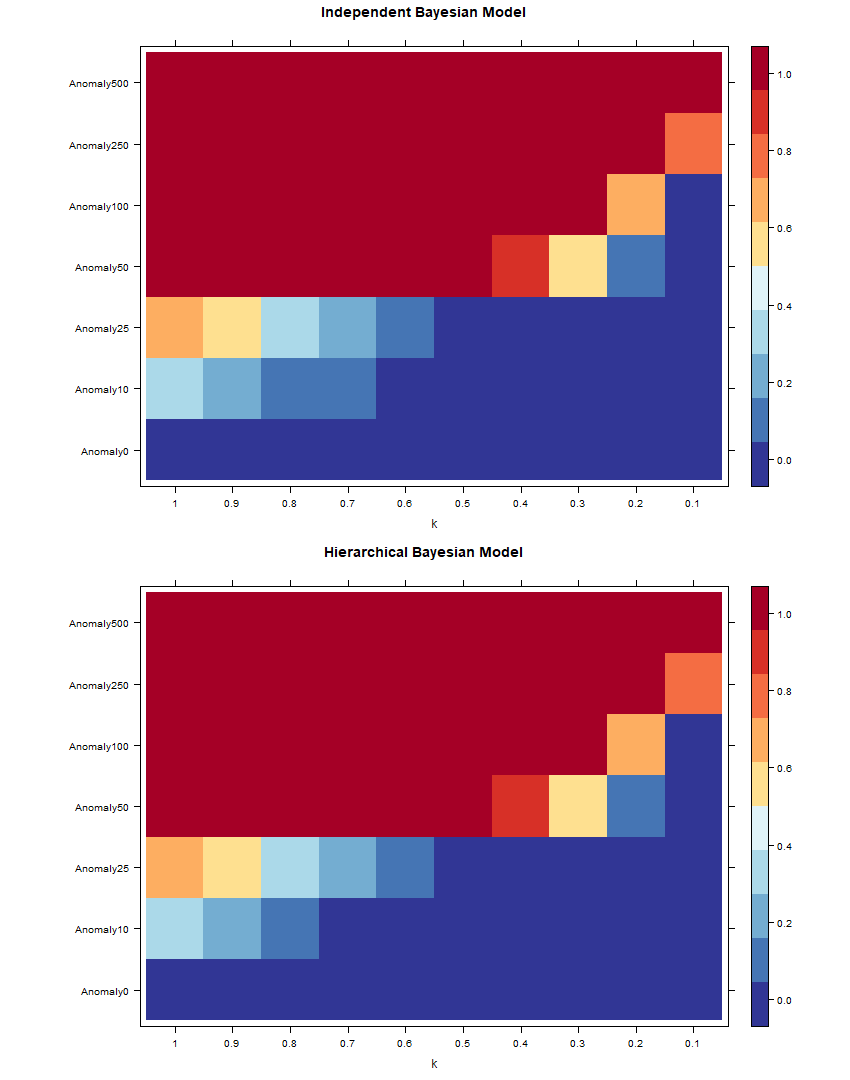


Figure 2.17: Heat-map of anomaly alarms for diff t size of anomaly under diff t k values, of IBM and HBM, at category A.

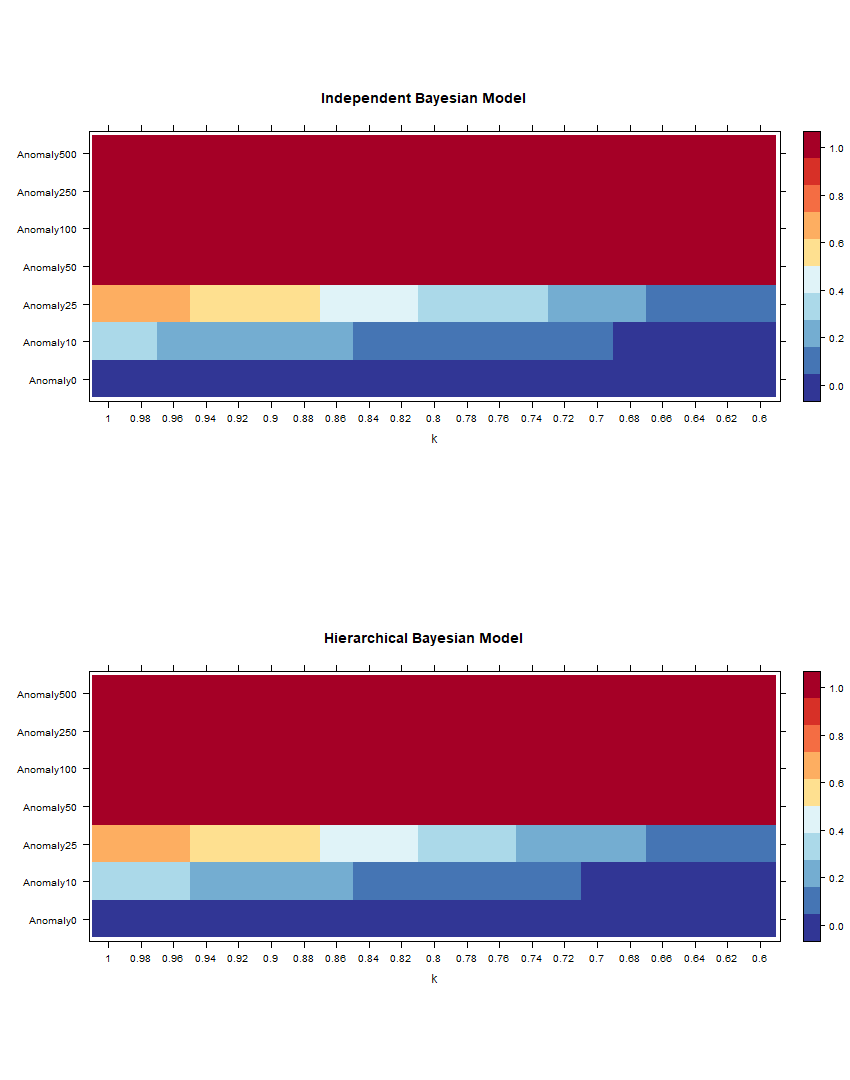


Figure 2.18: Heat-map of anomaly alarms for diff t size of anomaly under diff t k values, of IBM and HBM zoomed in at k values of 0.6 to 1, at category A.

The heat-map in fi re [2.17](#_bookmark74) presents the anomaly signals of category A for diff t increments of anomaly added, with IBM and HBM, against different increments of

*k*. Strong anomaly signals (in red) tend to appear toward the top left corner of our heap-map, this indicated that signals of anomaly tend to increase in strength with greater anomaly and higher capacity *k*.Zooming in at looking at k values from 0.6 to 1 in fi [2.18,](#_bookmark75) we observe a minimal diff e in anomaly detection rate between IBM and HBM for category A. Our observation suggests that the diff between anomaly detection rate is small or close to none for IBM and HBM, on level 2 of the hierarchy, for all sizes of anomaly, in a simulated setting with each category generated independently.

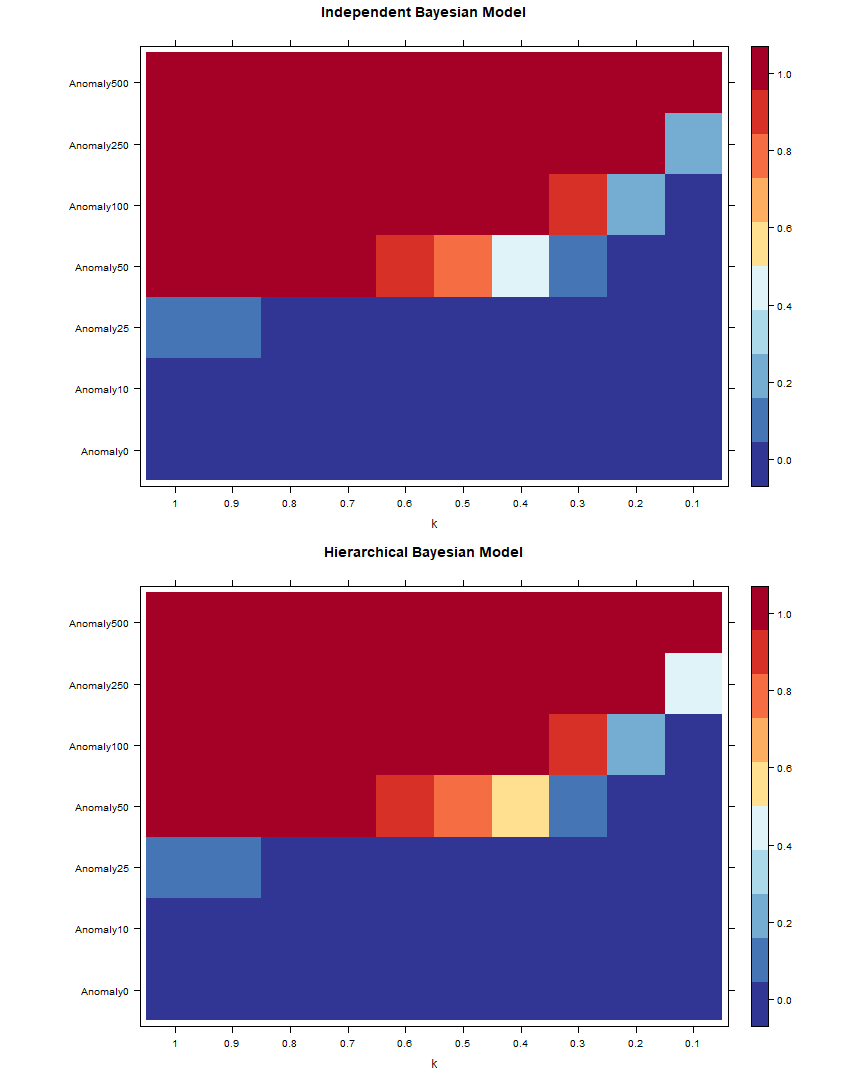


Figure 2.19: Heat-map of anomaly alarms for diff t size of anomaly under diff t k values, of IBM and HBM, at category AA.

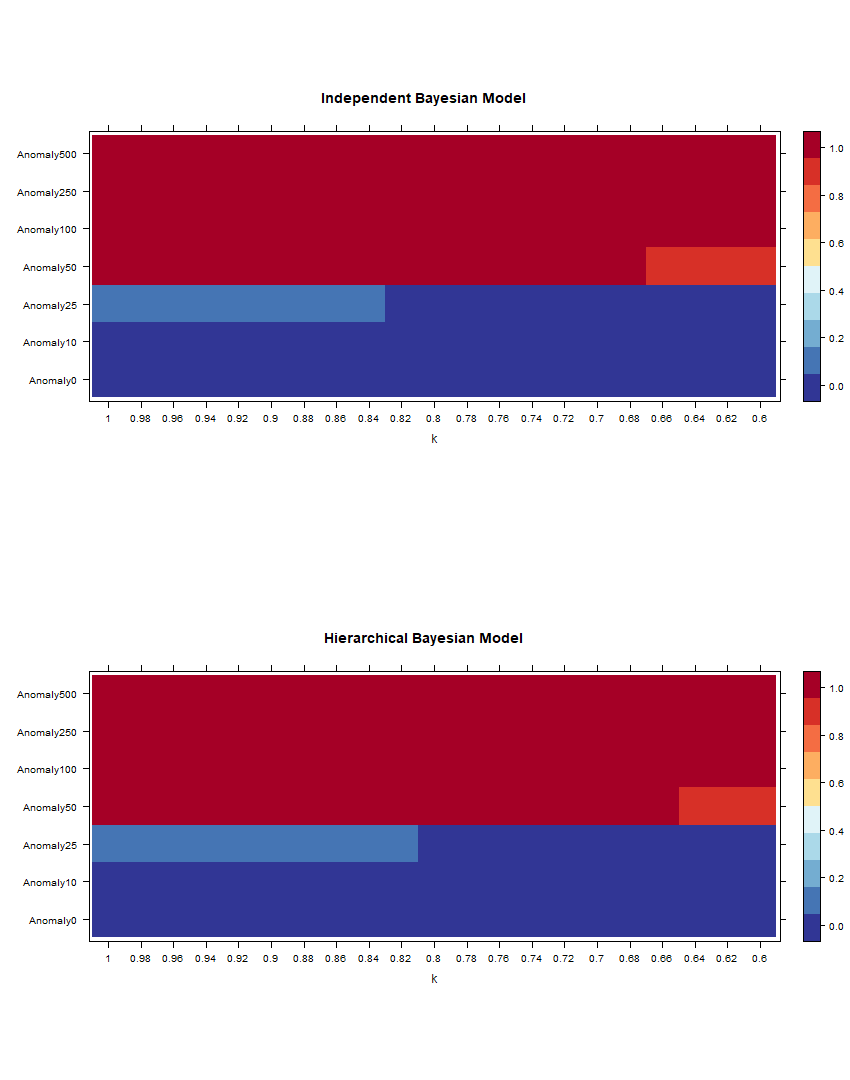


Figure 2.20: Heat-map of anomaly alarms for diff t size of anomaly under diff t k values, of IBM and HBM zoomed in at k values of 0.6 to 1, at category AA.

The heat-map in fi [2.19](#_bookmark76) presents the anomaly signals of category AA for diff t increments of anomaly added, with IBM and HBM, against diff t in- crements of *k*. Strong anomaly signals (in red) tend to appear toward the top left corner of our heap-map, this indicated that signals of anomaly tend to increase in strength with greater anomaly and greater capacity *k*. Zooming in at looking at k values from 0.6 to 1 in fi [2.20,](#_bookmark77) we observe a minimal diff in anomaly detection rate between IBM and HBM for category AA. Our observation suggests that the diff between anomaly detection rate is small or close to none for IBM and HBM, on the bottom of the hierarchy, for all sizes of the anomaly, in a simulated setting with each category generated independently.

Comparing between levels of hierarchies, we observe that for all levels of hierarchy, strong anomaly signals (in red) tend to appear toward the top left corner of our heap-map, confirming the fact that signals of anomaly tend to increase in strength with greater anomaly and higher capacity *k*, for all levels of hierarchy. For all levels of hierarchy, We observe powerful red signals for scenarios with 50% and more anomalies added, and very weak signals for the case of no anomaly, 10% anomaly and 25% anomaly. The overall signals for higher hierarchies tend to be stronger than overall signals in lower hierarchies.

## 2.4.3 Discussion

Our DIC results suggest that compared to IBM, HBM gives better goodness of fi and is the better model, for scenarios with diff t increments of anomalies added to a category at the bottom level of the hierarchy, this provided more evidence that suggests that HBM are superior compared to IBM. Although our DIC comparisons result varies every time we re-ran our JAGS models, the observation appears most of the time, and we have some evidence to back our claim. The fi is consistent with the fi in [Gelman](#_bookmark184) [(2006),](#_bookmark184) that hierarchical models tend to give a better prediction compare to no-pooling (independent) models in all levels. Although our DIC result suggests that HBM gives a diff t and better posterior distribution, it is kind of hard to tell the diff nces between posterior distributions for IBM and HBM with all the noise present, and this reduced out confidence to our claim. To rule out the noise from uncertainties in Bayesian posterior is a complex and calculation-intensive process and are out of our scope.

As the size of the anomaly increases, the number of ESS tend to decrease. The observation is due to the fact that a more substantial increase in anomaly results in a more substantial, more infl tial constant change in observed values, that produces a strong autocorrelation relationship and this will reduce the indecency

of the values, this is also the reason why the standard deviation of posterior also tend to increase as we increase the size of anomalies. The reasoning also applies to hierarchical levels, due to the aggregation constraint values at each level tend to increase, and sample distributions tend to approximate towards normal. So when we are dealing with hierarchical datasets, we must take into account the hierarchical structure, there are significant statistical information within the hierarchy and disregarding the hierarchical structure would significantly reduce the power of our models. A tell-tale sign that IBM had disregarded information about the hierarchical structure is that posterior distributions of IBM tend to have a greater ESS compare to HBM, information between levels often translate into correlations, and without this information, the posterior results in each category would be more independent have a greater ESS. The loss of information becomes more apparent for lower levels of hierarchy; this is because values of a category at a lower level of hierarchy often infl by an increasing number of more population-level parameter.

With our posterior distributions, we were able to set up a theoretical threshold and made attempts to detect anomalies. Signals of anomaly tend to increase in strength with larger anomalies and higher capacity (*k*), for all levels of hierarchy. Our observations made sense; it is easier to detect larger anomalies because the posterior will be larger and more likely to fall over the threshold. Also, as capacity increase, the average expected value closer to the threshold, and it becomes easier a higher than expected value to go over the threshold. I.e. for a capacity of 80%, we need a positive anomaly 1.25 times the expected value to meet the threshold, for a capacity of 40%, we need a positive anomaly 2.5 times the expected value to meet the threshold. Relating to the idea of congestion mentioned in the previous section, for a hospital to decrease the likelihood of encountering congestion, we would need to decrease the operation capacity of its facilities. Operating at a high capacity may save resource, but is the downside worth it? Did the policymaker take in an account of adverse effects such as doctors getting burnout and an increasing likelihood of congestion, and does the risk out weight the benefit?

# Simulation 3: Structure

## Simulation setups

In section 2.5, we present the results of the posterior distribution calculations for the synthesised data, with same 10% anomaly added on category AA on level 2 (leaf) of the hierarchy, but for hierarchies with a diff t structure. Most of the hierarchical structure is controlled to be the same, except for the brunching at category A. We create 12 scenarios with category A brunching into 1 to 12 subgroups equal in size. So for brunch = 1, category A will brunch into one subcategory AA, category AA will contain 100% of the size of the category. For brunch = 3, category A will brunch into three subcategory AA, AB and AC, category AA, AB and AC will each contain 33% of the size of category A.

## Results

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Brunch | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
| 1 | -7.033 | -1.960 | -0.809 | -1.527 | 0.294 | 1.172 |
| 2 | -4.435 | -2.722 | -0.994 | -0.591 | 0.214 | 6.308 |
| 3 | -2.710 | -1.013 | 0.554 | 0.521 | 1.440 | 4.575 |
| 4 | -3.418 | -0.451 | 0.370 | 0.305 | 1.422 | 3.356 |
| 5 | -2.715 | -1.447 | -1.065 | -0.318 | 1.022 | 2.791 |
| 6 | -3.082 | -0.733 | -0.129 | 0.121 | 1.042 | 4.216 |
| 7 | -4.280 | -2.843 | -1.286 | -1.581 | -0.213 | 0.790 |
| 8 | -4.349 | -2.215 | -0.967 | -0.422 | 1.065 | 7.971 |
| 9 | -4.931 | -1.649 | -0.335 | -0.303 | 1.014 | 4.038 |
| 10 | -5.247 | -1.578 | -0.234 | -0.668 | 0.618 | 3.247 |
| 11 | -2.597 | -1.126 | 1.002 | 0.466 | 1.655 | 3.751 |
| 12 | -3.340 | -1.786 | -0.102 | 0.223 | 1.838 | 4.623 |

Table 2.12: DIC comparisons for brunching number at A

Table [2.12](#_bookmark82) gives the comparison of DIC distributions of Independent Bayesian Model (IBM) and Hierarchical Bayesian Model (HBM) for scenarios with a diff t hierarchical structure. For all scenarios, the mean and median value is mostly negative; the negative signals suggest that HBM seem to perform better than IBM, most of the time. Uncertainty in Bayesian estimations still exist and have impacts on the reliability of our DIC comparison, although we mostly see negative values for median and mean of the DIC diff distribution, the signal still shifts every time our JAGS model is re-run. Our observation suggests that although there might be a diff between IBM and HBM, it might not be a very big diff

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Brunch | mean | sd | 2.5% | 50% | 97.5% | Rhat | n.eff |
| 1 | 21.2367 | 4.7008 | 12.9239 | 20.9013 | 31.1209 | 1 | 2787 |
| 2 | 21.4198 | 4.6439 | 13.5516 | 20.9515 | 31.6895 | 1 | 2904 |
| 3 | 30.4154 | 5.6673 | 20.1950 | 30.2022 | 42.1275 | 1 | 2877 |
| 4 | 13.1540 | 3.5731 | 6.9948 | 12.7660 | 21.1359 | 1 | 3000 |
| 5 | 26.4296 | 5.0584 | 17.2806 | 26.1034 | 37.3137 | 1 | 2815 |
| 6 | 22.3157 | 4.7457 | 14.0816 | 22.0346 | 32.1219 | 1 | 2881 |
| 7 | 17.2269 | 4.1264 | 10.2189 | 16.7961 | 26.4545 | 1 | 2587 |
| 8 | 19.1497 | 4.3697 | 11.4783 | 18.8391 | 28.4958 | 1 | 3000 |
| 9 | 27.4028 | 5.2818 | 18.0539 | 26.9926 | 38.6413 | 1 | 2854 |
| 10 | 18.2805 | 4.2982 | 10.6334 | 18.0157 | 27.3239 | 1 | 3028 |
| 11 | 22.2082 | 4.7100 | 13.9406 | 21.8632 | 32.7639 | 1 | 3000 |
| 12 | 23.2141 | 4.8762 | 14.6652 | 22.8895 | 33.4077 | 1 | 3090 |

Table 2.13: Posterior distributions of diff t models for Total, with diff t brunch number at A, and calculated with Independent Bayes model

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Brunch | mean | sd | 2.5% | 50% | 97.5% | Rhat | n.eff |
| 1 | 21.1342 | 4.6356 | 13.0372 | 20.8723 | 31.0156 | 1 | 2899 |
| 2 | 21.4399 | 4.6658 | 13.3765 | 21.0743 | 31.3564 | 1 | 2888 |
| 3 | 30.4796 | 5.5918 | 20.4537 | 30.1087 | 42.6573 | 1 | 3069 |
| 4 | 13.0428 | 3.6218 | 6.9370 | 12.7050 | 21.1728 | 1 | 2798 |
| 5 | 26.4792 | 5.0783 | 17.5450 | 26.1769 | 37.5123 | 1 | 2903 |
| 6 | 22.5513 | 4.7242 | 14.3816 | 22.2968 | 32.5958 | 1 | 3000 |
| 7 | 17.2175 | 4.1172 | 9.9473 | 16.9176 | 25.9870 | 1 | 3000 |
| 8 | 19.2860 | 4.3666 | 11.7869 | 19.0001 | 28.9699 | 1 | 2712 |
| 9 | 27.4363 | 5.2930 | 18.1801 | 27.1123 | 38.8600 | 1 | 2719 |
| 10 | 18.3153 | 4.2436 | 10.9404 | 18.0915 | 27.1782 | 1 | 2961 |
| 11 | 22.2699 | 4.7638 | 13.7242 | 21.8257 | 32.6838 | 1 | 3000 |
| 12 | 23.3112 | 4.9101 | 14.5832 | 22.9914 | 33.6660 | 1 | 3006 |

Table 2.14: Posterior distributions of diff t models for Total, with diff t brunch number at A, and calculated with Hierarchical Bayes model

Table [2.13](#_bookmark83) and table [2.14](#_bookmark84) gives the posterior distribution of the category total, for scenarios with diff t hierarchical structure. Our results show a large variability between posterior distribution for category total at the top of the structure hierarchy, but there are no obvious trends for diff t hierarchical structures and IBM and HBM models in any of the values due to the presence of noise.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Brunch | mean | sd | 2.5% | 50% | 97.5% | Rhat | n.eff |
| 1 | 10.2222 | 3.2257 | 4.9491 | 9.9313 | 17.3630 | 1 | 3000 |
| 2 | 7.2101 | 2.5820 | 3.0653 | 6.9089 | 13.0457 | 1 | 3135 |
| 3 | 13.3965 | 3.5903 | 7.1940 | 13.0422 | 21.4547 | 1 | 2843 |
| 4 | 5.2344 | 2.2966 | 1.7544 | 4.9212 | 10.8095 | 1 | 2659 |
| 5 | 12.2653 | 3.5442 | 6.3801 | 11.9468 | 20.0809 | 1 | 2831 |
| 6 | 12.4534 | 3.5617 | 6.4566 | 12.0511 | 20.3078 | 1 | 2826 |
| 7 | 11.3345 | 3.3140 | 5.6904 | 11.0780 | 18.6164 | 1 | 2763 |
| 8 | 11.3301 | 3.4455 | 5.6405 | 10.9504 | 18.8542 | 1 | 3000 |
| 9 | 15.2998 | 3.9278 | 8.5925 | 15.0144 | 24.1089 | 1 | 3000 |
| 10 | 10.2282 | 3.2404 | 4.9712 | 9.8590 | 17.4283 | 1 | 2719 |
| 11 | 12.2384 | 3.5230 | 6.2765 | 11.9505 | 19.9642 | 1 | 3000 |
| 12 | 16.2605 | 4.1472 | 9.1019 | 15.9211 | 25.5411 | 1 | 2614 |

Table 2.15: Posterior distributions of diff t models for A , with diff t brunch number at A, and calculated with independent Bayes model

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Brunch | mean | sd | 2.5% | 50% | 97.5% | Rhat | n.eff |
| 1 | 10.3658 | 3.2745 | 4.9563 | 9.9946 | 17.5591 | 1 | 2480 |
| 2 | 7.1699 | 2.7092 | 2.8318 | 6.7969 | 13.3853 | 1 | 2763 |
| 3 | 13.5106 | 3.6263 | 7.2349 | 13.2756 | 21.5433 | 1 | 2721 |
| 4 | 4.9821 | 2.1601 | 1.7617 | 4.6571 | 9.9815 | 1 | 2848 |
| 5 | 12.2966 | 3.5260 | 6.2470 | 11.9827 | 20.1420 | 1 | 2691 |
| 6 | 11.9993 | 3.3669 | 6.1917 | 11.7362 | 19.3448 | 1 | 2515 |
| 7 | 11.1468 | 3.3194 | 5.5362 | 10.8140 | 18.2800 | 1 | 2791 |
| 8 | 10.9205 | 3.2059 | 5.5478 | 10.6286 | 17.9246 | 1 | 2596 |
| 9 | 15.1743 | 3.7615 | 8.6417 | 14.8581 | 23.3316 | 1 | 2371 |
| 10 | 9.4961 | 2.9241 | 4.6206 | 9.1667 | 16.0461 | 1 | 2730 |
| 11 | 11.7279 | 3.3009 | 5.9251 | 11.4201 | 18.8472 | 1 | 2389 |
| 12 | 15.4642 | 3.7186 | 9.0081 | 15.1212 | 23.5432 | 1 | 2361 |

Table 2.16: Posterior distributions of diff t modelsfor A , with diff t brunch number at A, and calculated with Hierarchical Bayes model

Table [2.13](#_bookmark83) and table [2.14](#_bookmark84) gives the posterior distribution of the category A, for scenarios with diff t hierarchical structure. Our results show a large variability between posterior distribution for category total at the top of the structure hierarchy, but there are no obvious trends for diff t hierarchical structures and IBM and HBM models in any of the values due to the presence of noise.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Brunch | mean | sd | 2.5% | 50% | 97.5% | Rhat | n.eff |
| 1 | 10.3390 | 3.3313 | 4.9841 | 9.9711 | 18.0423 | 1 | 3015 |
| 2 | 2.2864 | 1.4796 | 0.4436 | 1.9694 | 5.9125 | 1 | 2383 |
| 3 | 4.3819 | 2.1286 | 1.2499 | 4.0705 | 9.5828 | 1 | 2021 |
| 4 | 2.2971 | 1.5500 | 0.3511 | 1.9390 | 6.0469 | 1 | 1776 |
| 5 | 4.3729 | 2.0611 | 1.2799 | 4.0873 | 9.4255 | 1 | 1568 |
| 6 | 5.1595 | 2.0347 | 1.8054 | 4.9425 | 9.5605 | 1 | 1448 |
| 7 | 3.2759 | 1.6852 | 0.7437 | 3.0322 | 7.3844 | 1 | 1259 |
| 8 | 3.2901 | 1.6666 | 0.7674 | 3.0726 | 7.0195 | 1 | 1255 |
| 9 | 3.1983 | 1.5386 | 0.8286 | 3.0263 | 6.7680 | 1 | 1424 |
| 10 | 4.3662 | 1.5571 | 1.6687 | 4.1853 | 7.6161 | 1 | 1110 |
| 11 | 2.1852 | 1.3714 | 0.2855 | 1.9361 | 5.4025 | 1 | 1077 |
| 12 | 2.9119 | 1.4034 | 0.7174 | 2.6996 | 6.1571 | 1 | 947 |

Table 2.17: Posterior distributions of diff ent models for AA, with diff t brunch number at A, and calculated with independent Bayes model

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Brunch | mean | sd | 2.5% | 50% | 97.5% | Rhat | n.eff |
| 1 | 10.3563 | 3.2788 | 4.9509 | 10.0156 | 17.7970 | 1 | 2540 |
| 2 | 2.1951 | 1.5255 | 0.3203 | 1.8362 | 5.9132 | 1 | 1978 |
| 3 | 4.5592 | 2.2061 | 1.2605 | 4.2379 | 9.5741 | 1 | 1910 |
| 4 | 2.2663 | 1.5396 | 0.3203 | 1.9350 | 5.9029 | 1 | 1540 |
| 5 | 4.4613 | 1.9943 | 1.3097 | 4.2350 | 8.8343 | 1 | 1354 |
| 6 | 5.2129 | 2.0225 | 1.8571 | 5.0210 | 9.6588 | 1 | 1023 |
| 7 | 3.3193 | 1.6606 | 0.7695 | 3.0899 | 7.1315 | 1 | 1053 |
| 8 | 3.3147 | 1.6493 | 0.7960 | 3.0817 | 7.1798 | 1 | 1023 |
| 9 | 3.2724 | 1.6032 | 0.7633 | 3.0402 | 6.8838 | 1 | 711 |
| 10 | 4.4763 | 1.6416 | 1.7625 | 4.3287 | 8.0546 | 1 | 737 |
| 11 | 2.2327 | 1.2427 | 0.3545 | 2.0587 | 4.9807 | 1 | 952 |
| 12 | 3.0880 | 1.4240 | 0.8039 | 2.9188 | 6.2747 | 1 | 816 |

Table 2.18: Posterior distributions of diff t modelsfor AA, with diff t brunch number at A, and calculated with Hierarchiacl Bayes model

Table [2.17](#_bookmark87) and table [2.18](#_bookmark88) gives the posterior distribution of category A, for scenarios with a diff t hierarchical structure. Our results show a large variability between posterior distribution for category total at the top of the structure hierarchy, unlike what we observe in previous tables, the trend in the number of brunch and diff between IBM and HBM is now much more apparent. Increasing the size of brunch at category A decreases mean standard deviation and ESS. Also, posterior

mean results for HBM is slightly larger compared to mean results for IBM.

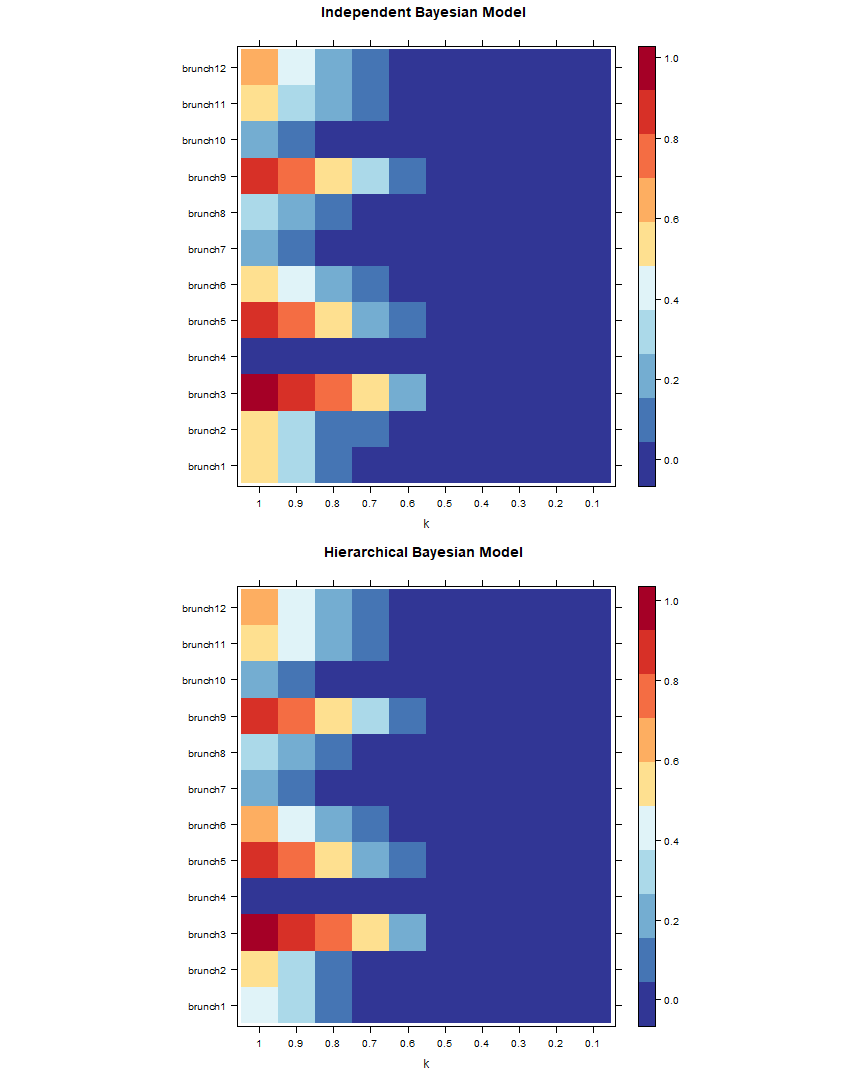


Figure 2.21: Heat-map of anomaly signals for diff t diff t hierarchical structure, of IBM and HBM at category total.

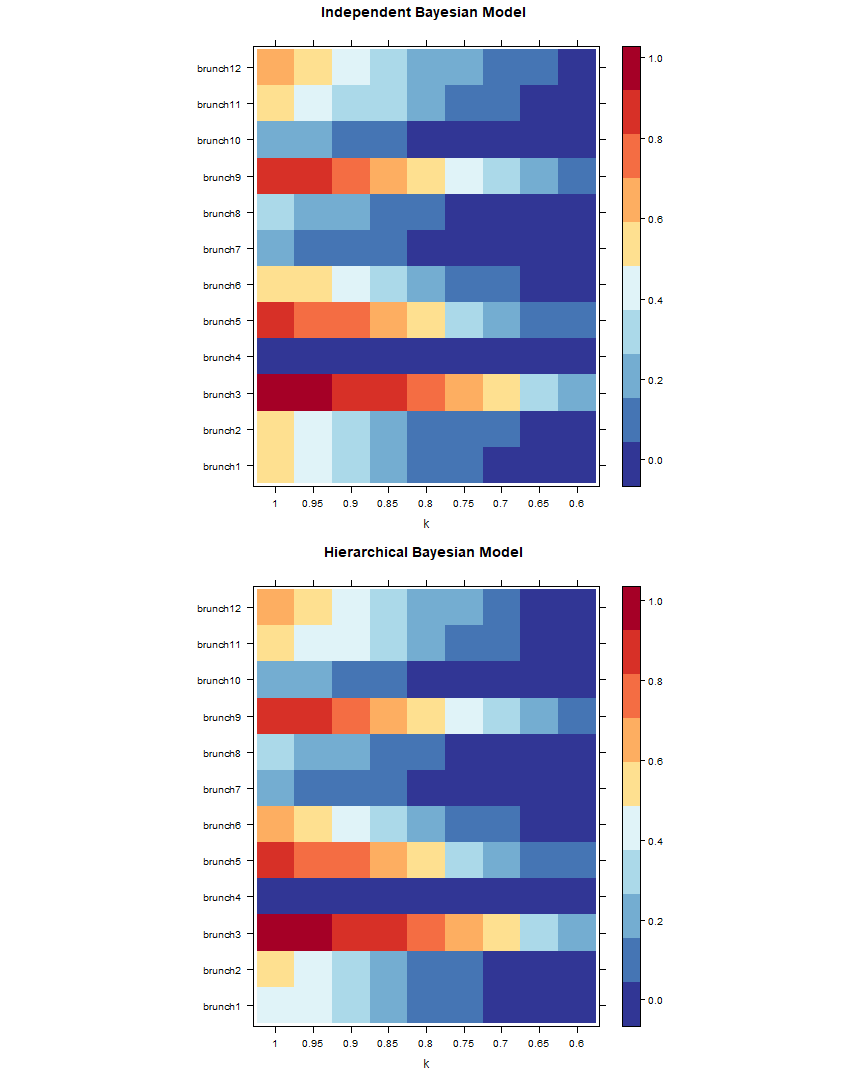


Figure 2.22: Heat-map of anomaly signals for diff t diff t hierarchical structure, of IBM and HBM zoomed in at k values of 0.6 to 1, at category total.

The heat-map in fi [2.21](#_bookmark89) presents the anomaly signals of category total for diff t hierarchical structures with the addition of a constant anomaly, with IBM and HBM, against diff t increments of *k*. There is no apparent trend in anomaly signals terms of the number of brunch at category AA, of category total. However, it is relatively clear that stronger signals of anomaly seem to locate at left sider of the heat-map, indicating that higher the *k*, stronger the signal. Zooming in at looking at k values from 0.6 to 1 in fi [2.22,](#_bookmark90) we observe a noticeable small diff in anomaly detection rate between IBM and HBM for category total. Our observation suggests that the diff between anomaly detection rate is small or close to none for IBM and HBM, on the top level of the hierarchy, for anomaly with a diff t hierarchical structure, in a simulated setting with each category generated independently.

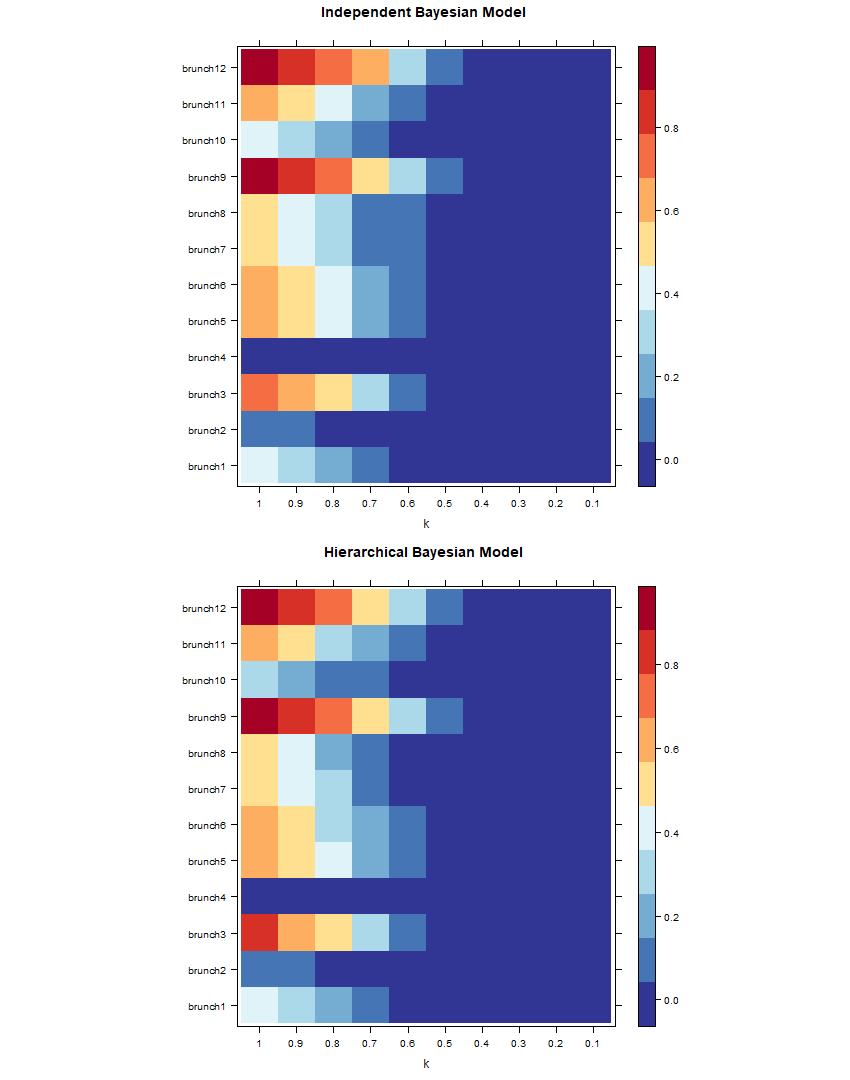


Figure 2.23: Heat-map of anomaly signals for diff t diff t hierarchical structure, of IBM and HBM, at category A.

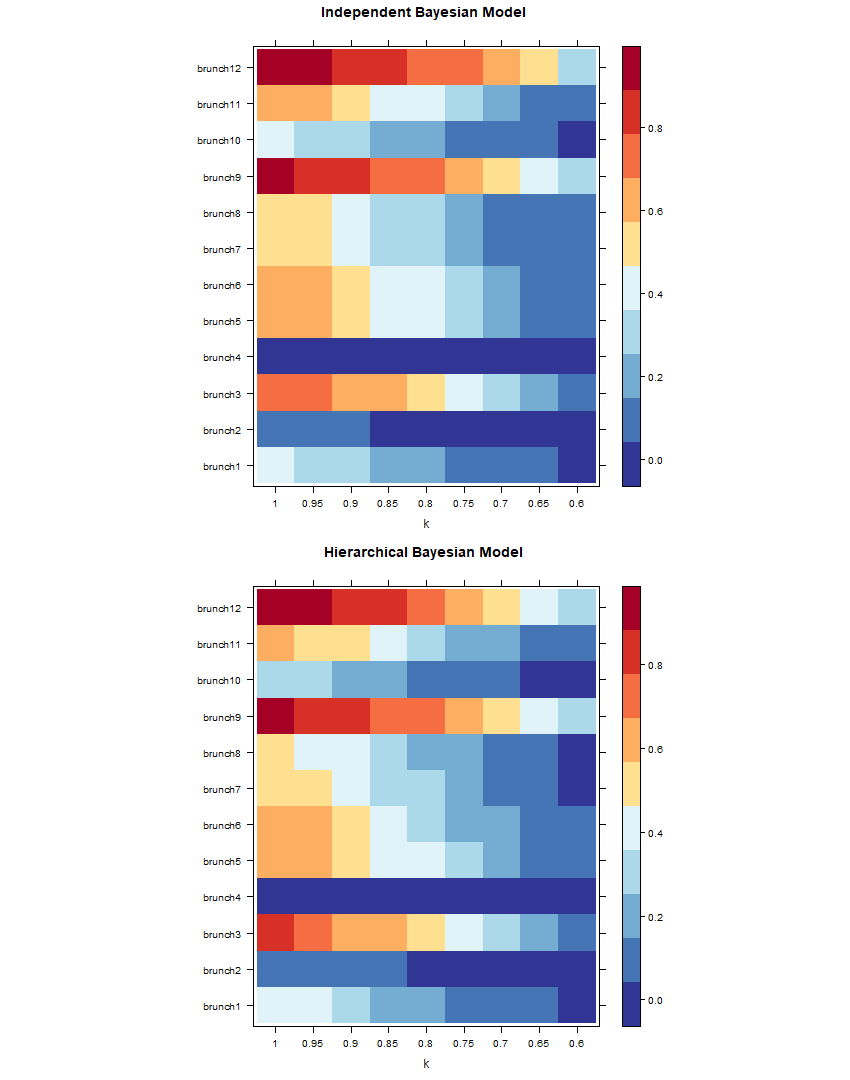


Figure 2.24: Heat-map of anomaly signals for diff t diff t hierarchical structure, of IBM and HBM zoomed in at k values of 0.6 to 1, at category A.

The heat-map in fi re [2.23](#_bookmark91) presents the anomaly signals of category A for diff t hierarchical structures with the addition of a constant anomaly, with IBM and HBM, against diff t increments of *k*. There is no apparent trend in anomaly signals terms of the number of brunch at category AA, of category A. However, it is relatively clear that stronger signals of anomaly seem to locate at left sider of the heat-map, indicating that higher the *k*, stronger the signal. Zooming in at looking at k values from 0.6 to 1 in fi [2.24,](#_bookmark92) we observe a noticeable small diff in anomaly detection rate between IBM and HBM for category A. Our observation suggests that the diff e between anomaly detection rate is small or close to none for IBM and HBM, on level 2 of the hierarchy, for anomaly with a diff t hierarchical structure, in a simulated setting with each category generated independently.

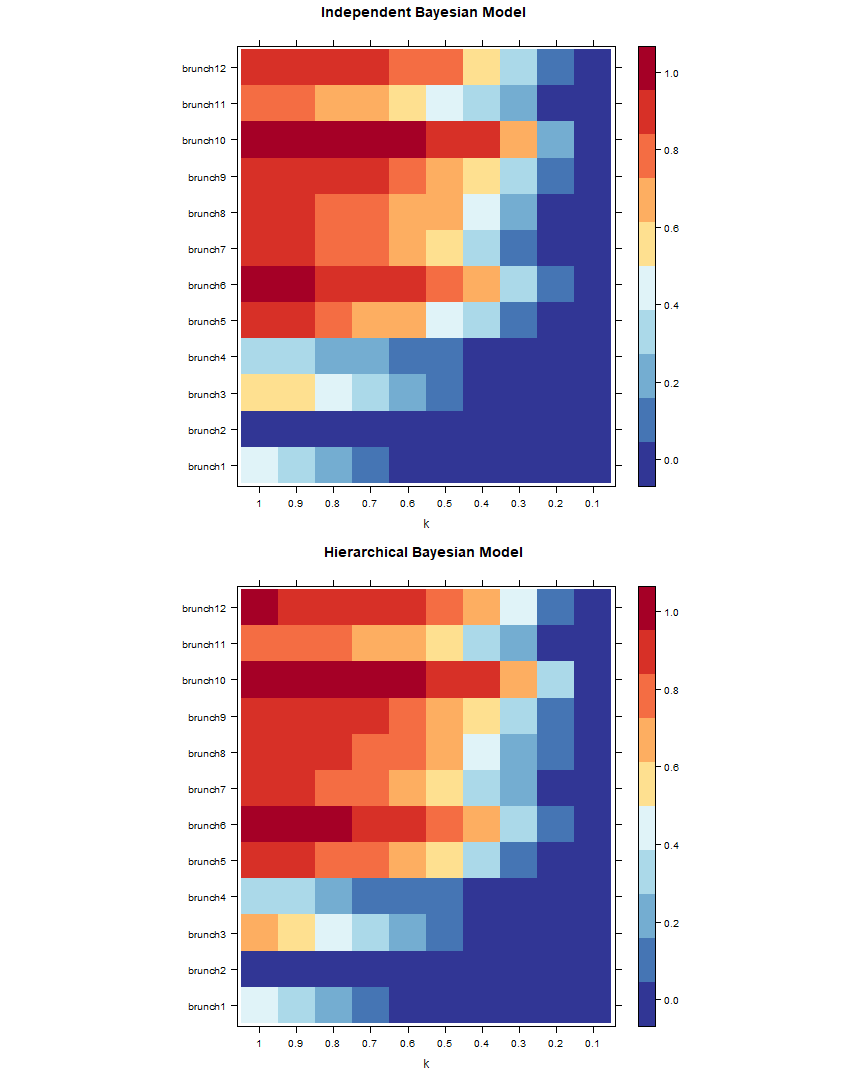


Figure 2.25: Heat-map of anomaly signals for diff t diff t hierarchical structure, of IBM and HBM, at category AA.

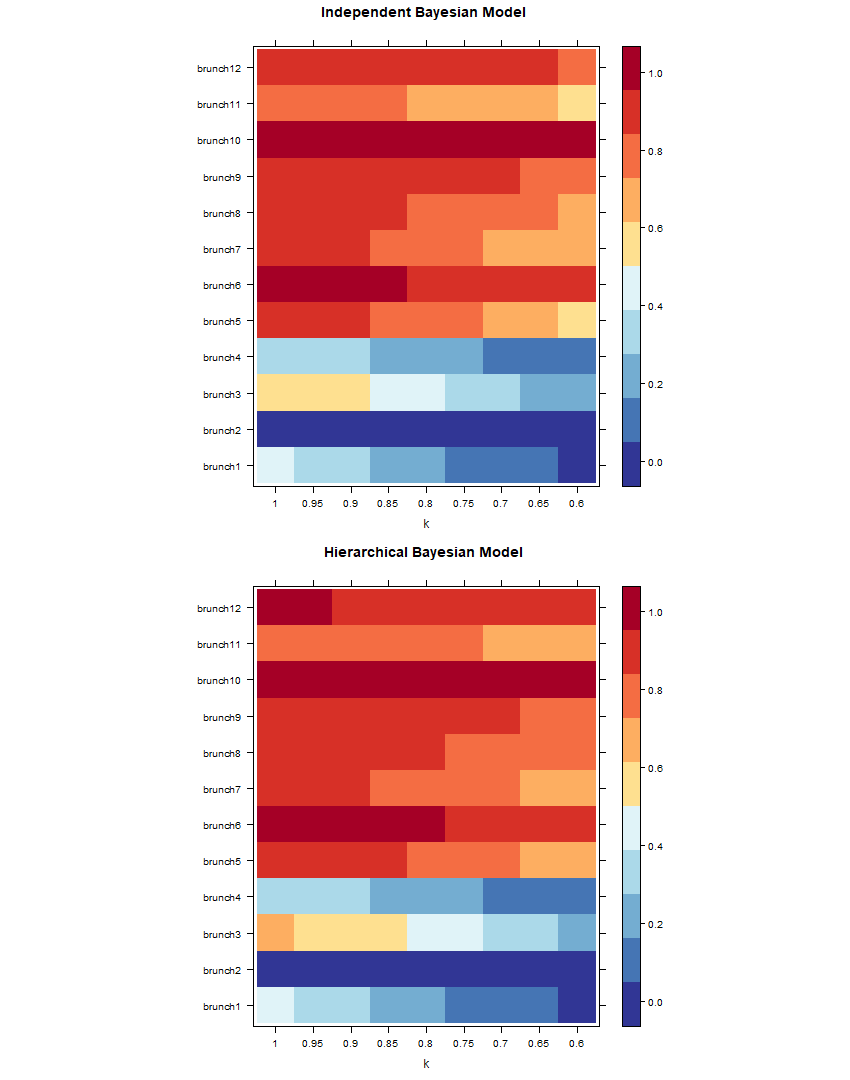


Figure 2.26: Heat-map of anomaly signals for diff t diff t hierarchical structure, of IBM and HBM zoomed in at k values of 0.6 to 1, at category AA.

The heat-map in fi ure [2.25](#_bookmark93) presents the anomaly signals of category A for diff t hierarchical structures with the addition of a constant anomaly, with IBM and HBM, against diff t increments of *k*. Unlike previous categories, anomaly signals tend to get stronger as the number of brunching increases. It is also very clear that stronger signals of anomaly seem to locate at left side of the heat-map, indicating that higher the *k*, stronger the signal. Zooming in at looking at k values from 0.6 to 1 in figure [2.26,](#_bookmark94) we observe a noticeable small diff ence in anomaly detection rate between IBM and HBM for category AA. Our observation suggests that the diff between anomaly detection rate is small or close to none for IBM and HBM, on the bottom of the hierarchy, for anomaly with a diff t hierarchical structure, in a simulated setting with each category generated independently.

Overall we see that there are slight diff in signals fro IBM and HBM results, and if we zoom in and look at a smaller interval of *k*, the diff are more apparent. HBM tend to give strong signals than IBM. Because we simulated datasets used for each scenario independently, there is a lot of uncertainty and noise and made it hard to distinguish diff and trends. However, at the bottom level, where the diff in structure takes place, the effect of noise does not seem to mask the trends any more. Comparing across levels, we see that the top hierarchy have the lowest signal overall, and the bottom hierarchy have the strongest signal overall.

## Discussion

Our DIC results yield similar results as what we have in the previous simulation, that HBM gives better goodness of fi and is the better model, for scenarios with diff t hierarchical structures. Our fi could suggest that HBM gives better goodness of fi consistently, for hierarchies with diff nt hierarchical structures. These fi

add to the reliability of HBM and suggesting it to be a versatile model that is superior to IBM in diff t settings and for diff t hierarchical structures. Because we had synthesised datasets used for our scenarios independently, we have a higher level of uncertainty compare to our previous simulation. It adds to the uncertainty we already had from Bayesian estimations and made it even harder to explore for diff

and trends. A possible remedy is to use the same dataset, but for brunching that take place in category AA, we could try to generate values for its children categories randomly, with statistical techniques such as bootstrapping.

An interesting observation is that trends and diff in category AA seem to have overcome the making effect by the noise and are a lot more clear than what we observe in higher categories. It is partly due to the fact that HBM tend to retain more information ( there is more information at a lower level of hierarchy) at lower

levels of hierarchy and the amount of information is significant enough for HBM to produce results that have a stronger presence than randomness. Another possibility is that it has something to do with infl of extra information that we created for Category A and it’s children during our data synthesis process, which is retained by HBM but not at all present in IBM.

**Chapter 3 MIMIC-III**

# Introduction

We carried-out simulations with synthetic patient arrival datasets, and compares the posterior distribution results and prediction rate between Independent Bayes Model (IBM) and Hierarchical Bayes Model (HBM) for-real life data, for the scenarios of common, rare and extremely rare diseases. Sections 3.2 to 3.8 provides background information about the MIMIC-III database. Sections 3.9 gives the result of our analyses. Lastly, Section 3.10 forms our discussion.

# Source of Data

For this thesis, datasets from the Medical Information Mart for Intensive Care version 1.4 (MIMIC-III v1.4) database, available on PhysioNet Clinical Databases (h[ttps://mimic.physionet.org),](https://mimic.physionet.org/) was used. MIMIC-III is a large, open-access, well maintained database with a collection of de-identified health-related data associated with over 58,000 hospital admissions for 38,645 adults and 7,875 neonates, who was admitted to critical care units of the Beth Israel Deaconess Medical Center (Boston, MA, USA) between the period of June 2001 and October 2012 and is hosted by the Laboratory for Computational Physiology (LCP) at the Massachusetts Institute of Technology (MIT) [(Johnson et al.,](#_bookmark199) [2016;](#_bookmark199) [Pollard and Johnson,](#_bookmark224) [2016;](#_bookmark224) [Goldberger et al.,](#_bookmark189) [2000).](#_bookmark189) MIMIC-III contains clinical, physiological, and mortality data aggregated from ICU information systems, digital hospital archives, bedside monitors and Social Security Administration Death Master Files. MIMIC-III includes health-related information such as demographics, vital sign measurements, laboratory test results, procedures, medications, caregiver notes, imaging reports, and in and out of hospital mortality registry. Data were collected continuously with non-invasive procedures during routine clinical care and data collection and was blinded from patient and

staff, minimising interference with medical routine. Private health information was also de-identified to protect patient privacy [(Mark,](#_bookmark206) [2016).](#_bookmark206) For example, all dates in the database have been shifted to periods between 2100 to 2200 while conserving seasonal trends.

Access and use of the MIMIC-III database for this thesis have been authorised with a restricted-access to PhysioNet clinical databases and bound by a data use agreement, approved on 9th April 2019.

# ICD-9

According to the World Health Organisation [(Organization](#_bookmark217), [2019),](#_bookmark217) International Statistical Classification of Diseases and Related Health Problems (ICD) is the international standard for clinical diagnosis. ICD defi diseases, disorders, injuries, and other related health conditions in a hierarchical fashion, this allows for storage and access of health-related information internationally, in other words, ICD is the standard for clinical diagnosis. Though there is a subtle variation to the standard as some country adopt diff t versions. For example, Clinical Modifi ation (ICD-9-CM) is an adaption used in assigning diagnostic in the United States [(Centers for Disease](#_bookmark172) [Control and Prevention,](#_bookmark172) [2019),](#_bookmark172) and in New Zealand, the Australian Modifi

(ICD-10-AM) were used [(Ministry of Health,](#_bookmark213) [2019).](#_bookmark213) The version of ICD used for medical diagnosis in the MIMIC dataset is ICD-9, the codes are assigned at the end of the patient’s stay and are used by the hospital to bill for care provided [(Johnson](#_bookmark199) [et al.,](#_bookmark199) [2016).](#_bookmark199)

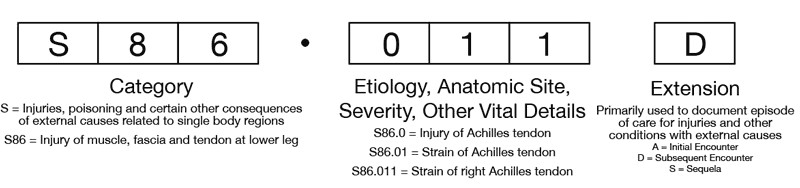


Figure 3.1: Visualisation of ICD-10 Code Structure, from [Andrus and Bohnett](#_bookmark163) [(201](#_bookmark163)3).

As shown in fi [3.1,](#_bookmark100) the fi three characters of an ICD codes designate the category of the diagnosis, followed by supplementary information of aetiology (i.e., the cause of condition), anatomic site, severity, or other vital clinical details after a full stop, and may contain an extension used to document episodes of care, this was implemented for ICD 10. For some instances, there may be multiple codes entries for a single condition, like in the case of an Achilles tendon strain, secondary external cause code may be provided along with the primary diagnosis.

|  |  |  |
| --- | --- | --- |
| **Series ICD-9 Label Description** | | |
| *Total*  1 | - | International Classification of Diseases, 9th Revision |
| *Category lv1* | 001-139 | Infectious And Parasitic Diseases |
| 1 |
| 2 | 140-239 | Neoplasms |
| 3 | 240-279 | Endocrine, Nutritional And Metabolic Diseases, And |
|  |  | Immunity Disorders |
| ... | ... | ... |
| 17 | 800-999 | Injury And Poisoning |
| 18 | V01-V91 | Supplementary Classification Of Factors Influencing |
|  |  | Health Status And Contact With Health Services |
| 19 | E000-E999 | Supplementary Classifi Of External Causes Of |
| Injury And Poisoning |
| *Category lv2* | *(001-999)* | Intestinal Infectious Diseases |
| 1 | 001-009 |
| 2 | 010-018 | Tuberculosis |
| 3 | 020-027 | Zoonotic Bacterial Diseases |
| ... | ... | ... |
| 35 | 980-989 | Toxic Effects Of Substances Chiefly Nonmedicinal As |
|  |  | To Source |
| 36 | 990-995 | Other And Unspecified Effects Of External Causes |
| 37 | 996-999 | Complications Of Surgical And Medical Care, Not |
| Elsewhere Classified |
| *Category lv3* | *(001-999)* | Cholera |
| 1 | 001 |
| 2 | 002 | Typhoid and paratyphoid fevers |
| 3 | 003 | Other salmonella infections |
| ˙.. | ... | ... |
| 89 | 997 | Complications affecting specified body system not |
|  |  | elsewhere classified |
| 89 | 998 | Other complications of procedures not elsewhere |
|  |  | classified |
| 89 | 999 | Complications of medical care not elsewhere classified |

Table 3.1: ICD-9 Diagnosis Codes

ICD 9 categories are coded from 001 to 999, with two extra supplementary class of categories starting with V and E. For example, 001 code for Cholera, and E880 is the code for Accidental fall on or from stairs or steps. Several categories are aggregated to form a larger category, and there are two aggregations which made ICD a three-level hierarchical structure. Following Cholera at level 3, it is aggregated into Intestinal Infectious Diseases for categories from 001 to 009 at level 2, and Infectious and Parasitic Diseases for the categories from 001 to 139 at level 1. Level 1 categories are sometimes referred to as chapters. ICD groupings help to provide a picture of the health situation of the general populations. This information is useful in health care management, and allocation of resources, and other health-related decision-making processes, such as billing of services. Table [3.1](#_bookmark101) provides some examples of ICD codes in diff t level of ICD categories.

# Study Population

Study inclusion criteria were ICU admission of patients of all age with available documentation of primary ICD-9 diagnosis, excluding supplementary diagnoses (ICD codes start with V or E). Repeated admission to ICU by the same patient were included. The fi dataset used contain 50959 uniquely identified Hospital admission event with 38869 uniquely identified patients.

# Study Variables

Three datasets were merged, including the admissions table, the diagnoses icd table and the d icd diagnoses table. Hospital Admission ID and Subject Id were used to as key identifiers. No demographic information was collected, some event-related infor- mation were included, these include Admission time, admission type and admission location. Admission time was converted to admission date. ICD-9 codes used only to contain the fi three digits of the full ICD code, the reason for it is that further digits do not reflect any categorical characteristics and will only increase the number of non-signifi t sub-categories. Supplementary ICD codes that start with V or E were not used, which meant we have ICD codes range from 001 to 999. We defi level 2 and level 3 ICD categories according to the standard ICD-9 defi and created fl my variables for each category. There are 132 unique level 2 ICD categories and 631 unique level 3 ICD categories that were created. Short and long title associated with ICD codes were also included.

# Model

This section presents the Independent Bayesian Model(IBM) and Hierarchical Bayesian Model(HBM) that is used to calculate posterior distributions.

Predictions about the expected daily arrival are used as our prior values, with data for all observed values for the dates prior to the prediction date, using predict.hts Function inside the R environment. Forecast method used is Autoregressive integrated moving average (ARIMA), prediction function used is tbats, and parallel processing is activated to speed up the prediction process. Examples of R-codes used for the time series prediction can be found in Appendix [A.3.](#_bookmark155)

For the IBM model, predictions are made with the bottom-up approach by setting argument method = "bu", the reason for it is that the bottom-up approach makes independent predictions at the leaf level, and does not take into account the hierarchical structure of the data and are suitable for IBM model. In the other hand, predictions for the IBM model are made with the optimal reconciliation approach by setting argument method = "comb", optimal reconciliation approach reconciles prediction from individual time forecasts from all levels and gives an optimum solutions as the forecast, which meant that the approach takes into account the hierarchical structure of the data, and are suitable for HBM model. However, current optimum solutions will give nonsensical negative predictions. However, these negative solutions are mostly extremely small and close to 0 and are therefore simply converted to 0. So the optimal reconciliation forecasts used for our HBM model are not the actual optimum solution, but the adjusted optimal solution. Non-negativity optimal reconciliation for the predict.hts the function is currently being developed and will be released on CRAN in late July, after the submission date of this thesis.

A Normal(1,0.1) distribution is multiplied with the predicted values to add distribution properties to our prior as our weakly-informative prior. For IBM posterior distributions, posteriors of each category are independently calculated with an IBM model. Whereas for HBM posterior distributions, posterior for each category is calculated with an HBM model that considers all of the levels within the hierarchy. JAGS-codes used to build the IBM and HBM are provided in Appendix [A.5](#_bookmark158) and Appendix [A.6.](#_bookmark159)

The likelihood model used for our IBM is:

*yi,t ∼ Poisson*(*µi,t*) *log*(*µi,t*) = *log*(*ρi,tλi,t*)

(3.1)

Priors for model parameters for our IBM is:

*λi,t ∼ Normal*(*µi,t, σi,t*)*T*[*λ>*0] (3.2) Hyper-priors for model parameters for our IBM is:

*µi,t ∼ Normal*(1*,* 0*.*1)

*i,t ∼ Normal*(0*.*1*,* 0*.*1) The likelihood model used for our HBM is also:

*σ*2

*yi,t ∼ Poisson*(*µi,t*) *log*(*µi,t*) = *log*(*ρi,tλi,t*)

(3.3)

(3.4)

But *λi,t* used for diff t levels of the hierarchy are interconnected with priors and hyper-priors and hyper-hyper priors and so on, that captures the hierarchical structure of our ICD-9 categories.

For level 0 category(total):

*λ*(*lv*0)*i,t ∼ Normal*(*µ*(*lv*0)*i,t, σ*(*lv*0)*i,t*)*T*[*λ>*0]

*µ*(*lv*0)*i,t ∼ Normal*(1*,* 0*.*1)

(*lv*0)*i,t ∼ Normal*(0*.*1*,* 0*.*1) For level 1 categories:

*σ*2

*λ*(*lv*1)*i,t ∼ Normal*(*µ*(*lv*1)*i,t, σ*(*lv*1)*i,t*)*T*[*λ>*0]

*µ*(*lv*1)*i,t ∼ Normal*(*mu*(*lv*0)*i,t,* 0*.*1)

(*lv*1)*i,t ∼ Normal*(0*.*1*,* 0*.*1) For level 2 categories:

*σ*2

*λ*(*lv*2)*i,t ∼ Normal*(*µ*(*lv*2)*i,t, σ*(*lv*2)*i,t*)*T*[*λ>*0]

*µ*(*lv*2)*i,t ∼ Normal*(*µ*(*lv*1)*i,t,* 0*.*1)

(*lv*2)*i,t ∼ Normal*(0*.*1*,* 0*.*1) For level 3 (leaf-level) categories:

*σ*2

*λ*(*lv*3)*i,t ∼ Normal*(*µ*(*lv*3)*i,t, σ*(*lv*3)*i,t*)*T*[*λ>*0]

*µ*(*lv*3)*i,t ∼ Normal*(*µ*(*lv*2)*i,t,* 0*.*1)

(*lv*3)*i,t ∼ Normal*(0*.*1*,* 0*.*1)

*σ*2

(3.5)

(3.6)

(3.7)

(3.8)

# Anomaly

The date of 10 November 2190 is set up to be the default date for anomalies additions, therefore, for yearly data anomaly is added on the year of 2190, for monthly data the anomaly is added on 2090Nov, and so on. Due to limitations in sample size and complexity of the ICD-9 hierarchy, it is decided that the time period that we would use is yearly periods between years of 2100 to 2290, the mimic-3 dataset provides an average of 500 arrivals per year.

We are interested in how the anomaly detection rates of Independent Bayesian Model (IBM) and Hierarchical Bayesian Model(HBM) will perform for arrivals of common, rare and extremely rare disease diagnoses. **Common** disease refers to a disease that public hospitals will likely to always encounter in a time period, in our case we decided to choose category 410, which codes for Acute myocardial infarction, commonly known as a heart attack. There are 3307 cases of Acute myocardial infarction in the MIMIC-III data, making it one of the largest category. Considering the pre-perturbation time period used in MIMIC-III is actually a 12 year period. We can suggest that Acute myocardial infarction as a common disease that the ICU department almost sees daily. **Rare** disease refers to a disease that public hospitals will unlikely to encounter in a time period, in our case we decided to choose category 415, which codes for Acute pulmonary heart disease, also known as cor pulmonale, is the enlargement and failure of the right ventricle of the heart as a response to increased vascular resistance (such as from pulmonic stenosis) or high blood pressure in the lungs.(Medlin[ePlus Medical Encyclopedia,](#_bookmark208) [2019).](#_bookmark208) There are 373 cases of Acute pulmonary heart disease in the MIMIC-III data, considering the pre-perturbation time period of MIMIC-III is actually a 12 year period, we can suggest that Acute myocardial infarction as a rare disease that the ICU department almost sees monthly. **Extremely rare** disease refers to a disease that public hospitals will extremely unlikely to encounter in a time period, in our case we decided to choose category 452, which codes for Portal vein thrombosis, it is blockage or narrowing of the portal vein by a blood clot (P[onziani et al.,](#_bookmark225) [2010).](#_bookmark225) There are only 12 cases of Acute pulmonary heart disease in the MIMIC-III data, considering the pre-perturbation time period of MIMIC-III is actually a 12 year period, we can suggest that Acute myocardial infarction as an extremely rare disease that the ICU department are lucky to see more than once per year.

Three abnormal yearly observations are created to test anomaly detection of the three scenarios when there is defi anomaly (we added anomaly) present at the lowest level of the hierarchy. Due to the aggregation constraint, the addition of anomaly at the lowest level of the hierarchy will result in the addition of the same

number of anomalies to its parent and ancestors. This has an effect on the detection rate in all levels of the hierarchy that relates to the specific category. Figure **??** shows how the addition of an anomaly of 100% of the observation at the lowest level does to all related categories if the observation is 0, a default anomaly of 1 is added.

Overall we are looking at four scenarios. Scenario 1 is the anomaly detection rate of common, rare and extremely rare disease categories of the observed data, without the addition of an anomaly. Scenario 2 is the anomaly detection rate of common disease categories of the observed data, with an addition 100% anomaly at the lowest hierarchy level. Scenario 3 is the anomaly detection rate of rare disease categories of the observed data, with an addition 100% anomaly at the lowest hierarchy level. Lastly, in scenario 4, we look at the anomaly detection rate of extremely rare disease categories of the observed data, with an addition 100% anomaly at the lowest hierarchy level.

# Statistical Analyses

A DIC comparisons table was fi used to compare the diff in DIC distributions of Independent Bayesian Model (IBM) and Hierarchical Bayesian Model (HBM). Posterior distributions of our categories of interest for original and anomaly data are then provided in tables and density plots. Lastly, signals of the anomaly are visualised with a heat-map. SAS version 9.4 [(SAS Institute Inc,](#_bookmark230) [2013)](#_bookmark230) was used during the data cleaning process, and analyses were performed with R version 3.3.2 (R [Core](#_bookmark226) [Team,](#_bookmark226) [2013)](#_bookmark226) in the R Studio environment Version 1.1.463 [(RStudio Team,](#_bookmark228) [2015).](#_bookmark228) An overview of functions created and models built in this thesis is provided in the Appendix. The full collection of R, JAGS, and SAS codes used to synthesise datasets, clean and merge datasets, run the simulation, and draw table and plots and are available on GitHub repository [https://github.com/jungxue/research-masters-Jung.](https://github.com/jungxue/research-masters-Jung)

# Results

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA’s |
| No amomaly | -5.635 | -0.443 | -0.004 | 0.095 | 0.494 | 9.212 | 109 |
| 410 | -5.579 | -0.416 | 0.005 | 0.118 | 0.547 | 6.933 | 109 |
| 415 | -7.660 | -0.438 | -0.010 | -0.007 | 0.444 | 8.195 | 109 |
| 452 | -6.540 | -0.466 | -0.002 | 0.048 | 0.496 | 7.628 | 109 |

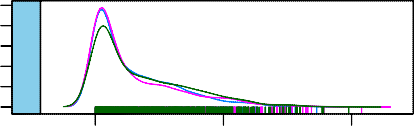
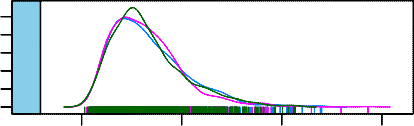
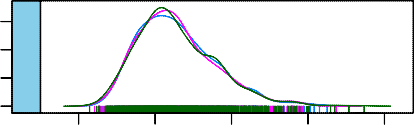
Table 3.2: DIC comparisons for Independent Bayesian Model and Hierarchical Bayesian model for no anomaly, and anomaly added for common, rare and extremely rare diseases

Table [3.2](#_bookmark108) gives the comparison of DIC distributions of Independent Bayesian Model (IBM) and Hierarchical Bayesian Model (HBM) for no anomaly, and anomaly added for common, rare and extremely rare diseases. It is tough to tell whether or not HBM perform better than IBM, because the negative signal is very hard to detect, and the mean and median value is very close to 0. Multiple samples of DIC are also not feasible as sampling process took a very long time. Conclusions about goodness-of-fi performance between IBM and HBM can not be made.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | sd | 2.5% | 50% | 97.5% | Rhat | n.eff |
| total | 526.647 | 22.818 | 482.137 | 526.444 | 573.014 | 1.000 | 2912.000 |
| 390-459 | 195.270 | 13.662 | 169.633 | 194.978 | 222.341 | 1.000 | 3000.000 |
| 410-414 | 78.372 | 8.798 | 62.315 | 78.048 | 95.974 | 1.000 | 2888.000 |
| 415-417 | 3.240 | 1.815 | 0.715 | 2.917 | 7.654 | 1.000 | 2415.000 |
| 451-459 | 6.412 | 2.555 | 2.477 | 6.048 | 12.171 | 1.000 | 2191.000 |
| 410 | 37.218 | 6.211 | 26.305 | 36.663 | 50.562 | 1.000 | 2879.000 |
| 415 | 3.293 | 1.833 | 0.800 | 2.932 | 7.714 | 1.000 | 1977.000 |
| 452 | 0.158 | 0.174 | 0.001 | 0.087 | 0.597 | 1.010 | 445.000 |

Table 3.3: Posterior distributions for ICD 410,415 and 452 and its ancestors, with no added anomalies, and calculated with Independent Bayes Model

0 5 10 15



0.00 0.10

0.00 0.15

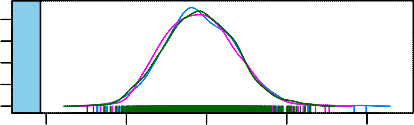
451−459

415

0 1 2 3 4 5

452

0 5 10 15 20

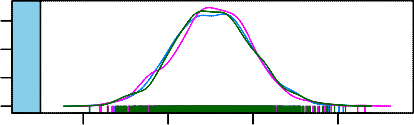


Density

0.00 0.02 0.04

410−414

60 80 100 120



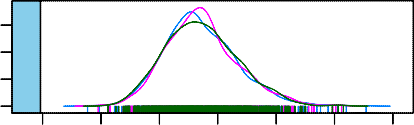
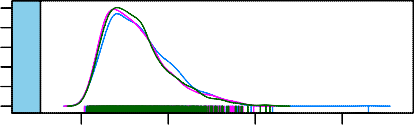
0.000 0.010

total

450 500 550 600

0.0 0.5 1.0

20 30 40 50 60 70



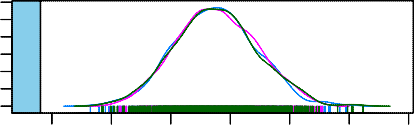
0.000.100.20

0.00 0.04

415−417

410

0 5 10 15



0.000 0.015

390−459

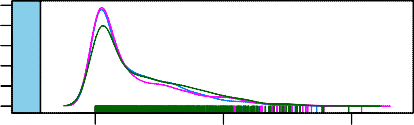
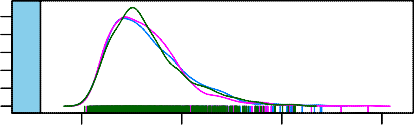
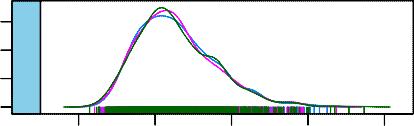
140 160 180 200 220 240

Figure 3.2: Posterior distributions for ICD 410,415 and 452 and its ancestors, with no added anomalies, and calculated with Independent Bayes Model

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | sd | 2.5% | 50% | 97.5% | Rhat | n.eff |
| total | 525.996 | 22.917 | 482.084 | 525.396 | 572.556 | 1.000 | 3000.000 |
| 390-459 | 194.540 | 14.077 | 167.744 | 194.275 | 223.211 | 1.000 | 3276.000 |
| 410-414 | 78.348 | 8.939 | 62.045 | 77.995 | 96.766 | 1.000 | 3160.000 |
| 415-417 | 3.116 | 1.734 | 0.692 | 2.817 | 7.201 | 1.000 | 2174.000 |
| 451-459 | 6.288 | 2.344 | 2.516 | 6.053 | 11.522 | 1.010 | 1401.000 |
| 410 | 37.678 | 6.166 | 26.738 | 37.342 | 50.959 | 1.000 | 3000.000 |
| 415 | 3.319 | 1.949 | 0.703 | 2.954 | 8.441 | 1.000 | 1990.000 |
| 452 | 0.150 | 0.169 | 0.001 | 0.082 | 0.607 | 1.050 | 279.000 |

Table 3.4: Posterior distributions for ICD 410,415 and 452 and its ancestors, with no added anomalies, and calculated with Hierarchical Bayes model

0 5 10 15



0.00 0.10

0.00 0.15

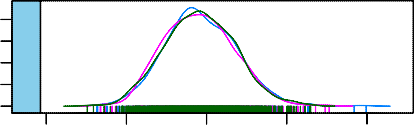
451−459

415

0 1 2 3 4 5

452

0 5 10 15 20

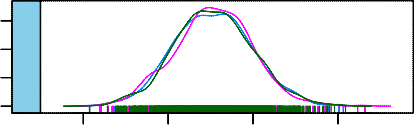


Density

0.00 0.02 0.04

410−414

60 80 100 120



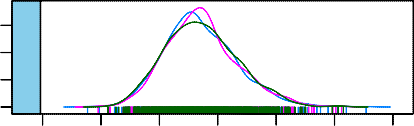
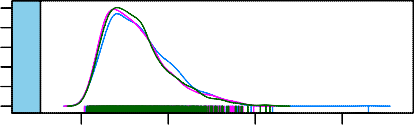
0.000 0.010

total

450 500 550 600

0.0 0.5 1.0

20 30 40 50 60 70



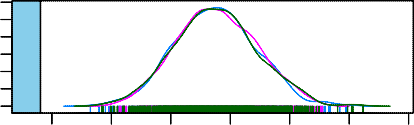
0.000.100.20

0.00 0.04

415−417

410

0 5 10 15



0.000 0.015

390−459

140 160 180 200 220 240

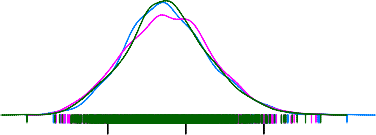
Figure 3.3: Posterior distributions for ICD 410,415 and 452 and its ancestors, with no added anomalies, and calculated with Hierarchical Bayes model

Table [3.3](#_bookmark109) and fi [3.2](#_bookmark110) shows the posterior for IBM and table [3.4](#_bookmark111) and fi [3.3](#_bookmark112)

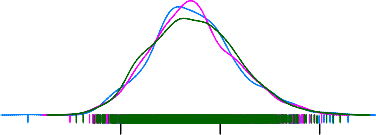
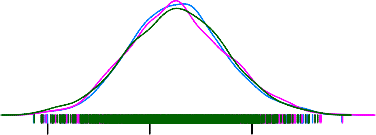
shows the posterior for HBM, for ICD 410,415 and 452 and its ancestors, with no added anomalies. There are no obvious diff between the distribution values and distribution shape for IBM and HBM. Distribution for categories at a higher level, and categories with higher mean values tend to approximate to normal, for the case of category 415-417, even though it is at a higher hierarchy level it has a low mean, and the distribution is not normal.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | sd | 2.5% | 50% | 97.5% | Rhat | n.eff |
| total | 563.697 | 23.816 | 518.412 | 563.222 | 612.356 | 1.000 | 3111.000 |
| 390-459 | 232.494 | 15.041 | 204.686 | 231.814 | 263.502 | 1.000 | 2905.000 |
| 410-414 | 115.402 | 10.918 | 94.819 | 114.956 | 137.937 | 1.000 | 3000.000 |
| 410 | 74.630 | 8.617 | 58.748 | 74.137 | 92.840 | 1.000 | 3542.000 |

Table 3.5: Posterior distributions for ICD 410 and its ancestors, with added anomalies, and calculated with independent Bayes model



80 100 120 140 160



Density

0.00 0.02 0.04

410−414

0.00 0.02 0.04

390−459

410

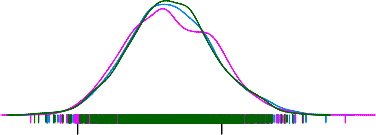
40 60 80 100

500 550 600 650

0.000 0.010

total

200 250 300



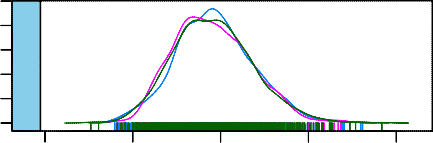
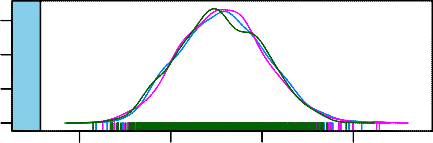
0.000 0.015

Figure 3.4: Posterior distributions for ICD 410 and its ancestors, with added anomalies, and calculated with independent Bayes model

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | sd | 2.5% | 50% | 97.5% | Rhat | n.eff |
| total | 528.324 | 23.296 | 485.619 | 527.677 | 574.402 | 1.000 | 2835.000 |
| 390-459 | 197.765 | 13.976 | 171.717 | 197.386 | 225.945 | 1.000 | 3014.000 |
| 410-414 | 78.351 | 8.886 | 62.382 | 77.809 | 96.776 | 1.000 | 2860.000 |
| 410 | 37.500 | 6.085 | 26.335 | 37.094 | 49.894 | 1.000 | 3138.000 |

Table 3.6: Posterior distributions for ICD 410 and its ancestors, with added anomalies, and calculated with Hierarchiacl Bayes model

60 80 100 120



Density

0.00 0.02 0.04

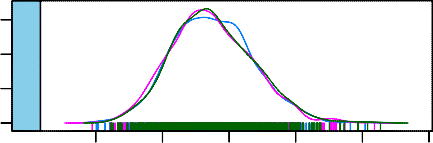
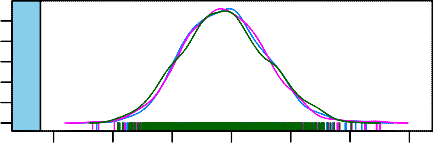
0.000 0.010

total

410−414

450 500 550 600

20 30 40 50 60



0.000 0.015

0.00 0.04

390−459

410

140 160 180 200 220 240 260

Figure 3.5: Posterior distributions for ICD 410 and its ancestors, with added anomalies, and calculated with Hierarchiacl Bayes model

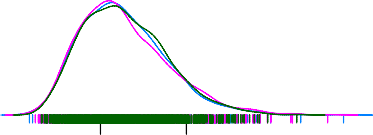
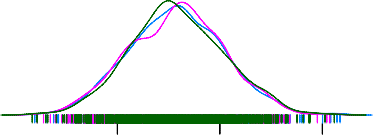
Table [3.5](#_bookmark113) and fi [3.4](#_bookmark114) shows the posterior for IBM and table [3.6](#_bookmark115) and fi [3.5](#_bookmark116)

shows the posterior for HBM, for ICD 410 and its ancestors, with added anomalies. There are no obvious diff ces between the distribution values and distribution shape for IBM and HBM.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | sd | 2.5% | 50% | 97.5% | Rhat | n.eff |
| total | 528.937 | 23.037 | 484.629 | 528.703 | 574.379 | 1.000 | 2859.000 |
| 390-459 | 198.553 | 14.315 | 171.867 | 198.445 | 227.649 | 1.000 | 3000.000 |
| 415-417 | 6.364 | 2.494 | 2.470 | 6.059 | 11.999 | 1.000 | 2419.000 |
| 415 | 6.436 | 2.571 | 2.373 | 6.117 | 12.070 | 1.000 | 2139.000 |

Table 3.7: Posterior distributions for ICD 415 and its ancestors, with added anomalies, and calculated with independent Bayes model

0 5 10 15 20



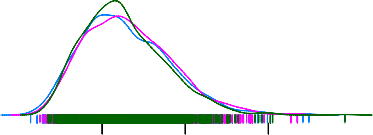
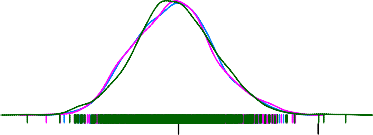
Density

0.00 0.10

total

415−417

0 5 10 15 20



0.00 0.10

390−459

415

450 500 550 600

0.000 0.010

150 200 250

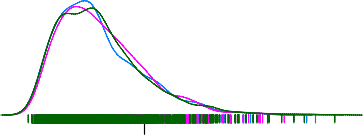
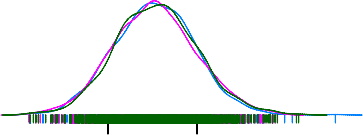
0.000 0.015

Figure 3.6: Posterior distributions for ICD 415 and its ancestors, with added anomalies, and calculated with independent Bayes model

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | sd | 2.5% | 50% | 97.5% | Rhat | n.eff |
| total | 526.519 | 23.086 | 481.720 | 526.254 | 572.876 | 1.000 | 3000.000 |
| 390-459 | 195.702 | 13.884 | 169.817 | 195.471 | 223.270 | 1.000 | 3165.000 |
| 415-417 | 3.146 | 1.784 | 0.721 | 2.827 | 7.459 | 1.000 | 1876.000 |
| 415 | 3.308 | 1.953 | 0.680 | 2.942 | 7.934 | 1.000 | 1730.000 |

Table 3.8: Posterior distributions for ICD 415 and its ancestors, with added anomalies, and calculated with Hierarchical Bayes model

0 5 10



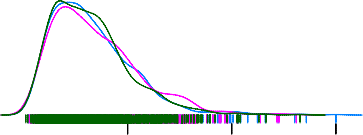
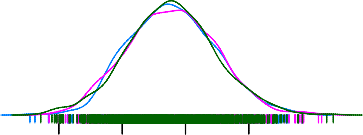
Density

0.00 0.10 0.20

total

415−417

0 5 10 15



0.000 0.015 0.030

0.00 0.10 0.20

390−459

415

450 500 550 600 650

0.000 0.010

140 160 180 200 220 240 260

Figure 3.7: Posterior distributions for ICD 415 and its ancestors, with added anomalies, and calculated with Hierarchical Bayes model

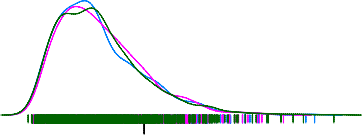
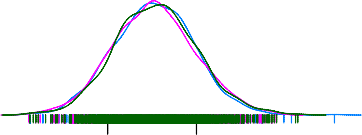
Table [3.7](#_bookmark117) and fi [3.6](#_bookmark118) shows the posterior for IBM and table [3.8](#_bookmark119) and fi [3.7](#_bookmark120)

shows the posterior for HBM, for ICD 415 and its ancestors, with added anomalies. There are no obvious diff ces between the distribution values and distribution shape for IBM and HBM.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | sd | 2.5% | 50% | 97.5% | Rhat | n.eff |
| total | 526.983 | 23.297 | 480.980 | 526.758 | 573.138 | 1.000 | 3114.000 |
| 390-459 | 196.224 | 13.951 | 169.986 | 195.689 | 224.270 | 1.000 | 3294.000 |
| 451-459 | 7.382 | 2.629 | 3.052 | 7.077 | 13.190 | 1.000 | 2322.000 |
| 452 | 0.346 | 0.208 | 0.033 | 0.323 | 0.812 | 1.000 | 517.000 |

Table 3.9: Posterior distributions for ICD 452 and its ancestors, with added anomalies, and calculated with independent Bayes model

0 5 10



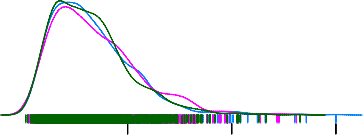
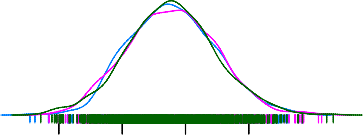
Density

0.00 0.10 0.20

total

415−417

0 5 10 15



0.000 0.015 0.030

0.00 0.10 0.20

390−459

415

450 500 550 600 650

0.000 0.010

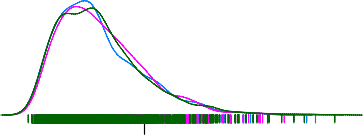
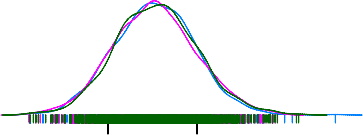
140 160 180 200 220 240 260

Figure 3.8: Posterior distributions for ICD 452 and its ancestors, with added anomalies, and calculated with independent Bayes model

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | sd | 2.5% | 50% | 97.5% | Rhat | n.eff |
| total | 526.447 | 23.045 | 482.679 | 526.250 | 572.946 | 1.000 | 3061.000 |
| 390-459 | 195.096 | 13.618 | 168.448 | 194.678 | 222.373 | 1.000 | 2950.000 |
| 451-459 | 7.476 | 2.495 | 3.299 | 7.282 | 13.055 | 1.000 | 1358.000 |
| 452 | 0.390 | 0.222 | 0.045 | 0.362 | 0.889 | 1.020 | 342.000 |

Table 3.10: Posterior distributions for ICD 452 and its ancestors, with added anomalies, and calculated with Hierarchical Bayes model

0 5 10



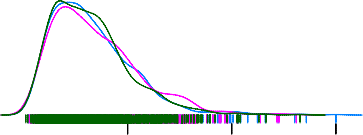
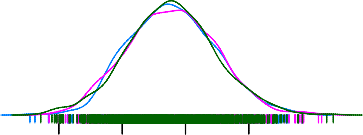
Density

0.00 0.10 0.20

total

415−417

0 5 10 15



0.000 0.015 0.030

0.00 0.10 0.20

390−459

415

450 500 550 600 650

0.000 0.010

140 160 180 200 220 240 260

Figure 3.9: Posterior distributions for ICD 452 and its ancestors, with added anomalies, and calculated with Hierarchical Bayes model

Table [3.9](#_bookmark121) and fi e [3.8](#_bookmark122) shows the posterior for IBM and table [3.10](#_bookmark123) and fi [3.9](#_bookmark124)

shows the posterior for HBM, for ICD 452 and its ancestors, with added anomalies. There are no obvious diff ces between the distribution values and distribution shape for IBM and HBM.

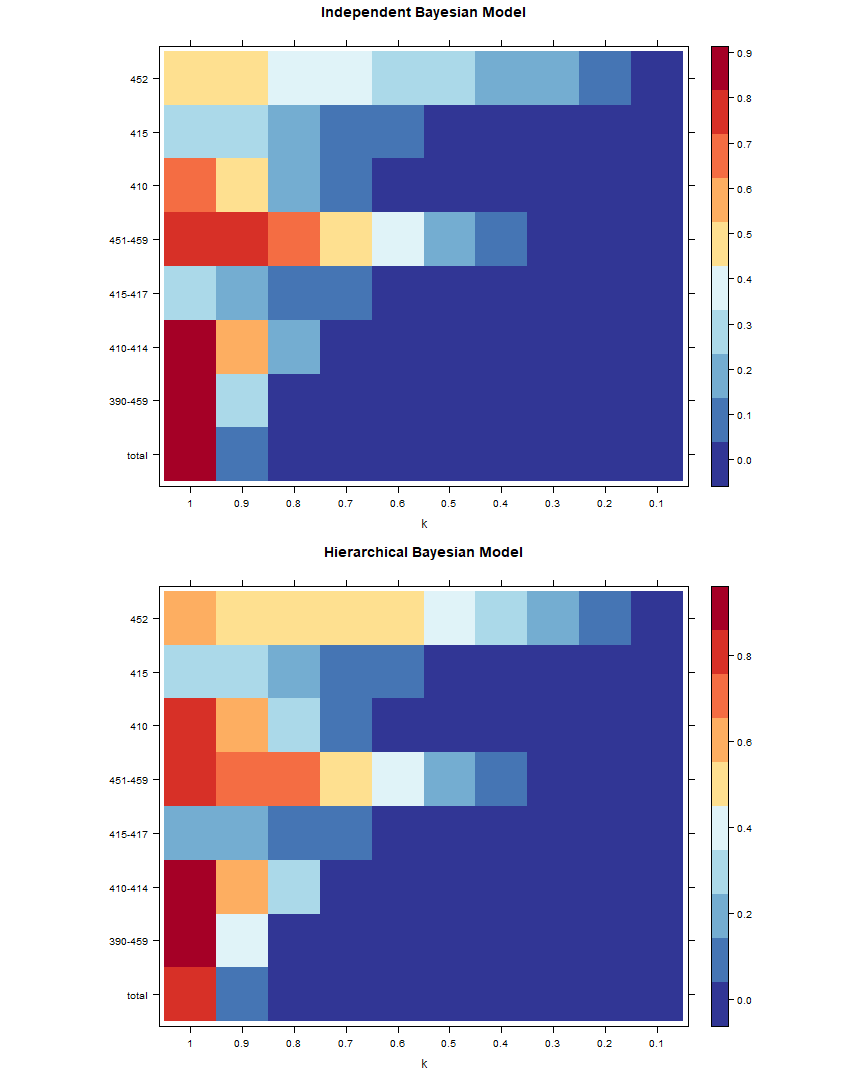


Figure 3.10: Heat-map of anomaly signals for ICD 410,415 and 452 and its ancestors, with no added anomalies, of IBM and HBM.

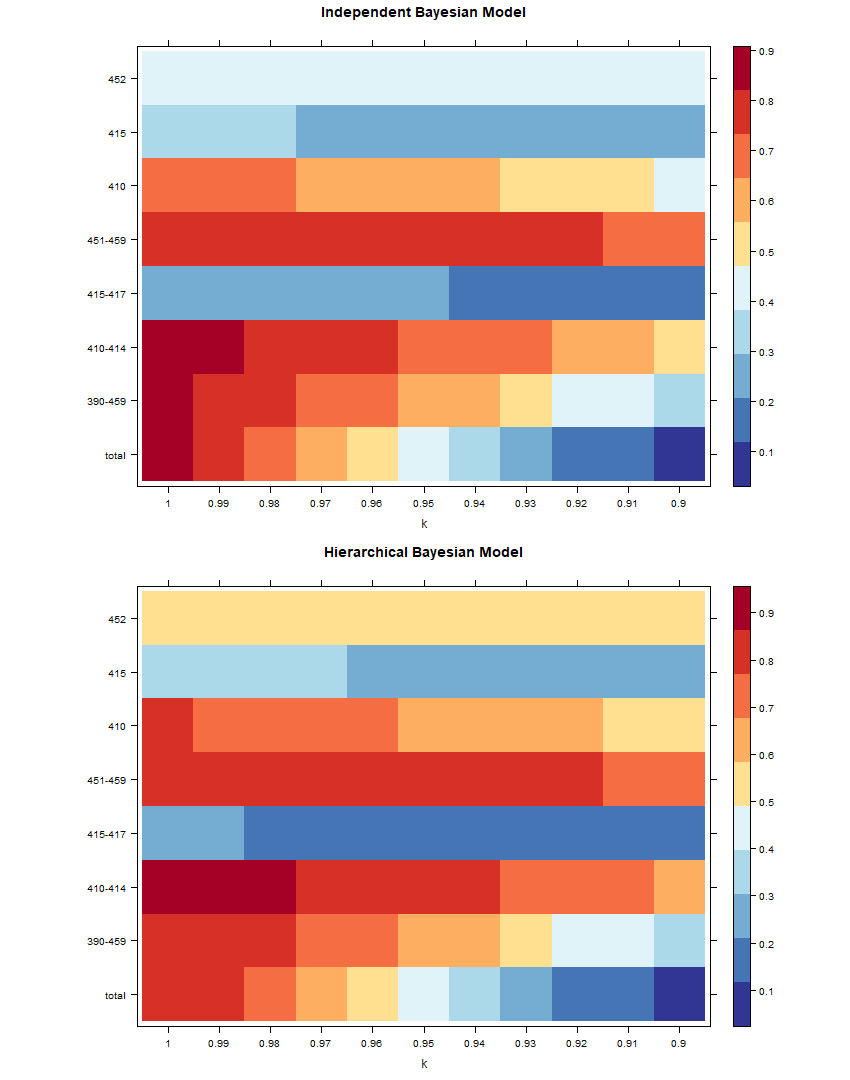


Figure 3.11: Heat-map of anomaly signals for ICD 410,415 and 452 and their ancestors, with no added anomalies, of IBM and HBM zoomed in at k values of 0.9 to 1.

The heat-map in fi [3.10](#_bookmark125) presents the anomaly signals of ICD 410,415 and 452 and their ancestors with no addition of a constant anomaly, with IBM and HBM, against diff t increments of *k*.Anomaly signals tend to be stronger as k increase, and stronger for lower hierarchies compare to higher hierarchies, the variability of signal strength also tend to increase with higher hierarchy. Zooming in at looking at k values from 0.6 to 1 in fi [3.11,](#_bookmark126) we observe a noticeable small diff in anomaly detection rate between IBM and HBM for category AA. Our observation suggests that the diff between anomaly detection rate is small or close to none for IBM and HBM, on the bottom of the hierarchy, for anomaly with a diff t hierarchical structure, in a simulated setting with each category generated independently.

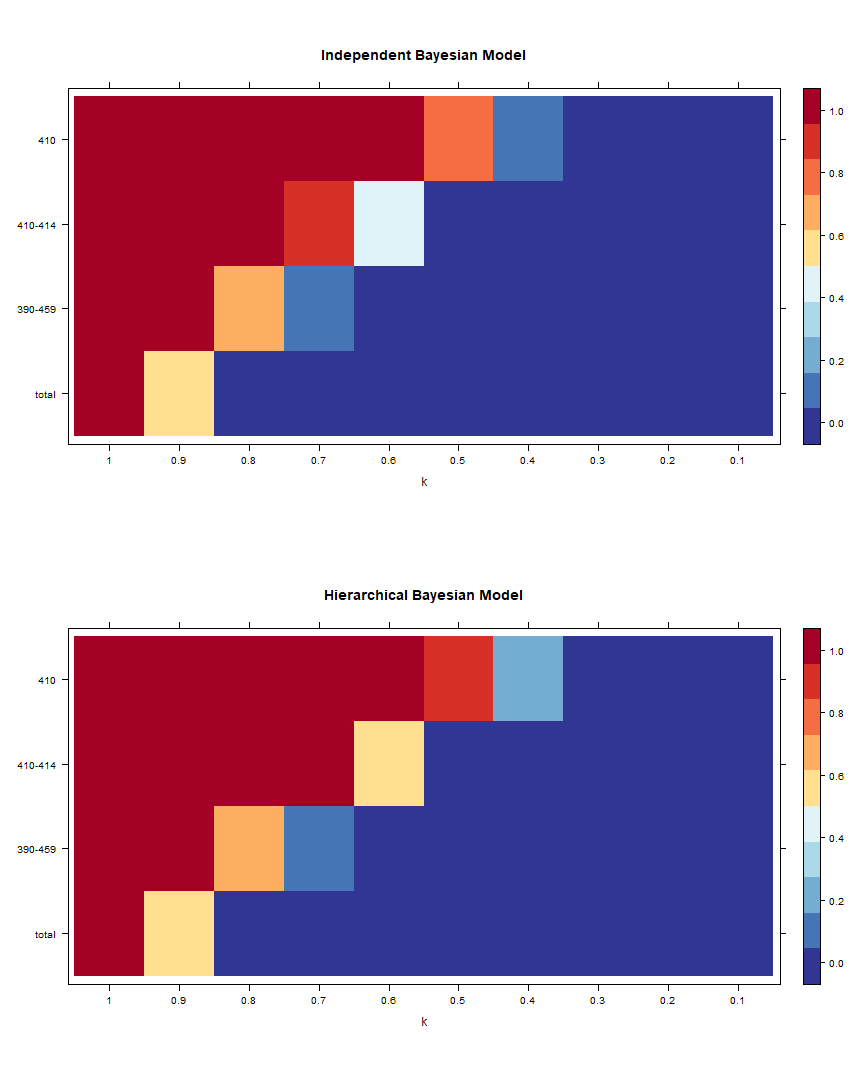


Figure 3.12: Heat-map of anomaly signals for ICD 410 and its ancestors, with no added anomalies, of IBM and HBM.

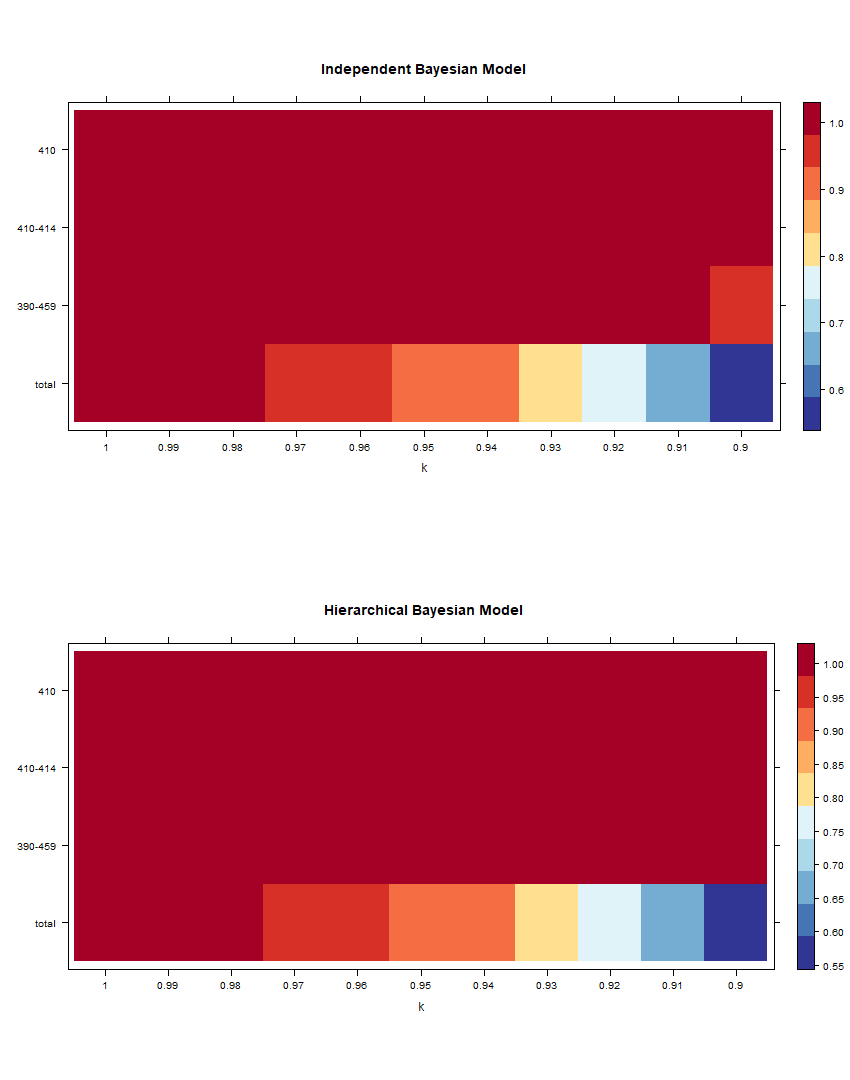


Figure 3.13: Heat-map of anomaly signals for ICD 410 and its ancestors, with no added anomalies, of IBM and HBM zoomed in at k values of 0.9 to 1.

The heat-map in fi [3.12](#_bookmark127) presents the anomaly signals of ICD 410 and their ancestors with no addition of a constant anomaly, with IBM and HBM, against diff t increments of *k*.Anomaly signals tend to be strong on the top left of the heat map. This suggests that signals of anomaly tend to increase in strength with increased k and higher hierarchy. Zooming in at looking at k values from 0.6 to 1 in fi [3.13,](#_bookmark128) we observe a noticeable small diff in anomaly detection rate between IBM and HBM for category AA. Our observation suggests that the diff between anomaly detection rate is small or close to none for IBM and HBM, on the bottom of the hierarchy, for anomaly with a diff t hierarchical structure, in a simulated setting with each category generated independently.

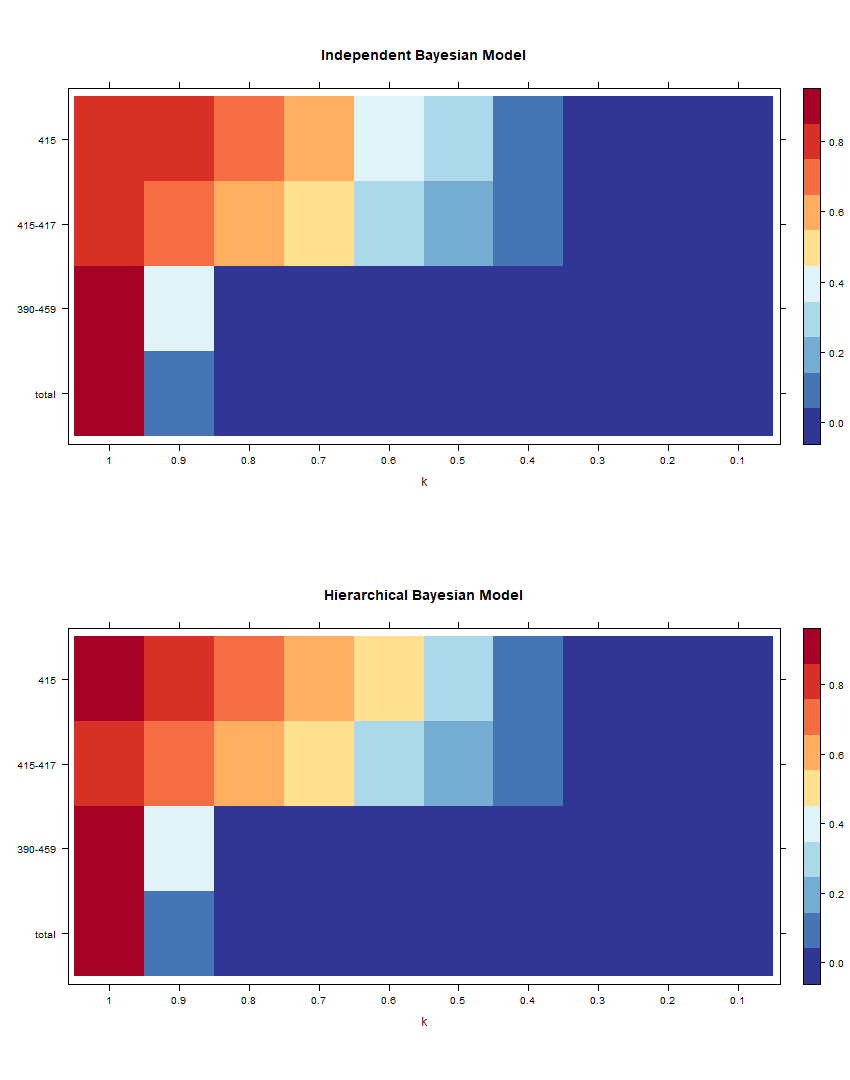


Figure 3.14: Heat-map of anomaly signals for ICD 415 and its ancestors, with no added anomalies, of IBM and HBM.

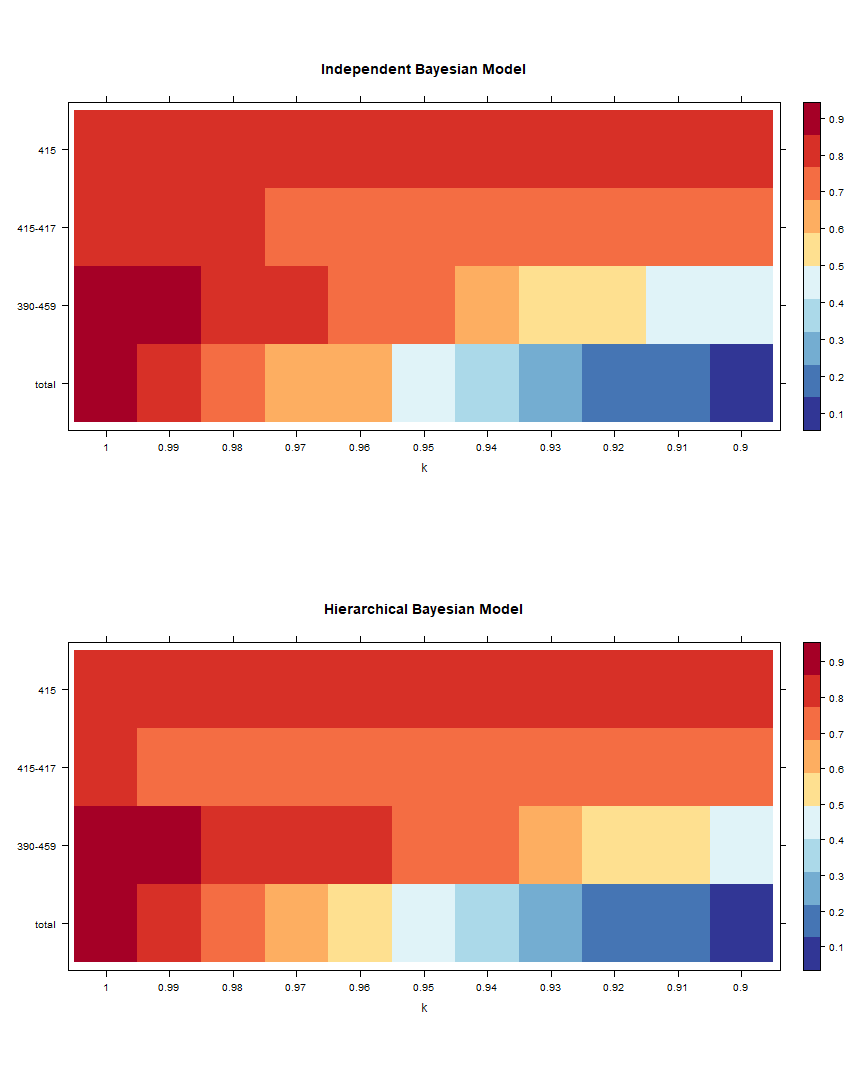


Figure 3.15: Heat-map of anomaly signals for ICD 415 and its ancestors, with no added anomalies, of IBM and HBM zoomed in at k values of 0.9 to 1.

The heat-map in fi [3.14](#_bookmark129) presents the anomaly signals of ICD 415 and their ancestors with no addition of a constant anomaly, with IBM and HBM, against diff t increments of *k*.Anomaly signals tend to be strong on the top left of the heat map. This suggests that signals of anomaly tend to increase in strength with increased k and higher hierarchy. Zooming in at looking at k values from 0.6 to 1 in fi [3.15,](#_bookmark130) we observe a noticeable small diff in anomaly detection rate between IBM and HBM for category AA. Our observation suggests that the diff between anomaly detection rate is small or close to none for IBM and HBM, on the bottom of the hierarchy, for anomaly with a diff t hierarchical structure, in a simulated setting with each category generated independently.

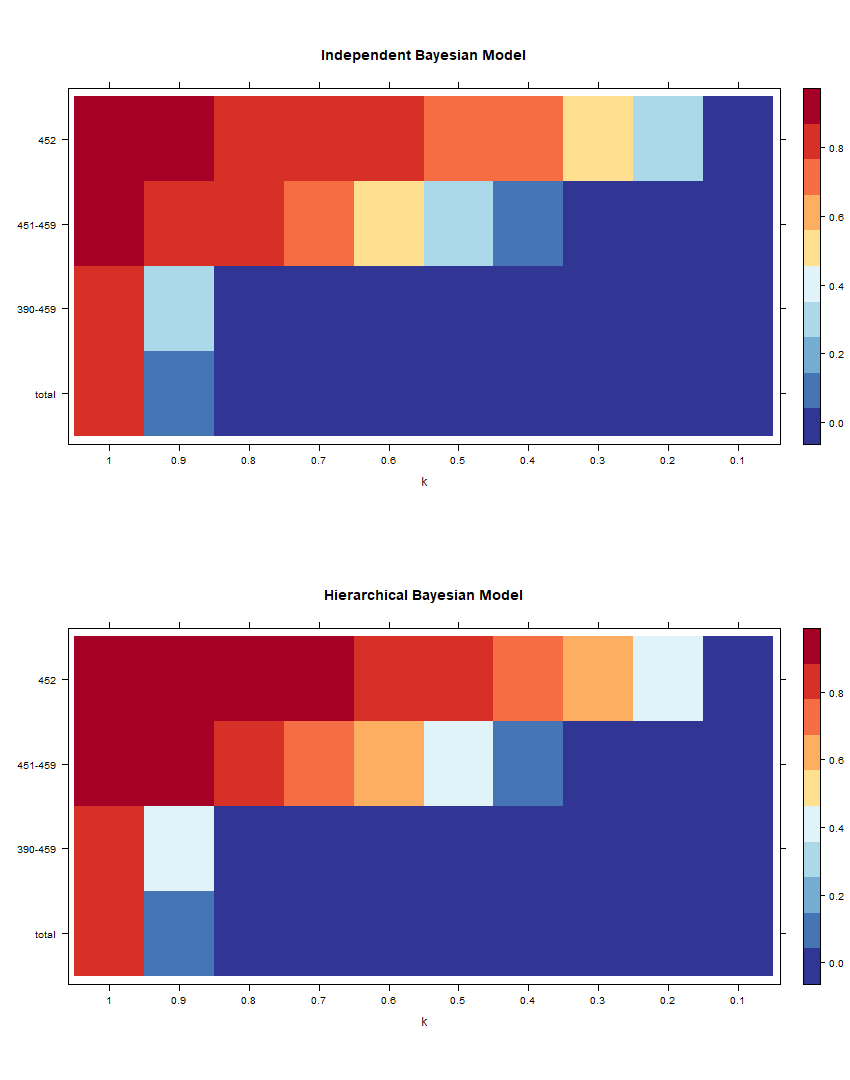


Figure 3.16: Heat-map of anomaly signals for ICD 452 and its ancestors, with no added anomalies, of IBM and HBM.

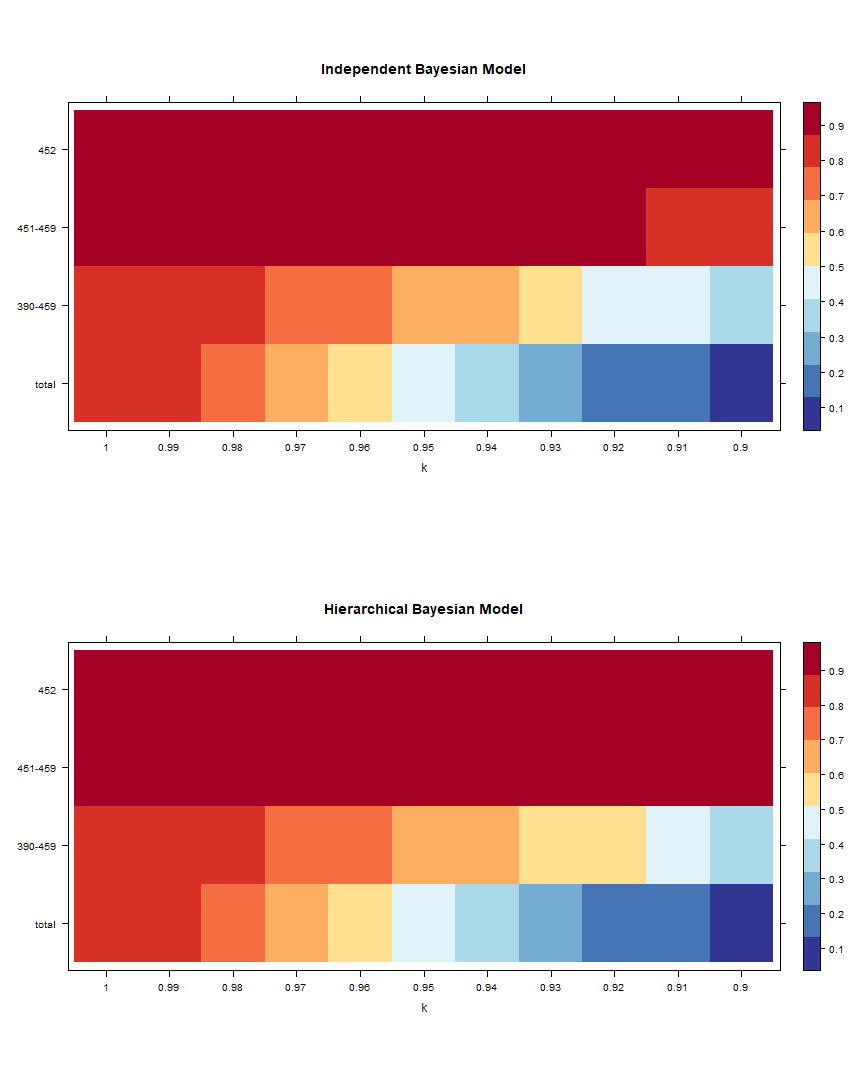


Figure 3.17: Heat-map of anomaly signals for ICD 452 and its ancestors, with no added anomalies, of IBM and HBM zoomed in at k values of 0.9 to 1.

The heat-map in fi [3.16](#_bookmark131) presents the anomaly signals of ICD 452 and their ancestors with no addition of a constant anomaly, with IBM and HBM, against diff t increments of *k*. Anomaly signals tend to be strong on the top left of the heat map. This suggests that signals of anomaly tend to increase in strength with increased k and higher hierarchy. Zooming in at looking at k values from 0.6 to 1 in fi [3.17,](#_bookmark132) we observe a noticeable small diff in anomaly detection rate between IBM and HBM for category AA. Our observation suggests that the diff between anomaly detection rate is small or close to none for IBM and HBM, on the bottom of the hierarchy, for anomaly with a diff t hierarchical structure, in a simulated setting with each category generated independently.

There are significant diff in anomaly signals of disease categories with diff t rate of occurrence. For common diseases (ICD 410)the signals seem to be more clear (change in colour for diff t increments of *k* is more abrupt) compare to rare (ICD 415) and extremely rare (ICD 452)disease categories. Signals seem to exist for categories at level 3 readily and level 2 of the hierarchy, for commonly diseases signals can be observed in level 1, but for rare and extremely rare diseases there is almost no signal at level 1. Signals at level 2 seem to be weaker than signals at level 3 overall, but it is still apparent and easily detectable.

# Discussion

We were unable to make a conclusion for the goodness-of-fi comparison between IBM and HBM because the mean and median values of diff in DIC are very close to 0. This suggests that for real-life data with large complex hierarchical structures, IBM and HBM performed similarly. The fi is unexpected, as our simulation studies and several academic papers all seem to have suggested HBM be the superior model. There are several possible explanations for this. The fi explanation is that for large complex hierarchical structures there are a huge amount of categories at each of the levels, and randomness from all these categories could have masked the small improvements in the goodness of fi for anomalies at one small category. Also that our HBM are quite primitive and there is no way for us to know whether or it really captures the hierarchical relationship, and the validity of our HBM is in question. If our HBM does not capture the hierarchical structure information, then there will not be any reason that HBM performs IBM. We also made observations that, for an extremely rare disease, the posterior distribution is not normal an seem to be skewed, but our model assumes it to be normal. We could possibly use diff t priors for categories with diff t magnitude of counts, for example, assigning a Half-Cauchy prior that was suggested by [Gelman et al.](#_bookmark185) [(2006)](#_bookmark185) for extremely rare categories.

There are significant diff in anomaly signals of disease categories with

diff t rate of occurrence. We noticed that for common disease category signals of the anomaly are still visible at level 1 of the hierarchy, but for rare and extremely are disease categories, there is almost no signal. However, signals can still be readily observed at level 2 and level 3 of the hierarchy for rare and extremely rare disease categories. Our observation made sense because rare and common disease have a much smaller magnitude of counts and as lower hierarchy aggregate into a higher hierarchy, smaller counts tend to get masked by randomness and uncertainty from other categories, so it would become harder to detect the anomaly at higher hierarchy levels. For ICD-9 disease categories, we are able to detect anomaly at the bottom two levels. This suggests that we should not look at anomaly at level 1 of hierarchy because we are likely to miss it. The rarity of disease also have a significant impact on anomaly detection, for common disease with large average values, signals of the anomaly are clear and easy to detect, but for rare and extremely rare disease categories it is unclear and harder to decide.

Overall we see that if the capacity is under 60% ish, there are almost no anomaly signal, this meant that if the hospitals are operating at capacities at or under 60%, they will almost never exceed their full capacity even if there are twice as many arrivals for certain disease categories compare to usual. It is a bit hard to make sense of this, but essentially what you need to know is that operating at a lower capacity should have a significant impact on the likelihood of anomalous arrivals. The obvious solution to congestion is to decrease the operation capacity (or increase overall capacity), but this is easier said than been done because the hospitals would have done it if they could. However, our results still high lighted the fact that having a lower operation capacity could effectively buff r against anomalous events and reduce the likelihood of congestion from occurring.

**Chapter 4 Conclusion**

# Introduction

For this thesis we ran several simulations that explore the performance of Hierarchical Bayesian Model (HBM) over the Independent Bayesian Model(IBM). We also inves- tigated whether taking existing hierarchical structures in ICD-9 disease diagnoses codes into account would affect the anomaly signal for diff nt occurrence levels of disease diagnoses. In this chapter, the fi from this thesis are summarised and discussed, extensions on our methodology and potential directions for future research are discussed.

* 1. **Summary of findings**

Models that used weakly-informed priors had outperformed the model with non- informative prior, in terms of model goodness-of-fi Assignment of prior is a prereq- uisite of a quality model, without any prior results of our Bayesian model will diff with models that have a prior by an substantial amount. However, small diff ces in the distribution of the prior does not create noticeable diff therefore we have great freedom in what prior to use. There are number of suggestion by academics in the fi such as Gilman and most are fairly robust and feasible options.

size of the anomalies have strong association with several posterior values. Increase of size of anomaly will decrease number of Effective Sample Size (ESS), and increase the size of standard deviation. This is mostly due to how the relative size of values diff for large and small values. Due to the aggregation constraint there are usually significant diff ence in size for different levels of hierarchy, higher the hierarchy the larger the mean size. For such reasons, when we are dealing with hierarchical datasets, we must take into account the hierarchical structure otherwise power of our models

will be greatly reduced. There are more information at lower levels of the hierarchy, therefore when informations about hierarchical structure is disregarded, the lower level of hierarchy often result in greatest diff between IBM and HBM.

For real-life datasets with complex hierarchical structures it was very diffi to prove superiority of HBM over IBM, and hence no conclusions can be made. there are several possible reasons that might contribute to the diffi y, such as uncertainty, and the validity of the HBM used. There are significant diff in anomaly signals of disease categories with diff ent rate of occurrence. for more common diseases it will have greater impact on the higher hierarchy and there maybe signals from level 3 to level 1, for rare and extremely rare cases we could not detect anomaly at level 1 and can only detect anomaly ay level 2 to level 3. There seem to be a dissipating effect, that anomalies occur at a level will lose it’s presences, for levels further away from the original level. Of course, this is due to aggregation of other groups masking the anomaly with their noise. For smaller categories dissipation effect are stronger.

# Extensions on Methodology

Hierarchical time-series techniques have worked fairly well for our studies, some of the down side is that the time-consumliess of the process and that the optimally reconciliation algorithm gives negative results. Unfortunately function for non- negative optimal reconciliation prediction will not be implemented until after the submission of the thesis.

Uncertainty in Bayesian estimations has certainly add diffi to the processes of model selections and result exploration. Uncertainty is a strength and a weakness of Bayesian inference, in our case diff in DIC are sometimes hard to distinguish, and we were unable to made strong claims to our fi Also with increased parameter that contained uncertainty, the computational cost increases dramatically with the increase in dimensionality, and limits the practicality. For simple models the problems isn’t so bad, but with large complex data, running models may take hours and hours, and we often cannot make significant observations from results.

# Implications

A possible implications of our work is to create an online signal detection alarm system that raises an alarm that notifi hospital management of congestion when and before it occurs, and models the possibility of abnormally high arrivals. With all the development in big data, data linkage and data streaming, there is enough resource for this idea to start in development in a practical setting. The idea is that by

CHAPTER 4. CONCLUSION 4.5. FUTURE DIRECTIONS

collecting data from various sources and send them to an online cloud platform, that integrates this information on arrival and exports signals of anomalies for diff t locations. If an alarm is raised at a location, steps could be taken at the earliest instance to deal with the problem. The essence of the idea is to detect and deal with the problem before it occurs. It does not take any effort to call for interference if congestion happened at a location, but this takes time and human, and by the time the problem is noticed, and the message got through to management, the negative effect would already be in effect. The system would, in theory, improve the efficiency of operation and reduce the chance of congestion of happening.

Another possible practical application of our techniques is to disaggregate the geo-location attribute from the New Zealand National Minimum Dataset (NMDS) data and try to monitor the anomalies for a geo-location hierarchy. Information analyst from the Ministry of Health has suggested that anomalies in health data are often due to diff in patient management strategies for diff t District Health Boards(DHB) from diff t parts of the country. For example, is there an after-hours clinic available or is the hospital the only available treatment provider? Alternatively, does this DHB manage patients by transferring them through diff t facilities or does it all take place in one facility (albeit potentially covering diff t sites)? Modelling and identifying anomalies at the individual doctor level, department level, individual facility level, DHB level, and island level with HBM could provide statistical evidence about how each doctor department, facility and DHB in New Zealand is performing independently and collectively. If a particular department always shows signals of abnormal arrivals, it could suggest that the department is operating at near full capacity and this particular department need more doctors and staff. If a high anomaly at particular DHB is observed, only in the evening, maybe an overall lack of staff for the after-hour clinic is the cause and roaster needs to be adjusted. The main idea is that observations of anomalies could translate into problematic operations and our HBM techniques have a significant advantage to, especially in small and time-related anomalies that are hard to detect by the intuition of hospital management, and informed decisions can be made to improve the current system and alleviate the problem of congestion.

# Future Directions

Future work will be aimed to extend the scope of the current simulations. For our exploratory analysis, HBM has shown to have superior performance over IBM, we could still think of other settings that could also test the robustness of HBM with datasets with diff nt qualities, and evaluate the performance of HBM from diff t perspectives of data. For example We could increase the number of levels of hierarchy,

and see how HBM perform with data with increasing complexity. We could try and test for hierarchical networks that do not necessarily follow the aggregation constraint, and see how HBM perform for networks. Another possible idea is to explore how HBM would perform if there are strong inter-level correlations, and see how multiplicity will affect our results.

There is also a hope to move the current research into a more practical setting, for current crisis in hospital congestions. In order to do this a great amount of background knowledge about specific areas such as operations research, public health and hospital management is required, so we would need to try and collaborate with expert in the fi and identify issues that need to be addresses and start forming ideas about how we could address the issue with our techniques

# Concluding remarks

Our fi suggest that Use of HBM provides better results than IBM for the case of anomaly detection for hierarchical datasets, HBM tends to have better goodness of fi and the diff in detection results are apparent. HBM is a feasible design that is well-suited for datasets that have attributes with large hierarchical structures and have an advantage over IBM in many areas. Hierarchical structures

**Appendix A Codes**

This appendix provides an overview of functions created and models built in this thesis. The full collection of R, JAGS, and SAS codes used to synthesise datasets, clean and merge datasets, run simulations, and draw table and plots and are available on GitHub repository: <https://github.com/jungxue/research-masters-Jung>

**Note:** All data set synthesised and used in this thesis are also uploaded on the GitHub repository, except for datasets from MIMIC Critical Care Database, more specifically three MIMIC data set were used (ADMISSIONS.csv, D ICD DIAGNOSES.csv and DIAGNOSES ICD.csv) for security and confidentiality reasons. To access datasets

in the MIMIC database, you must complete a training course “Data or Specimens Only Research” provided by CITI, and agree to a Data Use Agreement. The whole process takes around 2-3 hours. If you would like to request MIMIC data, please visit <https://mimic.physionet.org/gettingstarted/access/>

# A.1 Simulations

Custom R Functions were created for the synthesis of datasets. Here are the Each Function was saved on diff t R scripts and were simply loaded using source() function. The scripts can be found on the GitHub Repository.

rand.day.time Generation of single random time.

rand.day Generation of N random periods with custom length.

simdata Generation of random dataset.

cumdata Convert raw simulated data into a cumulative dataset.

tabulatedata Tabulate raw simulated data into daily counts.

ggcumplot Draw a line plot from cumulative dataset.

ggdailyplot Draw a line plot from daily counts dataset.

adddaily.anomaly Add custom anomalies to simulated data.

## rand.day.time

**Description:** Generation of single random time.

**Note:** originally by [Dirk Eddelbuettel](https://stackoverflow.com/questions/14720983/efficiently-generate-a-random-sample-of-times-and-dates-between-two-dates) 2012, however there were minor bugs and was improved by Thomas Lumley 2018.

### Usage:

rand . day . time (N, st=" 2006 / 01 / 01 00 : 00 : 01 ", et=" 2018 / 12 / 31 23 : 59 : 59 ")

### Arguments:

N number of generations, non-adjustable default = 1. st starting time

et ending time

## rand.day

**Description:** Generation of N random periods with custom length.

### Usage:

rand . day <- function ( N = 2 , st = " 2006 / 01 / 01 00 : 00 : 01 ", et = " 2018 / 12

/ 31 23 : 59 : 59 ", period = c( 5 ,3 ))

### Arguments:

N number of periods st starting time

et ending time

period a vector containing length of days for each period

## simdata

**Description:** Generation of random dataset.

### Usage:

simdata <- function ( sim . number = 1000000 , time . start = " 2006 / 01 / 01 00 : 00 : 01 ", time . end = " 2018 / 12 / 31 23 : 59 : 59 ", cat 2 . val = qpcR ::: rbind . na( c( 250 ,250 ),c( 250 ,250 )))

### Arguments:

sim.number number of simulated data, each number/row correspond to a single event.

time.start starting time time.end ending time

cat2.val a matrix containing default number in each leaf, must sum to 1000, each row correspond to a brunch, must use 0 for no leaf

## cumdata

**Description:** Convert raw simulated data into a cumulative dataset.

### Usage:

cumdata ( x)

### Arguments:

x a dataset, created using simdata function

## tabulatedata

**Description:** Tabulate raw simulated data into daily counts.

### Usage:

tabulatedata ( x= sample 1 )

### Arguments:

x a dataset, created using simdata function

## ggcumplot

**Description:** Draw a line plot from cumulative dataset.

### Usage:

ggcumplot ( data = cumdata . df , nameB 1 = " ggcumplot B 1 . jpg ", name Leaf = " ggcumplotleaf . jpg ")

### Arguments:

data cumulative dataset, created using simdata and cumdata function

nameB1 name for the jpg fi for line plot of 1st brunch nameLeaf name for the jpg fi for line plot of leaves

## ggdailyplot

**Description:** Draw a line plot from daily counts dataset.

### Usage:

ggdailyplot ( data = dailycount . df , nameB 1 = " ggdailyplot B 1 . jpg ", name Leaf = " ggdailyplotleaf . jpg ")

### Arguments:

data daily counts dataset, created using simdata and tabulatedata function

nameB1 name for the jpg fi for line plot of 1st brunch nameLeaf name for the jpg fi for line plot of leaves

## adddaily.anomaly

**Description:** Add custom anomalies to simulated data.

### Usage:

adddaily . anomaly ( data = daily 1 . df , usepercent = T, Total N = 100 , Total Npercent = 1000 , Anomalytype = " Period ", nperiod = c( 5 ,3 ), day = rand . day ( N= 2 , period = c( 5 , 3 )), Leafpercent = rbind ( c( 1 , 1 , 0 , 0

. 25 ,0 . 75 ,0 . 00 ,0 . 00 ),c( 1 ,1 ,0 ,1 ,0 ,0 ,0 )), Distribution = " Normal ",

output = " rdata / function / anomalytest . RData ")

### Arguments:

data daily counts dataset, created using simdata and tabulatedata function

usepercent using percent True or False

TotalN if usepercent = False, count in terms of absolute number TotalNpercent if usepercent = True, count in terms of percentage of mean

total

Anomalytype type of anomaly, Point or Period nperiod number of day of each period

day random periods with custom length.

Leafpercent percentage of anomaly added on total, brunch 1 and each leaf in a vector

Distribution Distribution used for period anomalies output Name of output fi as RData

# JAGS models

Table A.1: Code used to specify Independent Bayes Models for Chapter 2.3 a

Model Example JAGS code

* + 1. # Null model

model 1 . txt = " model {

# Likelihood

for ( j in 1: Nleaf ){ mu[ j] <- rho [ j] for ( i in 1: Nday ){

Y[ i, j] ~ dpois ( mu[ j])

}}} "

* + 1. # Normal (1 ,0.3) prior

model 2 . txt = " model {

# Likelihood

for ( j in 1: Nleaf ){

mu[ j] <- rho [ j]\* lambda for ( i in 1: Nday ){

Y[ i, j] ~ dpois ( mu[ j])

}}

### Prior

lambda ~ dnorm ( mu2 . y2 , sqrt ( sigma 2 . y2 )) T(0 ,)

### Hyper prior

mu2 . y2 ~ dnorm (1 ,0.1)

sigma 2 . y2 ~ dnorm (0.3 ,0.1)

}"

Table A.2: Code used to specify BHMs for Chapter 3 b

Model JAGS code

* + 1. # Normal (1 ,0.1) prior

model 3 . txt = " model {

# Likelihood

for ( j in 1: Nleaf ){

mu[ j] <- rho [ j]\* lambda for ( i in 1: Nday ){

Y[ i, j] ~ dpois ( mu[ j])

}

}

### Prior

lambda ~ dnorm ( mu2 . y3 , sqrt ( sigma 2 . y3 )) T(0 ,)

### Hyper prior

mu2 . y3 ~ dnorm (1 ,0.1)

sigma 2 . y3 ~ dnorm (0.1 ,0.1)

}"

* + 1. # Gamma (4 ,3) prior

model 4 . txt = " model {

# Likelihood

for ( j in 1: Nleaf ){

mu[ j] <- rho [ j]\* lambda for ( i in 1: Nday ){

Y[ i, j] ~ dpois ( mu[ j])

}

}

### Prior

lambda ~ dgamma ( alpha . y4 , beta . y4 ) T(0 ,)

### Hyper prior

alpha . y4 ~ dnorm (4 ,0.1)

beta . y4 ~ dnorm (3 ,0.1)

}"

Table A.3: Code used to specify BHMs for Chapter 3 c

Model JAGS code

* + 1. # laplace (1 ,1) prior

model 5 . txt = " model {

# Likelihood

for ( j in 1: Nleaf ){

mu[ j] <- rho [ j]\* lambda for ( i in 1: Nday ){

Y[ i, j] ~ dpois ( mu[ j])

}}

### Prior

lambda ~ ddexp ( mean . y5 , scale . y5 ) T(0 ,)

### Hyper prior

mean . y5 ~ dnorm (1 ,0.1)

scale . y5 ~ dnorm (1 ,0.1)

}"

* + 1. # 0.9 Mixture prior with Normal (1 ,0.1) prior

model 6 . txt = " model {

# Likelihood

for ( j in 1: Nleaf ){

mu[ j] <- rho [ j]\* lambda for ( i in 1: Nday ){

Y[ i, j] ~ dpois ( mu[ j])

}}

### Prior

lambda ~ dnorm ( spike \* 1 + (1 - spike )\* slab , 0.1) T(0 ,)

### Hyper prior spike ~ dbin (0.9 ,1)

slab ~ dnorm ( mu2 . y6 , sqrt ( sigma 2 . y6 ))

### Hyper - hyper prior mu2 . y6 ~ dnorm (1 ,0.1)

sigma 2 . y6 ~ dnorm (0.1 ,0.1)

}"

# Hierarchical Time-series Predictions

Table A.4: Example of R code used to make time-series prediction used in Chapter 3

fc 3 = forecast ( mimic 3 . yearly . hts , h = 1 , method = " comb ", fmethod = " arima ", FUN = function ( x) tbats (x, use . parallel =

TRUE ))

fc 7 = forecast ( mimic 3 . yearly . hts , h = 1 , method = " bu", fmethod = " arima ", FUN = function ( x) tbats (x, use . parallel = TRUE ))

# Independent and Hierarchical Models

modelmimic . txt = " model {

# Likelihood

for ( j

mu[ j] for ( i Y[ i, j]

}}

in

in

1: Nleaf ){

<- rho [ j]\* lambda [ j] 1: Nday ){

~ dpois ( mu[ j])

for ( j in 1: Nleaf ){

### Prior

lambda [ j] ~ dnorm ( mu2 . y2 [ j], sqrt ( sigma 2 . y2 [ j])) T(0 ,)

### Hyper prior

mu2 . y2 [ j] ~ dnorm (1 ,0.1)

sigma 2 . y2 [ j] ~ dnorm (0.1 ,0.1)

}}"

Table A.5: JAGS code used for the Independent Bayesian Model used in Chapter 3

modelmimich . txt = " model {

### Level 0

mu2 . y2 . lv0 ~ dnorm (1 ,0.1) sigma 2 . y2 . lv0 ~ dnorm (0.1 ,0.1)

lambda 1 ~ dnorm ( mu2 . y2 . lv0 , sigma 2 . y2 . lv0 ) T(0 ,)

### level 1

for ( l in 1: Nlv1 ){

mu2 . y2 . lv1 [ l] ~ dnorm ( lambda1 ,0 .1 ) sigma 2 . y2 . lv1 [ l] ~ dnorm (1 ,0.1)

lambda 2 [ l] ~ dnorm ( mu2 . y2 . lv1 [ l], sigma 2 . y2 . lv1 [ l]) T(0 ,)

}

### Level 2

for ( m in 1: Nlv2 ){

mu2 . y2 . lv2 [ m] ~ dnorm ( lambda 2 [ lV1b [ m ]] ,0.1) sigma 2 . y2 . lv2 [ m] ~ dnorm (1 ,0.1)

lambda 3 [ m] ~ dnorm ( mu2 . y2 . lv2 [ m], sigma 2 . y2 . lv2 [ m]) T(0 ,)

}

### level 3

for ( n in 1: Nlv3 ){

mu2 . y2 . lv3 [ n] ~ dnorm ( lambda 3 [ lV2b [ n ]] ,0.1) sigma 2 . y2 . lv3 [ n] ~ dnorm (1 ,0.1)

lambda 4 [ n] ~ dnorm ( mu2 . y2 . lv3 [ n], sigma 2 . y2 . lv3 [ n]) T(0 ,)

}

# prior \* Likelihood

lambda

<- c( lambda1 , lambda2 , lambda3 , lambda 4 )

for ( j

mu[ j] for ( i Y[ i, j]

}}

in

1: Nleaf ){

<- rho [ j]\* lambda [ j]

in 1: Nday ){

~ dpois ( mu[ j])

}"

Table A.6: JAGS code used for the Hierarchical Bayesian Model used in Chapter 3

**Bibliography**

Adams, T., O’Sullivan, M., and Walker, C. (2018). A simulation model of auckland city hospital’s emergency department. Unpublished report.

Akoorie, N. (2018). Underfunding of health system leads to $ 240m blowout. Retrieved from [https://www.nzherald.co.nz/nz/news/article.cfm?c\_id=1&objectid=](https://www.nzherald.co.nz/nz/news/article.cfm?c_id=1&amp;objectid=12128639) [12128639](https://www.nzherald.co.nz/nz/news/article.cfm?c_id=1&amp;objectid=12128639).

Allenby, G. M., Rossi, P. E., and McCulloch, R. E. (2005). Hierarchical bayes models: A practitioners guide. ssrn scholarly paper id 655541. *Social Science Research Network, Rochester, NY*.

Andrus, B. and Bohnett, C. (2013). Understanding the icd-10 code structure. Retrieved from [https://www.webpt.com/blog/post/understanding-icd-/](https://www.webpt.com/blog/post/understanding-icd-/10-code-structure) [10-code-structure](https://www.webpt.com/blog/post/understanding-icd-/10-code-structure).

Bauer, D., Higgs, B. W., and Mohtashemi, M. (2007). High performance computing for disease surveillance. In Zeng, D., Gotham, I., Komatsu, K., Lynch, C., Thurmond, M., Madigan, D., Lober, B., Kvach, J., and Chen, H., editors, *Intelligence and Security Informatics: Biosurveillance*, pages 71–78, Berlin, Heidelberg. Springer Berlin Heidelberg.

Bayley, M. D., Schwartz, J. S., Shofer, F. S., Weiner, M., Sites, F. D., Traber,

K. B., and Hollander, J. E. (2005). The fi burden of emergency department congestion and hospital crowding for chest pain patients awaiting admission. *Annals of emergency medicine*, 45(2):110–117.

Berry, S. M. and Berry, D. A. (2004). Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model. *Biometrics*, 60(2):418–426.

Betancourt, M. (2018). How the shape of a weakly informative prior af- fects inferences. Retrieved from [https://mc-stan.org/users/documentation/](https://mc-stan.org/users/documentation/case-studies/weakly_informative_shapes.html) [case-studies/weakly\_informative\_shapes.html](https://mc-stan.org/users/documentation/case-studies/weakly_informative_shapes.html).

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens,

M. H. H., and White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, 24(3):127–135.

B˚a˚ath, R. (2015). Introduction to bayesian data analysis with r. *Lund University*. Re- trieved from [http://www.sumsar.net/files/academia/user\_2015\_tutorial\_](http://www.sumsar.net/files/academia/user_2015_tutorial_bayesian_data_analysis_short_version.pdf) [bayesian\_data\_analysis\_short\_version.pdf](http://www.sumsar.net/files/academia/user_2015_tutorial_bayesian_data_analysis_short_version.pdf).

Brewer, B. J. (2017). Lecture notes in stats 331 introduction to bayesian statistics.

University of Auckland.

Carlin, B. P. and Louis, T. A. (2008). *Bayesian methods for data analysis*. Chapman and Hall/CRC.

Centers for Disease Control and Prevention (2019). International classification of diseases,ninth revision, clinical modifi (icd-9-cm). Retrieved from [https:](https://www.cdc.gov/nchs/icd/icd9cm.htm)

[//www.cdc.gov/nchs/icd/icd9cm.htm](https://www.cdc.gov/nchs/icd/icd9cm.htm).

Chambers, C. N., Frampton, C. M., Barclay, M., and McKee, M. (2016). Burnout prevalence in new zealand’s public hospital senior medical workforce: a cross- sectional mixed methods study. *BMJ open*, 6(11):e013947.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey.

*ACM computing surveys (CSUR)*, 41(3):15.

Cleland, I. D., Mason, A., and O’Sullivan, M. (2018). Improved staff rostering algorithms. 2018 Joint NZSA/ORSNZ Conference. Palmerston North, New Zealand.

Congdon, P. (2007). *Bayesian statistical modelling*, volume 704. John Wiley & Sons. Cook, J. D. (2017). Effective sample size for mcmc. Retrieved from [https://www.](https://www.johndcook.com/blog/2017/06/27/effective-sample-size-for-mcmc/)

[johndcook.com/blog/2017/06/27/effective-sample-size-for-mcmc/](https://www.johndcook.com/blog/2017/06/27/effective-sample-size-for-mcmc/).

Cornwall, J. and Davey, J. A. (2004). *Impact of population ageing in New Zealand on the demand for health and disability support services, and workforce implications*. Ministry of Health Wellington, New Zealand.

Das, S. (1994). Time series analysis.

Drechsler, J. (2011). *Synthetic datasets for statistical disclosure control: theory and implementation*, volume 201. Springer Science & Business Media.

Drechsler, J., Bender, S., and R¨assler, S. (2008). Comparing fully and partially synthetic datasets for statistical disclosure control in the german iab establishment panel. *Trans. Data Privacy*, 1(3):105–130.

Edgeworth, F. Y. (1887). Xli. on discordant observations. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 23(143):364–375.

Fewster, R. (2014). Lecture notes in stats 325 stochastic processes. University of Auckland.

Gelman, A. (2006). Multilevel (hierarchical) modeling: what it can and cannot do.

*Technometrics*, 48(3):432–435.

Gelman, A. 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, A. and Hill, J. (2007). Data analysis using regression and hierarchical/mul- tilevel models. *New York, NY: Cambridge*.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and computing*, 24(6):997–1016.

Golchi, S. (2016). Informative priors and bayesian computation. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 782–789. IEEE.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark,

R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples.

*Technometrics*, 11(1):1–21.

Grundmann, O. (2014). The current state of bioterrorist attack surveillance and preparedness in the us. *Risk management and healthcare policy*, 7:177.

HemOnc Today (2008). The psychology of oncology: Physician burnout is going unrecognized. Retrieved from [www.healio.com](http://www.healio.com/).

Hyndman, R., Lee, A., Wang, E., and Wickramasuriya, S. (2018). *hts: Hierarchical and Grouped Time Series*. R package version 5.1.5.

Hyndman, R. J. (2015). Optimal forecast reconciliation for big time series data. Retrieved from <https://robjhyndman.com/talks/Google-Oct2015-part2.pdf>.

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., and Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics*

*& Data Analysis*, 55(9):2579–2589.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*.

OTexts, Melbourne, Australia.

Hyndman, R. J., Athanasopoulos, G., et al. (2014). Optimally reconciling forecasts in a hierarchy. *Foresight: The International Journal of Applied Forecasting*, 1(35):42– 48.

Jia, Z. (2018). Posterior model odds estimation via gibbs sampler. University of Auckland. Unpublished dissertation.

Johnson, A. E., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035.

Jones, G. (2018). Diagnostic test evaluation or: How i learned to stop worrying and love bayesian statistics. 2018 Joint NZSA/ORSNZ Conference. Palmerston North, New Zealand.

Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. M. (1998). Markov chain monte carlo in practice: a roundtable discussion. *The American Statistician*, 52(2):93–100.

Khare, R. K., Powell, E. S., Reinhardt, G., and Lucenti, M. (2009). Adding more beds to the emergency department or reducing admitted patient boarding times: which has a more significant infl on emergency department congestion? *Annals of emergency medicine*, 53(5):575–585.

Kruschke, J. K. and Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. *The Oxford handbook of computational and mathematical psychology*, pages 279–299.

La Rosa, A. (2018). Anomaly detection in monitoring: Problems we can resolve with it. Retrieved from [https://pandorafms.com/blog/](https://pandorafms.com/blog/anomaly-detection-in-monitoring/) [anomaly-detection-in-monitoring/](https://pandorafms.com/blog/anomaly-detection-in-monitoring/).

Laplace, P.-S. (1998). *Pierre-Simon Laplace Philosophical Essay on Probabilities: Translated from the fifth French edition of 1825 With Notes by the Translator*, volume 13. Springer Science & Business Media.

Mark, R. (2016). *The Story of MIMIC*, pages 43–49. Springer International Publishing, Cham.

McGrayne, S. B. (2011). *The theory that would not die: how Bayes’ rule cracked the enigma code, hunted down Russian submarines, & emerged triumphant from two centuries of controversy*. Yale University Press.

MedlinePlus Medical Encyclopedia (2019). Cor pulmonale. Retrieved from [https:](https://medlineplus.gov/ency/article/000129.htm)

[//medlineplus.gov/ency/article/000129.htm](https://medlineplus.gov/ency/article/000129.htm).

Ministry of Health (2009). About emergency departments. Retrieved from [https://www.health.govt.nz/our-work/hospitals-and-specialist-care/](https://www.health.govt.nz/our-work/hospitals-and-specialist-care/emergency-departments/about-emergency-departments) [emergency-departments/about-emergency-departments](https://www.health.govt.nz/our-work/hospitals-and-specialist-care/emergency-departments/about-emergency-departments).

Ministry of Health (2011). Emergency department triage. Retrieved from [https://www.health.govt.nz/our-work/hospitals-and-specialist-care/](https://www.health.govt.nz/our-work/hospitals-and-specialist-care/emergency-departments/emergency-department-triage) [emergency-departments/emergency-department-triage](https://www.health.govt.nz/our-work/hospitals-and-specialist-care/emergency-departments/emergency-department-triage).

Ministry of Health (2015). Interface with primary health care. Retrieved from [https://www.health.govt.nz/our-work/hospitals-and-specialist-care/](https://www.health.govt.nz/our-work/hospitals-and-specialist-care/emergency-departments/interface-primary-health-care) [emergency-departments/interface-primary-health-care](https://www.health.govt.nz/our-work/hospitals-and-specialist-care/emergency-departments/interface-primary-health-care).

Ministry of Health (2016). How the electives process works. Retrieved from [https://www.health.govt.nz/our-work/hospitals-and-specialist-care/](https://www.health.govt.nz/our-work/hospitals-and-specialist-care/elective-services/how-electives-process-works) [elective-services/how-electives-process-works](https://www.health.govt.nz/our-work/hospitals-and-specialist-care/elective-services/how-electives-process-works).

Ministry of Health (2019). Icd-10-am/achi/acs. Retrieved from [https://www.](https://www.health.govt.nz/nz-health-statistics/classification-and-terminology/icd-10-am-achi-acs) [health.govt.nz/nz-health-statistics/classification-and-terminology/](https://www.health.govt.nz/nz-health-statistics/classification-and-terminology/icd-10-am-achi-acs) [icd-10-am-achi-acs](https://www.health.govt.nz/nz-health-statistics/classification-and-terminology/icd-10-am-achi-acs).

Morley, C., Unwin, M., Peterson, G. M., Stankovich, J., and Kinsman, L. (2018). Emergency department crowding: A systematic review of causes, consequences and solutions. *PloS one*, 13(8):e0203316.

Nickel, S. (2018). Health care logistics. 2018 Joint NZSA/ORSNZ Conference.

Palmerston North, New Zealand.

Open Science Collaboration and others (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Organization, W. H. (2019). Classification of diseases (icd),, 11th revision (icd-11).

Retrieved from <https://www.who.int/classifications/icd/en/>.

Park, D. K., Gelman, A., and Bafumi, J. (2004). Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis*, 12(4):375–385.

Patcha, A. and Park, J.-M. (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer networks*, 51(12):3448– 3470.

Paterson, R. and Adams, J. (2011). Professional burnout–a regulatory perspective.

*NZ Med J*, 124(1333):40–6.

Paulos, J. A. (2011). The mathematics of changing your mind. *New York Times (US)*.

Plummer, M. (2018). *rjags: Bayesian Graphical Models using MCMC*. R package version 4-8.

Plummer, M. et al. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, page 10. Vienna, Austria.

Pollard, T. J. and Johnson, A. E. W. (2016). The mimic-iii clinical database. [http://dx.doi.org/10.13026/C2XW26.](http://dx.doi.org/10.13026/C2XW26)

Ponziani, F. R., Zocco, M. A., Campanale, C., Rinninella, E., Tortora, A., Di Maurizio, L., Bombardieri, G., De Cristofaro, R., De Gaetano, A. M., Landolfi, R., et al. (2010). Portal vein thrombosis: insight into physiopathology, diagnosis, and treatment. *World journal of gastroenterology: WJG*, 16(2):143.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rahtz, M. (2015). Anomaly detection with k-means clustering. Retrieved from

<http://amid.fish/anomaly-detection-with-k-means-clustering>.

RStudio Team (2015). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.

Salway, R., Valenzuela, R., Shoenberger, J., Mallon, W., and Viccellio, A. (2017). Emergency department (ed) overcrowding: evidence-based answers to frequently asked questions. *Revista M´edica Cl´ınica Las Condes*, 28(2):213–219.

SAS Institute Inc (2013). *SAS 9.4 Statements: Reference.* SAS Institute Inc., Cary, NC.

Southland times (2015). Health board accepts delays caused woman’s death.

[*https://www.pr*](http://www.pressreader.com/)*essre*[*ader.c*](http://www.pressreader.com/)*om/*[.](http://www.pressreader.com/)

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fi *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4):583–639.

Stigler, S. M. (2002). *Statistics on the table: The history of statistical concepts and methods*. Harvard University Press.

Su, Y.-S. and Yajima, M. (2015). *R2jags: Using R to Run ’JAGS’*. R package version 0.5-7.

Target Internet (2018). What is data scraping and how can you use it? Retrieved from [https://www.targetinternet.com/what-is-data-scraping/](https://www.targetinternet.com/what-is-data-scraping/-and-how-can-you-use-it/)

-[and-how-can-you-use-it/](https://www.targetinternet.com/what-is-data-scraping/-and-how-can-you-use-it/).

The World Bank (2019). Current health expenditure %*ofgdp*, new zealand. Retrieved from <http://data.worldbank.org/indicator/SP.DYN.LE00.FE.IN>.

Wheeler, G. and Williamson, J. (2011). *Evidential probability and objective Bayesian epistemology*. Elsevier.

Wickramasuriya, S. L., Athanasopoulos, G., and Hyndman, R. J. (2019). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, 114(526):804–819.

Wolfgang, A. (2016). Hierarchical bayesian modeling. Penn State. Retrieved from

<https://astrostatistics.psu.edu/RLectures/hierarchical.pdf>.

Wong, H. J., Morra, D., Caesar, M., Carter, M. W., and Abrams, H. (2010). Un- derstanding hospital and emergency department congestion: an examination of inpatient admission trends and bed resources. *Canadian Journal of Emergency Medicine*, 12(1):18–26.

Wong, W.-K., Moore, A. W., Cooper, G. F., and Wagner, M. M. (2003). Bayesian network anomaly pattern detection for disease outbreaks. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 808–815.

Wong, Y. (2019). Lecture notes in stats 732 topic in statistical inference. University of Auckland.