Bayesian Fusion of Continuous-Valued Labels in Biomedical Applications

Tingting Zhu

Department of Engineering

University of Oxford

Supervised by

Prof. David A. Clifton

Prof. Gari D. Clifford

This thesis is submitted to the Department of Engineering Science, University of Oxford, in fulfilment of the requirements for the degree of *Doctor of Philosophy*

Kellogg College

Hilary Term, 2016

I would like to dedicate this thesis to my loving grandparents and parents.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university.

Tingting Zhu

Hilary Term, 2016

Acknowledgements

First, sincere thanks to my supervisors Prof. David Clifton and Prof. Gari Clifford, for the support and large degree of freedom I was given to pursue my own research interests during the DPhil. This has been instrumental in shaping me to become an independent researcher. Furthermore, I appreciate Gari for his academic advice, providing me numerous occasions to travel for research, setting up productive collaborations, and helping secure funding throughout my DPhil. David was supportive through all the challenges that arose in my DPhil, and provided mentorship through the journey with his personal experience and keen insight. I am grateful for all the guidance and opportunities that I was given.

It is needless to mention, but I would not have completed my DPhil without the help and support from my dearest friends and colleagues. Special thanks to those who edified me during my DPhil journey: Joachim Behar, for the power of mathematics; Alistair Johnson, for the importance of data pre-processing; Nick Dunkley, for the fun of modelling; Marco Pimentel, for the art of signal processing; Tasos Papastylianou, for the beauty of databases and web design. I would also like to extend my thanks, especially to Arvind Raghu, Athansios Tsanas, David Springer, Elina Naydenova, George Qian, Glen Colopy, Julien Oster, Kate Niehaus, Mauro Santos, and Maxim Osipov for the helpful discussion and support in my journey to make everything possible. I would also like to offer a more dedicated thanks to my thesis examiners, Prof. Maarten De Vos and Prof. Walter Karlen. I am also thankful for the support of the Research Councils UK Digital Economy Programme, and an ARM Scholarship through Kellogg College.

I would like to express the immense debt of gratitude I owe to my parents, Lino and Liya, for believing in me, encouraging me to pursue my dreams, and always being understanding. They are the best parents and more than I deserve. I would further thank my other family members for being there whenever I was in need.

At last, I would like to thank my husband, Aaron, for supporting me throughout the whole period with patience and love.

Tingting Zhu

Kellogg College

Doctor of Philosophy

Hilary Term, 2016

Bayesian Fusion of Continuous-Valued Labels in Biomedical Applications

Abstract

Expert labelling is the gold standard for diagnosing patient-specific diseases from medical data. However, experts are relatively scarce, their time is expensive, and the task of labelling is time-consuming. While automated algorithms offer advantages of time efficiency, repeatability, and cost-saving benefits compared to manual labelling, there remains a substantial discrepancy in their estimation as well as reliability that limit their use.

Two commonly-encountered and clinically-important examples were considered for the use of concept labels in clinical practice: the QT interval in the electrocardiogram and the respiratory rate estimation from the photoplethysmogram. This thesis sought to combine outputs of automated annotators (i.e., algorithms) to improve estimation of these exemplars in a principled manner.

A series of methods were proposed for inferring a consensus in a scenario when only noisy annotations of some presumed underlying ground truth are provided: (1) the Bayesian Continuous-Valued Label Aggregator (BCLA) with independent annotators was proposed to infer the ground truth while jointly predicting each algorithm's bias and precision. Both maximum-a-posteriori and Gibbs sampling approaches were derived to compute parameters in the BCLA model (denoted as BCLA-MAP and BCLA-Gibbs, respectively); (2) BCLA incorporating the notion of signal quality (i.e., BCLAs) was also proposed to provide a confidence measure relating to the quality of the input recordings; (3) BCLA with correlated annotators (BCLAc) model was further proposed to provide a more generalised framework for BCLA, in particular, two variants of BCLAc were considered to model the correlation of errors among annotators directly and indirectly. Furthermore, the proposed methods were compared to the mean, median, best-performing single annotator with least root-mean-squared-error, and two state-of-the-art benchmarking approaches described in the literature, namely "scalar Simultaneous Truth and Performance Level Estimation" and an Expectation-Maximisation label aggregation approach.

This thesis considered two exemplary medical datasets as well as simulated datasets to validate the proposed methods. The results of the proposed BCLA had optimal performance over the other comparison voting strategies in all datasets. Furthermore, BCLA-Gibbs as a fully-Bayesian approach was superior to BCLA-MAP. When incorporating an extension to include signal quality (i.e., the BCLAs model), results of BCLAs-MAP and BCLAs-Gibbs outperformed those of the basic BCLA methods consistently. Finally, we modelled the inclusion of potentially-correlated annotators using BCLAc.

In summary, the results of the methods proposed in this thesis improve accuracy by inferring a consensus in a scenario when only noisy labels of a group of imperfect algorithms are available. The proposed methods operate in an unsupervised Bayesian learning framework, and have the potential for real-time application to produce estimates that are more robust than any of the algorithms independently considered.

Relevant publications:

Journal articles

- **Zhu, T.**; Dunkley, N.; Behar, J.; Clifton, D. A. & Clifford, G. D. Fusing Continuous-Valued Medical Labels Using a Bayesian Model. *Annals of Biomedical Engineering*, 2015(43): 2892-2902.
- **Zhu, T.**; Johnson, A. E.; Behar, J. & Clifford, G. D. Crowd-Sourced Annotation of ECG Signals Using Contextual Information. *Annals of Biomedical Engineering*, 2014(42): 871-884.

Conference papers

- **Zhu, T.**; Pimentel, M. A.; Clifford, G. D. & Clifton, D. A. Bayesian fusion of algorithms for the robust estimation of respiratory rate from the photoplethysmogram. *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, 2015, 6138-6141.
- **Zhu, T.**; Osipov, M.; Papastylianou, T.; Oster, J.; Clifton, D. A. & Clifford, G. D. An intelligent cardiac health monitoring and review system. In *Appropriate Healthcare Technologies for Low Resource Settings*, 2014, 1-4.
- **Zhu, T.**; Behar, J.; Papastylianou, T. & Clifford, G. D. CrowdLabel: A crowdsourcing platform for electrophysiology. In *Computing in Cardiology Conference*, 2014, 789-792.
- Clifford, G. D.; Arteta, C.; **Zhu, T.**; Pimentel, M.; Santos, M.; Domingos, J.; Maraci, M.; Behar, J. & Oster, J. A scalable mHealth system for noncommunicable disease management. *IEEE Global Humanitarian Technology Conference*, 2014, 41-48.
- Behar, J.; **Zhu, T.**; Oster, J.; Niksch, A.; Mah, D.; Chun, T.; Greenberg, J.; Tanner, C.; Harrop, J.; Van Esbroeck, A. & others. Evaluation of the fetal QT interval using non-invasive foetal ECG technology. In *SMFM-34th Annual Meeting-The Pregnancy Meeting*, 2014.
- **Zhu, T.**; Johnson, A. E. W.; Behar, J. & Clifford, G. D. Bayesian voting of multiple annotators for improved QT interval estimation. In *Computing in Cardiology Conference*, 2013, 659-662.
- Silva, I.; Behar, J.; Sameni, R.; **Zhu, T.**; Oster, J.; Clifford, G. D. & Moody, G. B. Noninvasive fetal ECG: The PhysioNet/Computing in Cardiology Challenge 2013. In *Computing in Cardiology Conference*, 2013, 149-152.

Book chapter

- **Zhu, T.**; Clifford, G. D. & Clifton, D. A. Clifton, D. A. (Ed.) A Bayesian Model for Fusing Biomedical Labels. *Machine Learning for Healthcare Technologies*, The Institution of Engineering and Technology, 2016.

Table of contents

List of figures **xiii**

List of tables **xvii**

Nomenclature **xix**

1 Introduction **1**

1.1 Clinical Motivation 1

1.1.1 Exemplar I: QT Interval in the Electrocardiogram 2

1.1.2 Exemplar II: Respiratory Rate Estimation with the Photoplethysmogram 5

1.2 Outline of the Thesis 10

2 Background **13**

2.1 Introduction 13

2.2 Related Works 14

2.2.1 Ensemble Learning 14

2.2.2 Unsupervised Learning 15

2.2.3 Modelling Annotator Bias and Precision 21

2.3 Bayesian Probability in Parameter Estimation 26

2.4 Summary 29

3 Description of Datasets 30

3.1 Introduction . . . 30

3.2 The PhysioNet/Computing in Cardiology QT Dataset . . . 30

3.2.1 Dataset Description . . . 30

3.3 The Capnobase Respiratory Rate Dataset . . . 35

3.3.1 Dataset Description . . . 35

3.4 Simulated Datasets . . . 37

4 Non-Bayesian Continuous-Valued Label Aggregation 39

4.1 Introduction . . . 39

4.2 The EM-R Approach . . . 40

4.2.1 The Expectation-Maximisation Algorithm . . . 42

4.2.2 Learning from Incomplete Data . . . 45

4.3 Data Description and Methodology for Comparison . . . 45

4.4 Results . . . 52

4.5 Discussion . . . 58

4.6 Conclusion . . . 63

5 The Bayesian Continuous-Valued Label Aggregation Model 65

5.1 Introduction . . . 65

5.2 A Generative Model of Independent Annotators . . . 66

5.2.1 The Ground Truth Model . . . 66

5.2.2 The Annotator Model . . . 66

5.2.3 The Bayesian Continuous-Valued Label Aggregation (BCLA) Model 68

5.2.4 Simulated QT Dataset with Independent Annotators . . . 69

5.3 BCLA-MAP . . . 71

5.3.1 Convergence Criteria for BCLA-MAP using Extreme Value Theorem 74

5.3.2 Learning from Incomplete Data using the BCLA-MAP Model . . . 76
5.4 The scalar Simultaneous Truth and Performance Level Estimation (sSTAPLE) 76
5.4.1 Theory of sSTAPLE . . . 77
5.4.2 Learning from Incomplete Data using sSTAPLE . . . 80
5.5 Data Description and Methodology for Comparison . . . 81
5.6 Results . . . 85
5.6.1 Modelling of the Biases . . . 85
5.6.2 Simulated Dataset . . . 85
5.6.3 QT Dataset . . . 88
5.7 Summary . . . 93

6 A Fully-Bayesian BCLA Framework 96
6.1 Introduction . . . 96
6.2 Gibbs Sampling . . . 96
6.2.1 Theory . . . 97
6.2.2 Convergence Measures . . . 99
6.3 BCLA-Gibbs . . . 101
6.3.1 Learning from Incomplete Data using BCLA-Gibbs . . . 103
6.4 Data Description and Methodology of Comparison . . . 104
6.5 Results . . . 106
6.5.1 Convergence of the BCLA-Gibbs Model . . . 106
6.5.2 Simulated Dataset . . . 112
6.5.3 QT Dataset . . . 113
6.5.4 Capnobase RR Dataset . . . 119
6.6 Summary and Future Work . . . 126

7 BCLA with Signal Quality Extension 128

7.1 Introduction . . . 128
7.2 Background . . . 129
7.3 BCLA with Signal Quality (BCLAs) . . . 130
7.3.1 The Maximum-a-Posteriori (BCLAs-MAP) Approach . . . 131
7.3.2 The Gibbs Sampler (BCLAs-Gibbs) Approach . . . 134
7.4 Data Description and Methodology of Comparison . . . 135
7.5 Results . . . 138
7.5.1 QT Dataset . . . 138
7.5.2 Capnobase RR Dataset . . . 146
7.6 Summary and Future Work . . . 147

8 BCLA with Correlated Labellers 149

8.1 Introduction . . . 149
8.2 A Generative Model of Correlated Annotators . . . 150
8.2.1 The $\mathscr{M}_\Sigma$ Model . . . 150
8.2.2 The $\mathscr{M}_\rho$ Model . . . 152
8.2.3 Simulated QT Datasets with Correlated Annotators . . . 155
8.3 BCLA with Correlated Annotators (BCLAc) . . . 158
8.3.1 Gibbs Sampling for BCLAc with $\mathscr{M}_\Sigma$ (BCLAc-Gibbs) . . . 158
8.3.2 Background on Metropolis Hastings Algorithm . . . 160
8.3.3 The Metropolis Hastings Algorithm for BCLAc with $\mathscr{M}_\rho$ (BCLAc-MH) . . . 163
8.4 Data Description and Methodology of Comparison . . . 166
8.5 Results . . . 167
8.5.1 Simulated Datasets . . . 167
8.5.2 QT Dataset . . . 171
8.5.3 Capnobase RR Dataset . . . 176

8.6 Summary and Future Work 178

9 Conclusions and Future Work 180

9.1 Summary of Results 180

9.2 Future Work 187

References 192

Appendix A Expectation-Maximisation 210

A.1 Expectation-Maximisation of Maximum Likelihood 210

A.2 Expectation-Maximisation of Maximum-a-Posteriori 213

Appendix B Hyperparameter of the Conjugate Prior 215

B.1 Univariate Gaussian 215

B.1.1 Inferring the Posterior distribution of the Mean with Known Precision 216

B.1.2 Inferring the Posterior distribution of the Precision with Known Mean 218

B.2 Multivariate Gaussian 219

B.2.1 Inferring the Posterior distribution with Unknown Covariance and Mean 219

Appendix C Additional Results 223

C.1 Results of the BCLAs Model 223

C.2 Results of the BCLAc Model 228

List of figures

1.1 Example of a QT annotation in the electrocardiogram 3
1.2 Example of a 32s window of PPG used for RR estimation 9

2.1 Example of bias in the context of electrocardiogram QT interval labelling . 23
2.2 Examples of the ML and the MAP approaches 28

3.1 Example of Q and T annotations for a recording in the PCinC QT dataset . 31
3.2 Histograms of the QT annotations for all entries in the PCinC QT dataset . 33
3.3 Histograms of the percentage of QT recordings annotated per individual annotator for different divisions in the PCinC QT dataset 34
3.4 Histograms of the percentage of annotators per individual recording for different divisions in the PCinC QT dataset 35
3.5 Histogram and pie chart of the Capnobase RR dataset 36
3.6 Example of RR estimation using FFT and AR approaches 38

4.1 Graphical representation of the EM-R model 42
4.2 Example of a 5s noisy ECG segment of a patient with myocardial infarction 47
4.3 Pseudo code for the EM-R algorithm with features 51
4.4 Mean and standard error of RMSEs using EM-R with the bHRSQIs features, median and mean voting for different divisions in the PCinC QT dataset . . 56
4.5 Bland-Altman plots of the QT differences in the PCinC QT dataset Division 1 58

4.6 Bland-Altman plots of the QT differences in the PCinC QT dataset Division 4 59

5.1 Graphical representation of the ground truth model 67
5.2 Graphical representation of the annotator model 69
5.3 Graphical representation of the BCLA model 70
5.4 Simulated QT dataset and its distribution 71
5.5 Example of estimating the maxima of the precision distribution using GEVD 75
5.6 Graphical representation of the sSTAPLE model 77
5.7 Box plots of the error between the generated and *true* annotations for each of the simulated dataset for different mean bias values 83
5.8 A comparison of the simulated and inferred σ and bias of each annotator in the simulated dataset for different mean bias values 86
5.9 A comparison of the simulated and inferred σ and bias of each annotator in the simulated dataset [I] 87
5.10 A comparison of the reference and inferred σ and bias of each annotator for different divisions in the PCinC QT dataset [I] 89
5.11 The errors in annotations for Division 3 in the PCinC QT dataset 90
5.12 Mean and standard deviation of RMSEs of using different voting approaches as a function of the number of automated annotators [I] 94
5.13 The difference in RMSEs between BCLA-MAP and EM-R as a function of the number of automated annotators 95

6.1 An example of a 32s PPG window used for RR estimation and fusion using the BCLA model 105
6.2 Plot of samples of the parameters using the BCLA-Gibbs model 107
6.3 Convergence investigation of the BCLA-Gibbs model 108
6.4 Example of the sample draws of the posterior mean of the biases in the BCLA-Gibbs model for Division 4 in the PCinC QT dataset 110

6.5 Sample draws of the biases in the BCLA-Gibbs model 111
6.6 A comparison of the simulated and the inferred σ and bias of each annotator in the simulated dataset [II] 113
6.7 A comparison of the reference and inferred σ and bias of each annotator for different divisions in the PCinC QT dataset [II] 115
6.8 Mean and standard deviation of RMSEs of using different voting approaches as a function of the number of automated annotators [II] 118
6.9 The difference in RMSEs between BCLA-Gibbs and BCLA-MAP as a function of the number of automated annotators 119
6.10 Example of sample draws of the posterior mean of biases using the BCLA-Gibbs model for one subject in the Capnobase RR dataset 121
6.11 Differences in MAE between "Smart Fusion" and the BCLA approaches for 42 subjects 123
6.12 Box plot of MAEs across 42 subjects for different fusion approaches [I] . . . 124

7.1 Graphical representation of the BCLAs model 132
7.2 Example of errors and distribution plots for the PCinC QT dataset 139
7.3 The lower bound T on the bSQI results for the PCinC QT dataset 140
7.4 A comparison of the PCinC QT dataset reference, and inferred σ and bias of each annotator using the BCLAs model 142
7.5 Mean and standard deviation of RMSEs as a function of the number of annotators using the BCLAs model 145

8.1 Graphical representation of the BCLAc with $\mathscr{M}_\Sigma$ model 151
8.2 Graphical representation of the BCLAc with $\mathscr{M}_\rho$ model 153
8.3 pdf of a Beta distribution for modelling the correlation matrix 154
8.4 Simulated correlation of errors of the BCLAc with $\mathscr{M}_\Sigma$ Model 156
8.5 Simulated correlation of errors of the BCLAc with $\mathscr{M}_\rho$ Model 157

8.6 Bar plots of RMSE using different strategies for simulated datasets 168
8.7 A comparison of the simulated and the BCLAc inferred σ and bias of five annotators in the simulated datasets 168
8.8 A comparison of the simulated and the BCLAc inferred correlation of errors among five annotators in the simulated datasets 169
8.9 A comparison of the PCinCnew QT reference and inferred σ and bias of each annotator for different divisions 172
8.10 A comparison of the reference and the BCLAc inferred correlation of errors among 55 algorithms for Division 4 in the PCinCnew QT dataset 173
8.11 Box plots of RMSEs of using different voting approaches as a function of the number of automated annotators for PCinCnew QT dataset 174
8.12 A comparison of the reference and the BCLAc inferred correlation of errors among seven and nine annotators in the PCinCnew QT datasets 175
8.13 Box plot of MAEs across 42 subjects for different fusion approaches [II] . . . 177

C.1 The pSQI values for the Capnobase RR dataset 223
C.2 The difference in RMSEs between BCLA and BCLAs with the bHR feature versus No features as a function of the number of automated annotators . . 225
C.3 The difference in RMSE distribution between BCLA and BCLAs with the bHR feature as a function of the number of automated annotators 227
C.4 A comparison of the simulated and the BCLAc inferred σ and bias of 10 annotators in the simulated datasets 228
C.5 A comparison of the simulated and the BCLAc inferred correlation of errors among 10 annotators in the simulated datasets 229

List of tables

2.1 Survey of methodologies for fusing annotations to infer latent ground truth. 25

3.1 Summary of the diagnostic conditions of subjects in the PTBDB 31
3.2 Performance by participants for each division of the PCinC QT dataset . . . 32

4.1 RMSEs and MAEs for different voting strategies in the PCinC QT dataset . 53
4.2 RMSEs and MAEs of using three and nine manual/automated annotators for different voting strategies 54

5.1 The parameters of the BCLA model for modelling the PCinC QT dataset . . 82
5.2 RMSEs and MAEs of the inferred labels using different approaches in the simulated dataset $^{\mathrm{I}}$ 88
5.3 RMSEs and MAEs of inferred labels using different voting approaches in the PCinC QT dataset $^{\mathrm{I}}$ 91

6.1 RMSEs and MAEs of the inferred labels using different strategies in the simulated dataset $^{\mathrm{II}}$ 112
6.2 RMSEs and MAEs of inferred labels using different voting approaches in the PCinC QT dataset $^{\mathrm{II}}$ 116
6.3 The parameters of BCLA for modelling the Capnobase RR dataset 120
6.4 Comparison of MAE and standard error across 42 subjects in the Capnobase RR dataset 122

7.1 Optimisation of the lower bound T of the bSQI values for each division of the PCinC QT dataset. 141
7.2 RMSEs of inferred labels using different voting approaches with and without features in the PCinC QT dataset. 143
7.3 RMSEs (ms) of using three to eleven algorithms for different voting strategies. 146

8.1 The parameters of BCLAc for modelling the PCinCnew QT dataset. 171
8.2 RMSEs of the inferred labels using different strategies in the PCinCnew QT dataset. 173
8.3 The parameters of BCLAc for modelling the Capnobase RR dataset 176

C.1 MAEs of inferred labels using different voting approaches with and without features in the PCinC QT dataset. 224

Nomenclature

Acronyms / Abbreviations

AM	Amplitude Modulation
BCLA	Bayesian Continuous-Valued Label Aggregator
BCLAc with $\mathscr{M}_\rho$	BCLA with correlated annotators for modelling correlation
BCLAc with $\mathscr{M}_\Sigma$	BCLA with correlated annotators for modelling covariance
BCLAc	Bayesian Continuous-Valued Label Aggregator with Correlated Annotators
BCLAs	Bayesian Continuous-Valued Label Aggregator with Signal Quality Extension
BW	Baseline Wander
CVD	Cardiovascular Disease
ECG	Electrocardiogram
EM-R	Regression using Expectation Maximization proposed by Raykar *et al*.[89]
EM	Expectation Maximisation
FM	Frequency Modulation
GEVD	Generalised Extreme Value Distribution
MAP	Maximum-a-Posteriori
PCinC	PhysioNet/Computing in Cardiology
PPG	Photoplethysmography
PTBDB	Physikalisch-Technische Bundesanstalt Diagnostic Database
RR	Respiratory Rate
sSTAPLE	scalar Simultaneous Truth and Performance Level Estimation proposed by Warfield *et al*. [3]

Chapter 1

Introduction

1.1 Clinical Motivation

Expert labelling is the gold standard for diagnosing patient-specific diseases from medical data. For the purpose of this thesis, expert labelling is defined as being the result of experienced physicians annotating an area of interest in some data when the ground truth is not readily available. However, experts are relatively scarce, their time is expensive, and the task of labelling is time-consuming [1–3]. Expert labelling can be further complicated by the ambiguous definition of what it means to be an "expert". There is no standardisation for testing measures of clinical competency above a proficient level, both in terms of accuracy and consistency, and no empirical method for measuring levels of clinical expertise, even though the quality of annotation heavily relies on the expert's experience. Therefore, large inter- and intra- annotator variabilities occur among physicians, depending on their respective experience and level of training [2–9]. The "experts" could be either humans (such as trained clinicians), or automated algorithms, or some combination of both. Manual labels refer to annotations provided by human annotators, while automated labels may be estimates made/predicted by automated algorithms. This thesis will focus on the medical datasets where subjective continuous labels of some presumed underlying ground truth are

provided (where the desired ground truth is not readily available in clinical practice), and we seek to combine their outputs in a principled manner. To demonstrate the importance of formulating a consensus where the ground truth is not available, two commonly-encountered and clinically-important examples are described.

1.1.1 Exemplar I: QT Interval in the Electrocardiogram

Cardiovascular disease (CVD) is one of the leading causes of global mortality. The World Health Organisation estimated 17.5 million people died from CVD in 2012, comprising 31% of deaths worldwide [10]. Over three quarters of CVD deaths have taken place in low- and middle-income countries, and a further increase to 26.4 million deaths per year is predicted by 2030 [11].

The electrocardiogram (ECG) is a standard and powerful tool for assessing cardiovascular health as many detrimental heart conditions manifest as abnormalities in the ECG. Evidence-based literature on optimal methodologies for standardising the learning of, and testing of competency in ECG interpretation remains scarce [2]. Disagreements in ECG diagnostic annotations may be due to intrinsic difficulties in interpreting the signals that are linked to the level of training or experience of the annotators [2]. Disagreements may be exacerbated by significant amounts of noise such as motion artefacts, electrode contact noise, and baseline drift [12].

The QT interval is one particular measure of ECG morphology, which refers to the elapsed time between the onset of QRS complex and the T wave offset as it returns to the isoelectric line (see Fig. 1.1). The QT interval is a marker for ventricular depolarisation and repolarisation of the cardiac muscle cells [12]. The QT interval is sometimes adjusted for varying heart rates, commonly referred to as the "QT corrected" (QTc) interval. For the past two decades, the use of the QT interval as a biomarker has been intensively studied. It has been shown that shortening of the foetal QT interval measured using a scalp electrode is

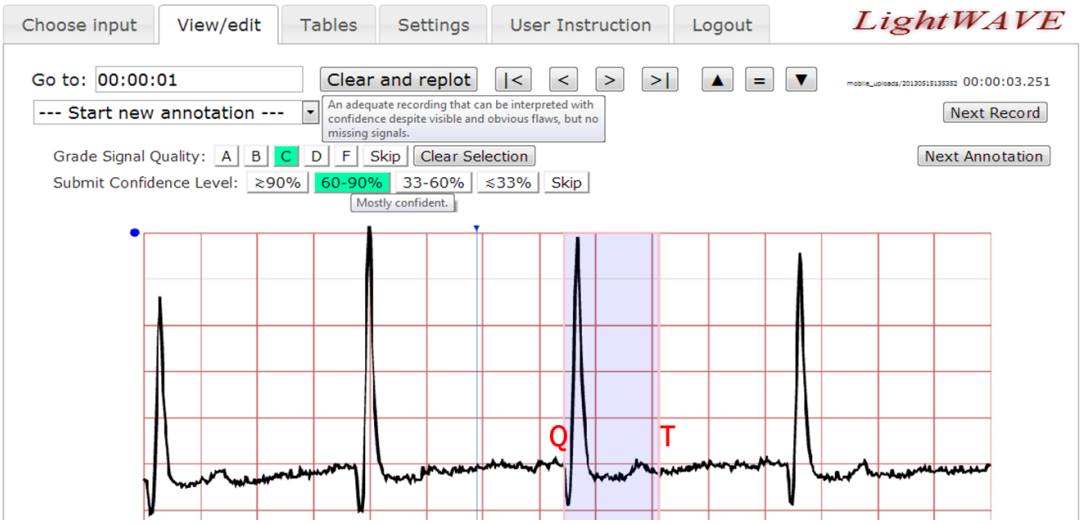


Fig. 1.1 Example of a QT annotation in Electrocardiogram on the CrowdLabel interface [13]. The shaded blue rectangle represents a QT annotation for a single beat, which starts at the beginning of the Q wave and terminates at the end of the T wave.

associated with intrapartum hypoxia resulting in metabolic acidosis [14]. Prolongation of the adult QT interval serves as an important risk factor for hypoglycaemia [15, 16], hypokalaemia [16–18], and hypomagnesaemia [17], which predisposes the individual to arrhythmias and sudden cardiac death [1, 19]. In particular, QT prolongation has been found in individuals after taking drugs that cause adverse effects including arrhythmia [19–21].

According to guidelines provided by the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH), thorough clinical studies need to be performed to evaluate QT prolongation for candidate drugs [1]. However, clinical studies frequently incur high costs related to ECG processing and measurement, particularly in the case of estimating QT intervals, because this process requires segmentation and analysis to be performed by experts [1]. Studies have reported significant intra- and inter-observer variability in QT annotations, ranging from 10 to 30 ms, where QT interval is usually around 400 ms [22, 23]. One of the challenges in measuring QT intervals is to identify correctly the end of the T wave [24–26]. Morphological abnormalities in the T wave may make annotations less accurate [25]. Manual annotators also appear to underestimate the true T wave end-point, due to various T wave morphologies and the

presence of different sources of noise [27]. Even for less ambiguous quantities estimated from the ECG (such as R-R intervals), intra- and inter-observer variability can still be as high as 30 ms [23]. In the context of foetal QT annotation, for which baseline wander is often evident, it was observed that manual annotators consistently over- or under-estimated the QT interval [13]. Al-Khatib *et al.* [28] performed a survey to assess healthcare practitioners' abilities to correctly measure the QT interval within a pre-specified measurement error of 20 ms from the actual value, and found only 43% of participants were successful. Viskin *et al.* [29] presented the ECGs recorded from two patients with long QT syndrome (LQTS) and from two healthy females to 902 physicians (25 QT experts who had published on the subject, 106 arrhythmia specialists, 329 cardiologists, and 442 noncardiologists) from 12 countries. No other details were given concerning the level of the training or the intrinsic accuracy of these annotators. For patients with LQTS, 80% of arrhythmia specialists calculated the QTc correctly, but only 50% of cardiologists and 40% of noncardiologists did so. Inter-observer agreement was assessed using Cohen's kappa coefficient, κ [30]: this was deemed to be "excellent" among QT experts ($\kappa = 0.82$), "moderate" among arrhythmia specialists ($\kappa = 0.44$), and "low" for cardiologists and noncardiologists ($\kappa < 0.3$ for both groups). Furthermore, the correct classification of all QT intervals as either "long" or "normal" was achieved by 96% of QT experts and 62% of arrhythmia specialists, but by less than 25% of cardiologists and noncardiologists who are otherwise considered to be experts. A follow-up study was conducted by Postema *et al.* [26] to assess the possibility of teaching non-experts to interpret QT intervals correctly: 71 medical students were trained and tested with four ECGs (which were the same as those from the study of Viskin *et al.* [29]). Results indicated that accurate QT interval interpretations were achieved by 77% of students, as compared with 62% by the arrhythmia specialists and less than 25% by the cardiologists and noncardiologists.

While automated algorithms offer advantages of time efficiency, repeatability, and cost-saving benefits compared to manual methods, there remains a substantial discrepancy in

their estimation of the QT interval that limits their use [31–36]. As there is no guidance for regulating automated algorithms in this context, it is difficult to assess the acceptability of automated annotation algorithms. The "PhysioNet/Computing in Cardiology 2006 Challenge" was based on competing algorithms estimating QT intervals from a publicly-available ECG database, with the best automated algorithm achieving a root-mean-squared error of 16.34 ms [37]. The "Meta-6" algorithm proposed by Moody [37] combined the three best-performing annotators from both the open-source and closed-source algorithms entered into the competition by taking a median voting strategy. However, the development of this algorithm required selecting annotators based upon a known ground truth, making this approach infeasible for data where the desired ground truth is not readily available (as is typically the case in clinical practice). Kligfield *et al.* [33] compared two commercial automated annotation systems and found less than 2 ms mean difference between them, but 12 ms standard deviation in their QTc measurements, which was a result of large variation in their measurements.

1.1.2 Exemplar II: Respiratory Rate Estimation with the Photoplethysmogram

The photoplethysmogram (PPG) is a waveform that is obtained from an optical technique based on the differential absorption of red and infra-red light by the blood, and is used to quantify blood-volume changes and blood oxygenation in biological tissue [38]. This is typically performed using a pulse oximeter, which is a non-invasive device that has been widely used to measure blood oxygen saturation and heart rate [39]. The PPG waveform is composed of AC and DC components: the former is due to the absorbance of light caused by pulsatile constituent of the arterial blood, whereas the latter is the absorbance of light due to non-pulsed blood and tissue [40].

Respiratory rate (RR) is an important physiological variable measured in a wide range of clinical scenarios, because it provides valuable diagnostic and prognostic information

[41]. Extreme values of RR can be associated with increased risks of adverse episodes of patients in hospitals [42–45]. A clinical diagnosis of pneumonia is made for children aged 1 to 5 years when their RR exceeds 40 breaths-per-minute (bpm) [46]. Any adult with a RR over 24 bpm is likely to be critically ill [42, 47]. Furthermore, RR is used to assist in the diagnosis diseases such as sepsis [48], hypercarbia [49], pneumonia in adults [50], and sudden unexpected death in epilepsy [51]. Measurement of RR is commonly performed by manually counting the chest movements of the individual over a period of approximately 30 - 60s [52]. This practice is time-consuming and prone to large inter- and intra-variation in RR estimates [53–57]. Typically, the number of breaths occurring within 15s is counted, and then multiplied by four to give an estimate of RR in bpm, which is therefore quantised to the nearest 4 bpm. This approach is particularly problematic with younger children, who have fast breathing rates [58] and where the potential for error is substantial. Lovett *et al*. [55] compared the accuracy in the RR estimates made by triage nurses and via transthoracic impedance plethysmography (TIP) in 159 patients (age ranges from 18 to 89), where the latter is the clinical standard for non-invasive monitoring. The authors found poor RR agreement between both sets of estimates and with the reference RR estimated using the gold-standard measurements (i.e., auscultation for one minute and observation for one minute [52]). The 95% limits of agreement in a Bland-Altman plot between the triage nurses' and the reference measurements of RR was -8.6 to 9.5 bpm, and was -9.9 to 7.5 bpm between the TIP and the reference. Gupta *et al*. [59] had also found that nurses in general medicine wards have systematic bias in reporting RR to be around 20 bpm. Furthermore, despite the importance of monitoring RR as one of the vital signs, for monitoring life-threatening conditions of patients, Cook and Smith [60] had found only four out of the 30 medical textbooks that they examined emphasised the importance of measuring RR. The applications of RR monitoring ranged from hospital to home-based monitoring, which are detailed in the following sub-sections.

A. Hospital Inpatient Monitoring

Early warning scores [61, 62] (EWS) are commonly used across hospitals and clinics in the UK to track patients' health conditions using observations of physiological data. Vital signs, including RR, are routinely observed throughout a patient's stay in most hospital wards, and scores are assigned to these observed values [63]. The scores for each physiological variable are summed, and the total is compared to a pre-defined threshold [64, 65]; when the threshold is exceeded, the physiological deterioration is deemed to have occured, and care is escalated. Although several studies have attempted to automate the use of the EWS systems [63, 66, 67], little research has been undertaken to improve the accuracy of the manual observations of the vital signs subsequently used by the EWS system. Furthermore, RR was identified as the one of the vital signs which nursing staff recorded less than 50% of the time despite its perceived importance [68]. Continuous patient monitoring using automated systems has the potential to improve on the use of manual observations, and thereby increase the potential for giving early warning of patient deterioration. The RR estimate obtained from impedance plethysmography via ECG sensors is often considered to be inaccurate [55], and an alternative solution using less obtrusive sensors may be to estimate RR from the PPG via a pulse oximeter.

B. Mobile Health Monitoring

Mobile health ("m-health") has received growing interest in providing point-of-care diagnostics and monitoring using wearable sensors. In some applications, these sensors may be linked to a smartphone; in others, the sensors may be linked by a so-called "body area network" [69]. Patients with chronic diseases such as asthma, diabetes, hypertension, chronic-obstructive pulmonary disease, and the like can benefit from using such systems for management of their condition [41, 70, 71]. Although wearable sensors such as the pulse oximeter have been a

popular choice for monitoring heart rate and SpO_2 [72, 73], there is a lack of robust methods for estimating of RR using the PPG waveform.

C. Non-contact Monitoring

The concept of measuring PPG signals using a camera positioned some metres away from the subject has been explored over the past decade [74–76]. More recent studies have demonstrated that contactless sensing of heart rate and RR is possible using data acquired from video cameras [77–79]. However, the video-based PPG technique for RR estimation commonly relies of the tracking of the small regions of the exposed skin of a subject, which produces a substantial challenge in them removing noise, such as ambient light interference and movement artefact. Robust methods for estimating the RR are therefore required, to improve upon the current state-of-the-art (using a autoregression model and this is described in more detail in Chapter 3).

RR Estimation

Although poorly understood, several different physiological mechanisms such as respiratory activity cause modulation of the PPG [78, 80, 81], which result in amplitude modulation (AM), baseline wander (BW), and modulation of beat-to-beat intervals (frequency modulation, FM) (see Fig. 1.2). The AM describes the change in the amplitude of the PPG signal with respiration, and is defined as the difference between successive peaks and troughs of the waveform. It is produced due to variation of cardiac stroke volume over the respiratory cycle [82, 83]. BW is the respiratory-induced intensity variation, which is estimated from the time-series of peak amplitudes in the PPG waveform. BW is caused by the intra-thoracic pressure variation which prompts changes in the baseline of the perfusion of peripheral tissues [81, 83, 84]. The FM is also named respiratory sinus arrhythmia, where respiration induces a change in heart rate: the heart rate increases during inspiration and decreases

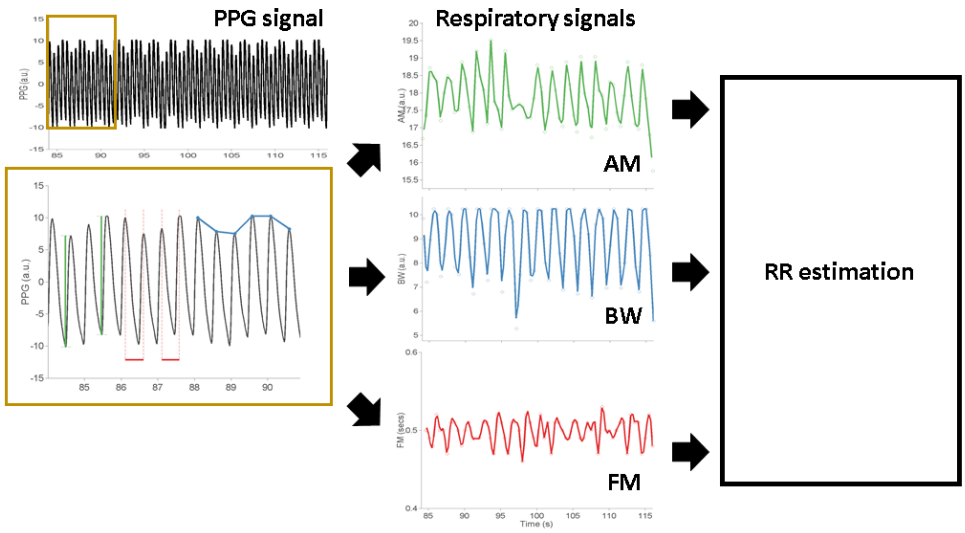


Fig. 1.2 Example of a 32s window of PPG used for RR estimation. The amplitude modulation (AM in green), baseline wander (BW in blue), and modulation of beat-to-beat intervals (FM in red) are extracted from the PPG.

during expiration, and thus this corresponds to frequency modulation in the PPG waveform [85].

Recently proposed methods for extracting RR from segments of PPG rely on estimating RR independently from each modulation (the derived respiratory signals), and fusing the three estimates into a final RR value [81, 86]. However, the common "fusion" approaches employed in such settings are typically very straightforward (and result in poor overall fusion results), and which tend to reduce substantially the number of windows for which an estimate of RR is computed. Some other studies have used more complex approaches to select or combine the three modulations: Johansson [87] combined the modulation measurements using neural networks to attempt to improve the accuracy in RR estimation; Leonard *et al.* [88] compared the time-series of the FM, the AM, and primary wavelet decomposition from 22 subjects, then manually selected the "optimal" waveform for providing the final RR estimate for each individual; Addison *et al.* [80] derived the three modulations using continuous wavelet decompositions, and the final RR estimate was computed through weighted averaging and a logical decision-making process. To date, there is no obvious optimal approach for combining

the three modulations. To make matters more challenging, these modulations are not always observed in all patients; i.e., some patients may exhibit only one or two types of modulation depending on their age and health [80]. Therefore, a robust framework is needed to combine these modulations in a principled manner, rather than using existing heuristic methods that are insufficiently robust for use in practice.

1.2 Outline of the Thesis

This thesis presents a series of methods for inferring a consensus in a scenario when only noisy annotations of some presumed underlying ground truth are provided. We will show that these methods can produce reliable and accurate labels for our two exemplary medical applications, focusing on aggregating the outputs of a group of mixed imperfect automated algorithms (the "experts" or "annotators") in an unsupervised manner. For the purpose of this thesis, each annotator is assumed to have a bias and a precision value. The precision value can be further represented in term of standard deviation of annotations from an annotator, which describes its variance of error in annotations around the underlying ground truth.

Chapter 2 provides a detailed survey of the works that related to the models proposed in the thesis. It reviews relevant models that aim to fuse noisy labels to infer the underlying true annotation via unsupervised learning.

Chapter 3 describes the two publicly-available datasets considered in the thesis, namely the 2006 PhysioNet/Computing in Cardiology Challenge QT dataset, and the Capnobase respiratory rate dataset. Both datasets are used to validate the models proposed in the thesis.

Chapter 4 investigates an existing maximum likelihood approach proposed by Raykar *et al.* [89] using the expectation-maximisation (denoted EM-R hereafter) for estimating consensus labels when only the noisy versions of the unknown true labels are available. Contextual features are introduced in this chapter to further improve the performance of

EM-R, and the EM-R algorithm subsequently serves as a benchmarking algorithm throughout the thesis, representing the existing state-of-art.

Chapter 5 proposes the novel Bayesian Continuous-Valued Label Aggregator (BCLA), which is a generative model that aggregates medical labels in an unsupervised manner, to estimate jointly the assumed bias and precision of each annotator. The method is then used to infer the underlying ground truth. This chapter also investigates the effect of the bias of individual annotators. In particular, the scalar version of the Simultaneous Truth and Performance Level Estimation (sSTAPLE) proposed by Warfield *et al.* [3] is described, and the latter serves as a second benchmarking approach for inferring the underlying true labels. A maximum-a-posterior (MAP) approach for formulating the BCLA framework (denoted hereafter as the BCLA-MAP) with independent annotators was applied to QT interval estimation using labels from a simulated dataset and from the QT dataset. The performance of the BCLA-MAP approach is compared to results from mean and median voting, as well as the two benchmarks: sSTAPLE and EM-R.

Chapter 6 introduces the Gibbs sampler for taking a fully-Bayesian approach for BCLA (denoted hereafter as BCLA-Gibbs) with independent annotators. It is applied to both the simulated and real QT datasets, as well as the Capnobase respiratory dataset. The performance of the BCLA-Gibbs method is compared to all methods considered in previous chapters.

Chapter 7 proposes an extension to BCLA with independent annotators incorporating the notion of signal quality (denoted hereafter as BCLAs); this involves a confidence measure relating to the quality of the input waveform. We illustrate the method using our two clinical exemplars.

Chapter 8 proposes the BCLA framework with correlated annotators (denoted hereafter as BCLAc). Two variations of the BCLAc framework are presented, where one involves modelling the correlation matrix between expert labels (denoted as BCLAc with $\mathscr{M}_\rho$), and

where the other involves modelling the covariance matrix between expert labels (denoted as BCLAc with $\mathscr{M}_{\Sigma}$). This is the most complex variant of BCLA, and represents the case that typically occurs in practice in which subsets of experts may exhibit some correlation. In the case of human experts, this could correspond to various levels of experiences; in the case in which the "experts" are automated algorithms (as in considered in this thesis), this could correspond to different types of modelling approaches taken by these algorithms.

Chapter 9 provides a summary of the work presented in this thesis with highlights of the key contributions, and an overview of potential future work.

Chapter 2

Background

"Essentially, all models are wrong, but some are useful."

– George E. P. Box

2.1 Introduction

In the healthcare domain, there are several approaches for fusing clinical information for diagnosing chronic or rare diseases when expert clinicians are not readily available [90–92]. Most of these approaches focus on providing a platform for exchanging hypotheses concerning the diagnosis, or which act as an intelligent search engine for querying relevant medical information. However, these tools face the common difficulty in deciding which diagnosis is probably appropriate when there is no ground truth. This raises a question about current medical practice as to whether simply majority consensus between candidate automated diagnoses is the best way to achieve a result, or if some other approach is better. In the context of labelling medical data, sole reliance on results from automated algorithms to produce a diagnosis would not be sufficient due to large inter- and intra-variabilities between recommendations from such algorithms. Furthermore, when the ground truth is

unknown, it is difficult to choose which algorithms to trust or discard, or even how to merge their recommendations to form a diagnostic output. Therefore, this chapter reviews related works for the use of voting algorithms in an unsupervised manner when ground truth is not readily available, and any relevant theoretical background that is necessary for the framework described by this thesis.

2.2 Related Works

2.2.1 Ensemble Learning

Boosting [93] is a popular technique used for combining the outputs of multiple classifiers, with the aim of producing a prediction that is better than any of the individual classifiers. This method involves resampling a subset of training data in a strategic manner, such that a more informative subset of the training data is used for each consecutive classifier. Similar to boosting is bagging [94] (i.e., bootstrap aggregating), which also trains different classifiers with different training sets that are bootstrapped with replacement from the entire training dataset. However, the result is performed by majority or average voting of the outputs of the classifiers. The commonly-used AdaBoost algorithm is named after adaptive boosting [93], and creates an ensemble of classifiers by resampling the data. The classifiers trained in sequence are then combined by weighted majority voting, where weight corresponds to accuracy when classifying the training data. Essentially, each classifier is trained using a weighted form of the total training set, in which the weights depend on the performance of the previous classifier [95]. In addition to boosting and bagging, another commonly-used ensemble learning algorithms is stacking, or stacked generalisation [96]. This has two tiers of classifiers: the tier 1 classifiers are trained using bootstrapped training samples, and their outputs when applied to a held-out dataset are considered to be inputs to train a tier 2 meta-classifier. Stacking can be used to combine classifiers of different types.

These ensemble learning methodologies require pre-labelled data to either train classifiers or to estimate the weighting of each classifier. In the case when the ground truth is not readily available, such approaches are problematic because of the *chicken or the egg* causality dilemma. In addition to the acquiring cost of labelled data, it is undesirable to optimise the weights of each classifier using only a specific dataset, as its ability to generalise to unseen datasets cannot be guaranteed.

2.2.2 Unsupervised Learning

Majority voting is a simple form of aggregation, and takes little account of the accuracy or consistency of each annotator for the assigned task; it also has no consideration of the difficulty of the task. This voting strategy is used in some crowd-sourcing projects such as reCAPTCHA [97], which implements the CAPTCHA (an acronym for "Completely Automated Public Turing test to tell Computers and Humans Apart") algorithm using old printed material to allow humans to annotate words which were difficult to label automatically. When two automated algorithms differently translated text from an image, humans are asked to translate the word in return for access to a website (such as an email account). Each human translation counts as one vote, and each automated algorithm receives half a vote. Translations receiving a cumulative vote (from both humans and automated programs) of 2.5 or more become the accepted result since 67.87% of the words encountered required only two human responses to be considered correct. A slightly improved version to this simplistic aggregation is weighted majority voting [98], which combines annotations of all labellers with weights that vary according to their individual accuracies. However, accuracy cannot readily be estimated in many real cases since there is no available ground truth. In a study of combining non-expert annotators' labels of sleep spindle location, which is a pattern in human electroencephalography, Warby *et al.* [99] demonstrated that fusing annotations using majority voting could reach similar results as those provided by experts. In the situation

where a majority does not exist, other methods of forming aggregate labels, such as mean and median voting strategies are considered. In the context of fusing QT interval estimates, as introduced earlier the "Meta-6" algorithm was proposed by Moody [37] for the 2006 PhysioNet/Computing in Cardiology Challenge, and which used a median voting strategy. In a data-mining application, Sheng *et al.* [100] showed that accuracy of classifiers could be improved by using a repeated labelling technique. However, they made a strong assumption that all of the annotators were of equal skill level (i.e., the same weights were used for all annotators).

A more effective and less biased probabilistic approach, combining annotations using an expectation maximisation (EM) algorithm (see Chapter 4 for details), was first proposed by Dawid and Skene [101]. They applied EM to classify the unknown true states of health (e.g., "fit enough to undergo a general anaesthetic") for 45 patients, given the decisions made by five anaesthetists. They assumed that each anaesthetist could have different types of error rate, and that the error rate for an anaesthetist could be modelled by a confusion matrix. Each component in the confusion matrix was the probability that each anaesthetist declared a patient to be in a particular state with respect to the true state of the patient. The confusion matrix and the true state of the patients could then be inferred simultaneously using EM, which iteratively (i) estimates the true state of the patients based on the weighted votes of anaesthetists according to their competence derived from the confusion matrix; and (ii) re-estimates the confusion matrix for each anaesthetist based on the estimated true state of the patients from the previous step. Spiegelhalter and Stovin adopted Dawid and Skene's model for evaluating the uncertainty in biopsy specimens taken from the heart following cardiac transplantation [102]. Smyth *et al.* [103] applied the same EM algorithm for image labelling tasks to infer the underlying ground truth by fusing labels collected from four planetary geologists examining five types of volcano images.

Jin and Ghahramani [104] used a similar EM approach in an information-retrieval setting but considered that only one of the candidate labels could be correct. In the field of binary medical image segmentation, Warfield *et al.* [105] proposed a simultaneous-truth-and-performance-level-estimation (STAPLE) based on the EM approach, which took a collection of binary segmentations of an image, and then simultaneously computed a probabilistic estimate of the true segmentation and a measure of the performance level represented by each segmentation. Cholleti *et al.* [106] adapted the STAPLE algorithm by including a confidence-rated boosting [107] (which is a generalisation of AdaBoost [108]) for the experts.

A full Bayesian version of the approach proposed by Dawid and Skene [101] was developed by Carpenter [109] in 2008, focusing on a binary model with two categories. A general framework using the concept of confusion matrix proposed by Dawid and Skene [101] for Bayesian model combination was described by Kim and Ghahramani [110], which explicitly modelled the relationship between each classifier's output and the unknown true categorical labels. Whitehill *et al.* [111] also modelled the task difficulty as well as labeller accuracy in a probabilistic framework called GLAD ("Generative model of Labels, Abilities, and Difficulties") using EM. They performed experiments using Amazon's "Mechanical Turk" [112] for face images that contained smiles as either *Duchenne* or *Non-duchenne*[1], and crowdsourced 3,572 binary labels from 20 Mechanical Turk labellers for 160 images. The inferred labels were compared to the annotations provided by two certified experts in the Facial Action Coding System, and gave a 6% increase in performance over the results obtained using the majority voting approach.

More recently, Raykar *et al.* [113] proposed a Bayesian version of the EM algorithm for aggregating labels. Instead of using full Bayesian inference, they considered a point estimate for the mode of the posterior. They described a binary EM method applied to the area of computer-aided diagnosis for classifying malignant or benign breast tissue

[1] A *Duchenne* smile refers to an "enjoyment" smile through the activation of the orbicularis oculi muscle around the eyes while a *Non-uuchenne* smile means the "social" smile that does not cause muscle movement around the eyes [111].

from a medical image. A set of digital mammography and breast magnetic resonance images (MRI) were collected from a hospital for training the classifier, and where the gold standard was available via biopsies to evaluate the classifier's performance. Binary annotations representing malignant or benign breast tissue were provided by computer simulated radiologists (i.e., automated algorithms) and four real radiologists for digital mammograms and MRI respectively. Where the results of Raykar *et al.* [113] were compared with the majority voting approach, a 3% improvement was observed in estimating the ground truth of digital mammography as expressed as the area-under-the-receiver-operator-curve (AUROC). Additionally, a 6% improvement was observed when estimating labels for classifying breast tissue in MRI when using eight morphological features with a leave-one-out methodology. They also validated their algorithm for recognising textual entailment (a natural language processing task of deciding whether two given text fragments have similar meaning; i.e., the meaning of one fragment can be inferred from another) [89]. The authors crowd-sourced 164 distinct annotators from Mechanical Turk with 8,000 annotations, and found that their EM-based algorithm outperformed majority voting, achieving 5% higher accuracy when using at least 60 annotators and similar accuracy (90%) was achieved when using 160 annotators.

Yan *et al.* [114] extended Raykar's model of the underlying ground truth by incorporating a graph prior, as well as modelling each annotator's expertise as a function of input feature in different datasets. Without a prior on the ground truth, 225 binary labels (e.g., whether or not a region is malignant) of 75 mammograms were combined from three clinicians, incorporating eight morphological features to model the expertise of each annotator. The results showed a slight improvement over Raykar's model (where annotator's expertise was not a function of the input feature) with AUROC of 0.83 and 0.81 respectively. In the case of fusing 220 binary labels for detecting abnormalities in heart-wall motion in ultrasound videos (with labels provided by five cardiologists), Yan's model with a graph prior of the

ground truth outperformed Raykar's when there were large number of missing annotations (i.e., using 20% of the annotations). It showed almost equal performance to Raykar's model when using all annotations. Similar results were also observed by Yan *et al.* when fusing information extracted from electronic medical records, for determining whether or not a patient has a history of atrial fibrillation. A similar model of measuring annotator's expertise, the "Aggregating Experts and Filtering Novices" (AEFN) approach was proposed by Zhang *et al.* [115] for binary biomedical text classification and protein disorder prediction. Different from Yan's model, Zhang *et al.* considered that features were from different components in a Gaussian mixture, and that annotators had a different sensitivity and specificity for each mixture component. Furthermore, the AEFN algorithm eliminated annotations provided by novice labellers (those with low sensitivities and specificities) prior to inferring the consensus. A 3% improvement in AUROC was observed using the AEFN algorithm when compared to Raykar's model [113]. Audhkhasi and Narayanan [116] presented a similar work, modelling annotator reliability as a variable over the complete input feature space, and assuming reliability was constant over each mixture component.

An approach for inferring the accuracy of experts while simultaneously training a classifier for prediction using a full Bayesian model was proposed by Johnson [117]. This used ordinal categorical data for the application of creating an automated essay grader. A total of 473 essays categorised into six levels were graded by six human raters, a generative model was then fitted to the data, and stochastic sampling was performed to infer the values of the model parameters. Although the task-specific biases of each rater was described in the model, they were assumed to be identically zero and hence disregarded when performing the model inference. The resulting classifier outperformed three of the human raters and was nearly as reliable as the remaining three.

Simpson *et al.* [118] proposed a variant of the independent Bayesian classifier combination method to infer the underlying true categorical labels by modelling a confusion matrix

(a similar form as described by Dawid and Skene [101]) that quantified the decision-making abilities of each base classifier. They proposed using variational Bayesian inference (i.e., a Bayesian generalisation of the EM algorithm) to compute the parameters of interest, and validated their model in a dataset collected from the Galaxy Zoo supernovae citizen science crowdsourcing project. From 1,705 base classifiers (i.e., volunteer citizen scientists) for classifying 963 images as either "very unlikely to be a supernova", "possibly a supernova", or " very likely a supernova", their proposed model outperformed the majority, weighted majority, weighted sum, and mean voting approaches (with an AUROC of 0.9840 versus 0.7809, 0.7378, 0.7722, and 0.7543, respectively). In addition, Simpson *et al*. presented a novel approach to cluster classifiers based on their inferred confusion matrices, and found five groupings of classifiers, each had a distinct approach to decision making. Furthermore, a dynamic version of their proposed model (i.e., modelling each classifier-specific confusion matrix changes over time) was detailed in a separate publication [119].

Chittaranjan *et al.* [120] proposed a two-stage probabilistic model for estimating the binary dominance level of participants, in small-group conversations using non-verbal audio and visual cues: (1) using Raykar's binary model [113] with raters' self-reported confidences as features to infer the ground truth; and (2) the inferred ground truth was used to train a classifier in a supervised manner, using audio-visual features as inputs to identify whether or not a participant in a meeting is dominant. The results of their proposed model outperformed the majority voting approach. Extending the concept of self-reported confidence scores, Oyama *et al.* [121] proposed integrating the confidence score into the Dawid and Skene model [101], and further created two probabilistic models to infer the unknown ground truth from noisy labellers using the EM algorithm. Each annotator was assumed to have a distinct conditional distribution for his confidence score given the true label and his annotations, and annotators were assumed to have either an independent or dependent confidence model. Annotations were provided by 100 annotators for labelling 10 images for identifying the

correct bird name when presented with a binary choice. The inferred results had demonstrated that only five or 10 annotators were required respectively to allow the two proposed models to surpass the performance of the majority voting approach, as well as that of the Dawid and Skene model. Furthermore, the choice for modelling the annotators' confidences as a dependent or independent factor in the model was task-specific.

Instead of modelling the parameters of interest (e.g., an annotator's expertise and model parameters for training a classifier), and then jointly inferring the underlying true annotations, Kajino *et al.* [122] proposed a "personal" classifier for each of the annotators and estimated the base classifier by relating it to the personal models without estimating the true labels. In a study of identifying whether or not a token in texts was a named entity, Kajino *et al.* fused annotations from 135 annotators for the 120,968 instances in the training set, and compared the performance of their model to that of Raykar's [113]. The proposed model had sub-optimal positive predictive value (0.651 versus 0.761) but higher sensitivity (0.766 versus 0.553) than Raykar's model. A similar model for building the base classifier was also proposed by Valizadegan *et al.* [123] for binary labels, which was extended for the use with categorical labels [124], where each personal classifier had an annotator-specific precision (whereas Kajino *et al.* assumed constant precision for all annotators in their model).

2.2.3 Modelling Annotator Bias and Precision

The probabilistic approaches mentioned previously focused on modelling the case with binary or categorical labels, which is not applicable to the continuous-valued labels that are commonly considered in many biomedical applications, and the subject of this thesis. Furthermore, empirical studies have shown that both human or automated annotators are likely to have their own bias values regardless of their "expertise" [3, 13, 125–127]. The bias of each annotator is defined as being the opposite of accuracy, where the latter is the average

difference between the estimation and the true value. An example of bias is demonstrated in Fig. 2.1 in the context of electrocardiogram labelling.

There exist some key works in literature for modelling continuous-valued labels, and the biases and expertise of each annotator, which are described below.

Raykar *et al.* [89] extended their binary EM-based method to fuse continuous-valued labels for measuring the diameter of a suspicious lesion from a medical image using a regression model. They considered that the discrepancies in the lesion diameter from different annotators were Gaussian-distributed and that they were noisy versions of the actual true diameter. Their proposed method jointly estimates the precision of each annotator and infers the underlying true annotation, as well as learning the parameter of the regressor which estimated the ground truth. In a study of automated colorectal polyp measurements, five annotators were simulated and their annotations were created for the 393 examples in the training set, and where the reference gold-standard annotations were provided by radiologists. The reference polyp diameters in the test set of 397 examples were compared with the diameters predicted by the regressor learnt using (a) the reference, (b) the proposed algorithm, and (c) the average of the simulated annotations. The regressor constructed using the proposed algorithm was almost as good as that obtained using the reference in the test set, and both models exhibited similar correlation coefficients (0.706 versus 0.715, respectively). This was found to be significantly better than the performance seen when using the mean voting of all annotations (correlation of 0.555). The root mean-squared-error of the proposed algorithm across the test set was 1.99mm, compared to that of the reference (1.97mm), and the mean voting strategy (2.89mm).

Welinder and Perona [126] devised a Bayesian EM framework for fusing binary, multi-valued, and continuous-valued labels, which explicitly modelled the precision of each annotator to account for their varying skill levels, without modelling their bias values. They created an online implementation of the model where the proposed algorithm actively asked

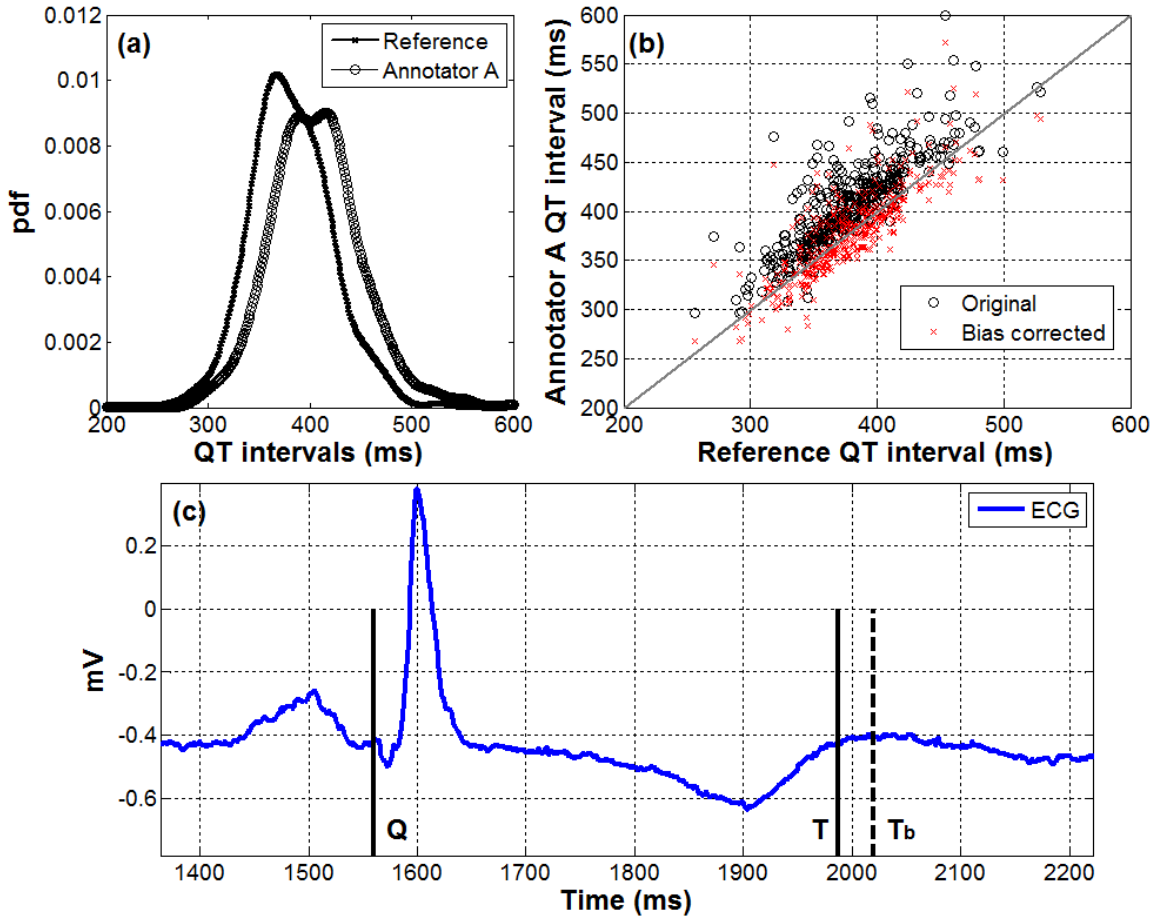


Fig. 2.1 An example of bias in the context of electrocardiogram (ECG) QT interval labelling. (a) The probability density function of the QT intervals for the reference (supplied by the human experts) annotation and annotator A (such as an automated algorithm). (b) A plot of QT intervals across different recordings: the diagonal (grey) line indicates a perfect match of QT intervals between the reference and annotator A; the 'o' indicates the original QT intervals provided by annotator A; the '×' indicates the bias-corrected QT intervals of annotator A, which fits closely to the diagonal line. (c) An example of bias that occurs in an ECG record for labelling QT interval. The reference QT interval on a single beat starts at the beginning of the Q wave and ends at the end of the T wave (denoted as Q and T), and the biased trend from annotator A is demonstrated as T_b.

for labels only when the target value was uncertain. Using a simple majority decision rule and GLAD [111] as the benchmark algorithms, Welinder and Perona demonstrated that their proposed model was superior when fusing labels provided by 47 Mechanical Turk annotators for a synthetic dataset. Welinder *et al.* [128] also presented a more specialised form of the Bayesian model which modelled the bias of an annotator and an estimate of the task difficulty, but it was focused on the modelling of binary tasks for classifying water birds in 40 images. Their study focused on fusing annotations provided by 40 Mechanical Turk labellers. Their inferred ground truth using the proposed model had 75.4% accuracy, compared to 68.3% accuracy when using the majority voting approach and 60.4% accuracy with the GLAD method proposed by Whitehill *et al.* [111]. However, the proposed model cannot account for more complex tasks such as continuous-valued labelling.

In medical imaging, Warfield *et al.* [3] proposed a method for validating segmentation by estimating the bias and variance of each annotator, serving as a scalar version of STAPLE [105]. They considered the label of each voxel in an image to be a continuous score, and that the annotations provided by each annotator were a noisy version of the true score (which can be obtained through distance transform for each voxel of an image). Through fusing annotations using the EM algorithm, they jointly inferred the unknown ground truth as well as the bias and variance (i.e., inverse precision) of the labels for each annotator. In their study of nine annotators, each was asked to provide up to three MRI segmentations for each of four brain tumours. The inferred segmentation was more consistent with that of the annotators than both the average of the segmentations and that of a single annotator segmentation. A similar model was described by Ouyang *et al*. [129] that tackled the quantitative ground truth (such as count and percentage estimation) measure in crowd sensing. Xing *et al*. had recently proposed using a Gaussian prior on the bias parameter for the identification of cardiac landmarks in two dimensional images [130]. However, their model does not cater

Table 2.1 Survey of probabilistic techniques for fusing annotations to infer latent ground truth.

Source	Data Type	Feature Incorporation in the Model	Modelling of Annotator's Expertise	Modelling of Annotator's Bias	Dealing with Missing Annotations
Wiebe *et al*. [131]	Categorical	✓	X	✓	N/A
Snow *et al*. [132]	Categorical, Continuous	X	✓	✓	N/A
Warfield *et al*. [3]	Continuous	X	✓	✓	N/A
Commowick and Waefield [133]	Continuous	X	✓	✓	N/A
Raykar *et al*. [89]	Binary, Ordinal, Continuous	✓	✓	X	✓
Ipeirotis *et al*. [134]	Binary	X	✓	✓	N/A
Welinder and Perona [126]	Binary, Multi-valued, Continuous	X	✓	X	✓
Welinder *et al*. [128]	Binary	✓	✓	✓	N/A
Baba and Kashima [135]	Categorical	X	✓	✓	N/A
Cabrera *et al*. [136]	Binary	✓	✓	✓*	N/A
Xing *et al*. [130]	Continuous	X	✓	✓	N/A
Xing *et al*. [137]	Continuous	X	✓	✓	N/A
Akhondi-Asl *et al*. [138]	Continuous	X	✓	✓	N/A
Ouyang *et al*. [129]	Continuous	X	✓	✓	✓
Nasir *et al*. [139]	Continuous	X	X	✓	✓
Kamar *et al*. [140]	Categorical	✓	✓	✓**	N/A
Proposed Method	Continuous	✓	✓	✓	✓

Note: N/A – not available as it was not modelled or discussed in the publication. The * means that the bias is modelled as observation-specific (i.e., dataset-specific), and the ** refers to both annotator- and dataset-specific.

for missing annotations and the possibility of incorporating physiological features into the model to further improve the estimation of ground truth as shown in [89, 127].

The aforementioned-studies [3, 89, 126, 128, 130] serve as the basis of the proposed model and its variants. Our method differs from the above by being considered in this thesis: a Bayesian generative model which aggregates continuous-valued labels in an unsupervised manner, to estimate jointly the precision and bias values of each annotator, and infer the underlying ground truth as well as training a classifier. A comprehensive comparison of similar models is presented in Table 2.1.

2.3 Bayesian Probability in Parameter Estimation

A generative model is commonly considered as a stochastic process that randomly simulates one or more synthetic datasets as observations, given some model parameter values. It is fully probabilistic as it models the joint probability of all parameters. Furthermore, it is also possible to infer unknown or latent parameters in a model given the observations (e.g., annotations provided by annotators). Assuming there are observed data $\mathbf{D}$ and the parameter θ is to be estimated for a model, then Bayes' theorem can be used to evaluate the posterior probability of θ after $\mathbf{D}$ has been observed:

$$p(\theta \mid \mathbf{D}) = \frac{p(\mathbf{D} \mid \theta)\, p(\theta)}{\int_{\theta} p(\mathbf{D} \mid \theta)\, p(\theta)\, d\theta}, \tag{2.1}$$

where the quantity $p(\mathbf{D} \mid \theta)$ is the likelihood function of observing data $\mathbf{D}$ given different values of θ, and $p(\theta)$ is the prior probability distribution over θ. Prior knowledge of θ can be obtained from expert knowledge, contextual information, and previous observations before seeing the current data $\mathbf{D}$. The denominator is the normalised constant described as the "evidence" or marginal likelihood, which ensures that $p(\theta \mid \mathbf{D})$ is a probability density that integrates to one [95]. In many applications where the interest lies in estimating the posterior with various values of θ, the denominator is considered to be fixed, and hence the posterior is proportional to the product of the likelihood and the prior:

$$\begin{aligned} p(\theta \mid \mathbf{D}) &\propto p(\mathbf{D} \mid \theta)\, p(\theta) \\ &\propto \prod_{i=1}^{N} p(\mathbf{y}_i \mid \theta)\, p(\theta), \end{aligned} \tag{2.2}$$

where $\mathbf{D}$ is assumed to have N independent observations, such as $\mathbf{D} = \{\mathbf{y}_{i=1}, \cdots, \mathbf{y}_{i=N}\}$, and $\mathbf{y}_i$ is a row vector for the ith observation.

In contrast to Bayesian methods, frequentist approaches can also be used to approximate the parameters of interest, where the posterior probability is determined entirely from the observations themselves. One of the most commonly-used methods for this purpose is maximum likelihood (ML), which assumes the following:

$$p(\theta \mid \mathbf{D}) = \prod_{i=1}^{N} p(\mathbf{y}_i \mid \theta). \tag{2.3}$$

The ML approach obtains a point estimate for θ that maximises the likelihood function; i.e., $\underset{\theta}{\operatorname{argmax}} \left\{ \prod_{i=1}^{N} p(\mathbf{y}_i \mid \theta) \right\}$. An example of the ML approach for a given set of observation $\mathbf{y}_1$ is shown in Fig. 2.2 (a), where it estimates the most probable value of θ that best explains $\mathbf{y}_1$; i.e., $\underset{\theta}{\operatorname{argmax}} \{ p(\mathbf{y}_1 \mid \theta) \}$. A similar example for two sets of observations $\{\mathbf{y}_1, \mathbf{y}_2\}$ is shown in Fig. 2.2 (b) where ML maximises the joint probability; i.e., $\underset{\theta}{\operatorname{argmax}} \{ p(\mathbf{y}_1, \mathbf{y}_2 \mid \theta) \}$. Note that the prior $p(\theta)$ is missing in the ML approach, or equivalently, is assumed to have a uniform prior of one (i.e., there is equal probability for each value of θ). However, because the ML approach produces a point estimate, it can be heavily biased when only a small set of data is observed, and is sensitive to the choice of starting values where a local maximum (instead of the global maximum) may be found. An example of inferring the parameters and the ground truth as a latent variable using the ML approach will be detailed in Chapter 4.

Beyond the ML approach, the maximum-a-posteriori (MAP) method incorporates a prior distribution over the data $\mathbf{D}$, which acts as a regularisation term to ensure that the posterior probability does not solely depend on a potentially-small number of observations. The posterior of θ for MAP is written as:

$$p(\theta \mid \mathbf{D}) = \prod_{i=1}^{N} p(\mathbf{y}_i \mid \theta)\, p(\theta). \tag{2.4}$$

Fig. 2.2 demonstrates the difference between the ML and the MAP approaches for one set or multiple sets of observations. The posterior of θ is estimated by the MAP ap-

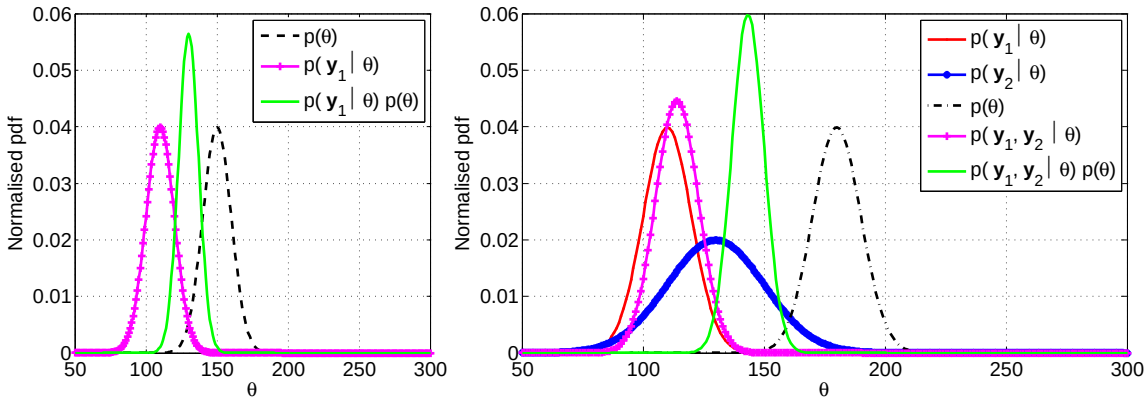


Fig. 2.2 Examples of the difference between the ML and the MAP approaches: (a) when there is a set of observations $\mathbf{y}_1$, the ML approach estimates the most probable value of θ that best explains these observations; i.e., $\underset{\theta}{\text{argmax}}\{p(\mathbf{y}_1 \mid \theta)\}$, whereas the MAP approach estimates the joint probability of the likelihood function and the prior by $\underset{\theta}{\text{argmax}}\{p(\mathbf{y}_1 \mid \theta)\,p(\theta)\}$; (b) when there are two sets of independent observations, $\mathbf{y}_1$ and $\mathbf{y}_2$, the ML estimation of the θ maximises their joint probability; i.e., $\underset{\theta}{\text{argmax}}\{p(\mathbf{y}_1,\mathbf{y}_2 \mid \theta)\}$, while MAP incorporates the prior as $\underset{\theta}{\text{argmax}}\{p(\mathbf{y}_1,\mathbf{y}_2 \mid \theta)\,p(\theta)\}$.

proach as maximising the joint probability of the likelihood function and prior distribution (i.e., $\underset{\theta}{\text{argmax}}\{\prod_{i=1}^{N} p(\mathbf{y}_i \mid \theta)\,p(\theta)\}$). This is demonstrated as $\underset{\theta}{\text{argmax}}\{p(\mathbf{y}_1 \mid \theta)\,p(\theta)\}$ for dataset $\mathbf{y}_1$ as shown in Fig. 2.2 (a), and $\underset{\theta}{\text{argmax}}\{p(\mathbf{y}_1,\mathbf{y}_2 \mid \theta)\,p(\theta)\}$ for datasets $\{\mathbf{y}_1,\mathbf{y}_2\}$ as shown in Fig. 2.2 (b). The estimated posterior distribution of θ (i.e., $p(\mathbf{y}_1 \mid \theta)$ or $p(\mathbf{y}_1,\mathbf{y}_2 \mid \theta)$) using the ML approach differs from that obtained using MAP (i.e., $p(\mathbf{y}_1 \mid \theta)\,p(\theta)$ or $p(\mathbf{y}_1,\mathbf{y}_2 \mid \theta)\,p(\theta)$). The value for θ is chosen at the mode of its posterior distribution: it is 110 and 130 for dataset $\mathbf{y}_1$, 114 and 143 for datasets $\{\mathbf{y}_1,\mathbf{y}_2\}$, using ML and MAP, respectively. The BCLA model will be solved using the MAP approach and detailed in Chapter 5.

In comparison to the use of frequentist approaches, Bayesian inference requires the computation of the full posterior distribution for the parameters of interest, particularly in the estimation of the marginalisation of the denominator in equation (2.1). Depending on the choice of the prior, the integral of the product between the likelihood function and the prior

can take different algebraic forms. To make this integral tractable, it is often convenient to choose so-called conjugate distributions to model the prior so that the posterior distribution would be in the same family as the prior (i.e., it has the same algebraic form as the prior). The prior in this form is called the conjugate prior for a given likelihood function. Defining the prior to be a conjugate simplifies the estimation of the posterior, as its parameters can be derived in closed form. The Gibbs sampler, a variety of Markov chain Monte Carlo algorithm, takes advantage of conjugate priors to approximate the posterior distribution. Details of the Gibbs sampler will be described in Chapter 6, where we introduce it in the context of inferring ground truth, and thereby for approximating associated parameters in the proposed models of the thesis.

2.4 Summary

This chapter has summarised the relevant voting algorithms in the literature which infer a consensus in the scenario where only noisy annotations of some underlying true value are provided.

Raykar's regression model [89], denoted as EM-R, will be detailed in Chapter 4. However, the EM-R does not model the bias of each annotator explicitly. In comparison, the scalar version of the STAPLE (denoted as sSTAPLE) proposed by Warfield *et al.* [3] (see Chapter 5) considered the bias of each annotator but did not model the ground truth by means of learning a classifier as a function of the features. Following the concept of constructing a Bayesian framework by Welinder *et al.* [126], the BCLA model will later be introduced in the thesis (see Chapter 5) to combine both EM-R and sSTAPLE to form a novel Bayesian generative model for fusing medical labels.

Chapter 3

Description of Datasets

3.1 Introduction

This chapter outlines the medical datasets considered in the work described by this thesis: the QT interval dataset of ECG and the respiratory rate dataset of PPG.

3.2 The PhysioNet/Computing in Cardiology QT Dataset

3.2.1 Dataset Description

The 2006 PhysioNet/Computing in Cardiology (PCinC) QT dataset [37] provides an excellent opportunity to assess the feasibility of crowd-sourced annotations with large amounts of human or algorithmic annotations. The data were drawn from the QT interval annotations generated by participants in the 2006 PCinC Challenge [37]. Each participant in the challenge was required to submit a Q onset time with an accompanying T offset time for one "representative" beat[1] (see Fig. 3.1), for lead II of each of the 549 recordings in the

[1]The Q and T timestamps for a record were obtained directly from the competition: they were the interval in milliseconds from the beginning of the record to the beginning of the measured QT interval, and the interval from the beginning of the record to the end of the measured QT interval.

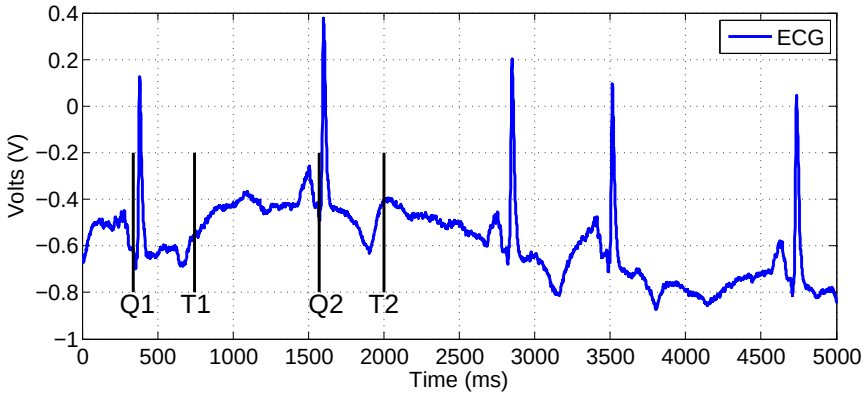


Fig. 3.1 Example of Q and T annotations from two annotators for a recording in the PCinC QT dataset, where Q1 and T1 were provided by one annotator, and Q2 and T2 were provided by the second annotator.

Table 3.1 Summary of the diagnostic conditions of subjects in the PTBDB.

Diagnosis	**Number of subjects**
Healthy controls	52
Myocardial infarction	148
Cardiomyopathy/heart failure	18
Bundle branch block	15
Dysrhythmia	14
Myocardial hypertrophy	7
Valvular heart disease	6
Myocarditis	4
Miscellaneous	4
N/A	22

Note: N/A refers to subjects included in the PTBDB, but where their clinical summaries are missing.

Physikalisch-Technische Bundesanstalt Diagnostic ECG Database (PTBDB) [141]. Each ECG lead II (up to two minutes in length) was digitised at 1000Hz, with 16-bit resolution, over a range of $\pm$16.384mV. The records were obtained from 290 subjects (209 men with mean age 55.5, and 81 women with mean age 61.6), each represented by between one and five recordings. About 20% of the subjects were healthy controls. The PTBDB contained records of patients with a variety of ECG morphologies having different QT intervals ranging from 256 ms to 529 ms. Diagnostic classifications are detailed in [141] and summarised in Table 3.1.

Table 3.2 Performance by participants for each division of the PCinC QT dataset.

	Manual annotators	**Automated algorithms**		
	Division 1	**Division 2 (closed-source)**	**Division 3 (open-source)**	**Division 4 (closed- & open-source)**
Number of annotators	20	48	21	69
Average annotations per 5s segment	18	39	15	54
Average annotations per 2min segment	18⋆	41⋆	21⋆	62⋆
RMSE of 5s segment (ms)	6.65	16.36	17.46	16.36
RMSE of 2min segment (ms)	6.67⋆	16.34⋆	17.33⋆	16.34⋆

Note: The manual/automated annotator having the lowest RMSE over a 5 s segment was selected to represent the best score. The results annotated ⋆ were published in the competition for a 2min segment.

There were two categories of annotations: manual and automated (see Table 3.2). 89 entries to the competition were submitted, including revised submissions for a total of 38,621 annotations sourced from: 20 human annotators in Division 1; 48 automated algorithms in Division 2 (closed-source); and 21 in Division 3 (open-source). An additional division, Division 4, was created in this thesis so as to combine all automated algorithms from Divisions 2 and 3, and to infer a potentially better estimation of QT intervals. The distribution of QT annotations for each division is shown in Fig. 3.2, where it maybe be seen that QT annotations from all entries are not approximately Gaussian-distributed (Jarque-Bera test [142] with $p < 0.01$). A single record, "patient285/s0544re", was excluded as it did not contain any recognisable ECG signals. Annotations for 548 records of the PTBDB were processed using different voting strategies. The maximum number of annotators per division and averaged number of annotations per record are listed in Table 3.2.

The competition score for each entry was calculated from the root-mean-square-error (RMSE) between the submitted and the reference QT intervals. The reference annotations were generated from Division 1 entries using a maximum of 15 participants by taking the "median self-centering approach" as detailed in [143]. The best-performing annotator with least RMSE score for each division is also listed in Table 3.2. Furthermore, the majority of

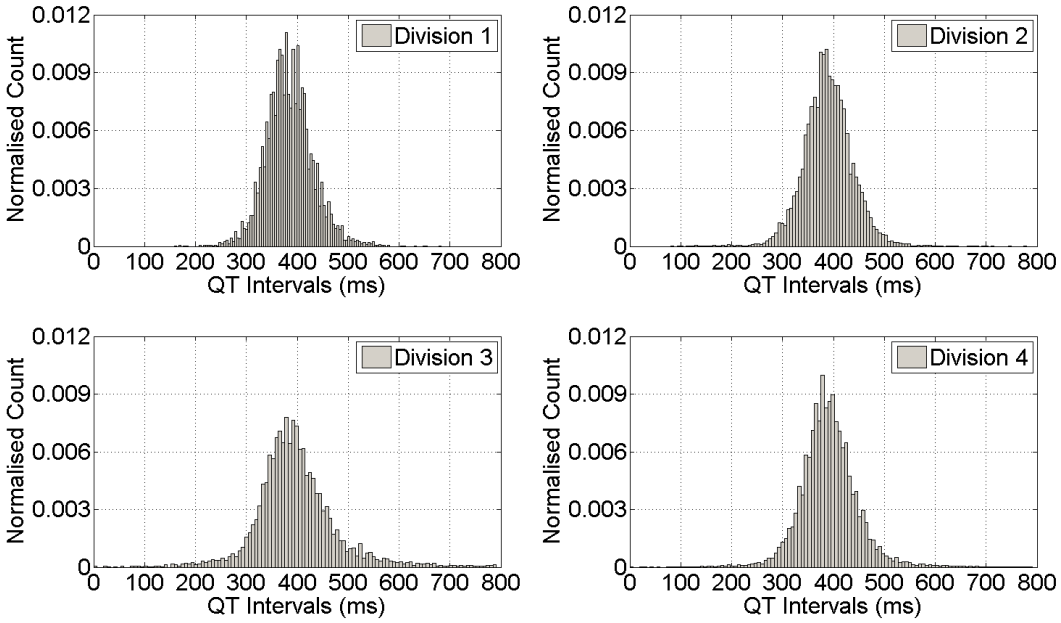


Fig. 3.2 Histograms of the QT annotations for all entries including human annotators (Division 1) and automated algorithms (Divisions 2-4). QT annotations from all entries are not Gaussian-distributed (Jarque-Bera test [142] with $p < 0.01$).

the QT annotations of each 2min record occurred within the first 5s period, and the best scores in the first 5s segment were similar to those of the 2min segment (denoted by $\star$ in Table 3.2). To reduce any possible inter-beat variations, only the annotations within the first 5s segment of each record were chosen, ensuring that all annotators had approximately labelled the same region of a record, with similar QT morphologies. Therefore, the motivation for choosing the first 5s segment of each record was to consider a short segment where the QT interval is not changing dramatically (with respect to any particular beat that an annotator chose to view), while retaining the highest number of annotations. Those that fell outside this segment were considered to be missing information and discarded in the process of the QT estimation.

As not all annotators had provided complete annotations (i.e., 548 QT interval annotations for 548 recordings), the histogram of the percentage of recordings per annotator in each division is shown in Fig. 3.3. Note that 95% (i.e., 19 out of 20) manual annotators had labelled at least 50% of the recordings, but only 81.2% of the automated algorithms had done so in Division 4. The percentage of annotators per recording is also plotted for each division

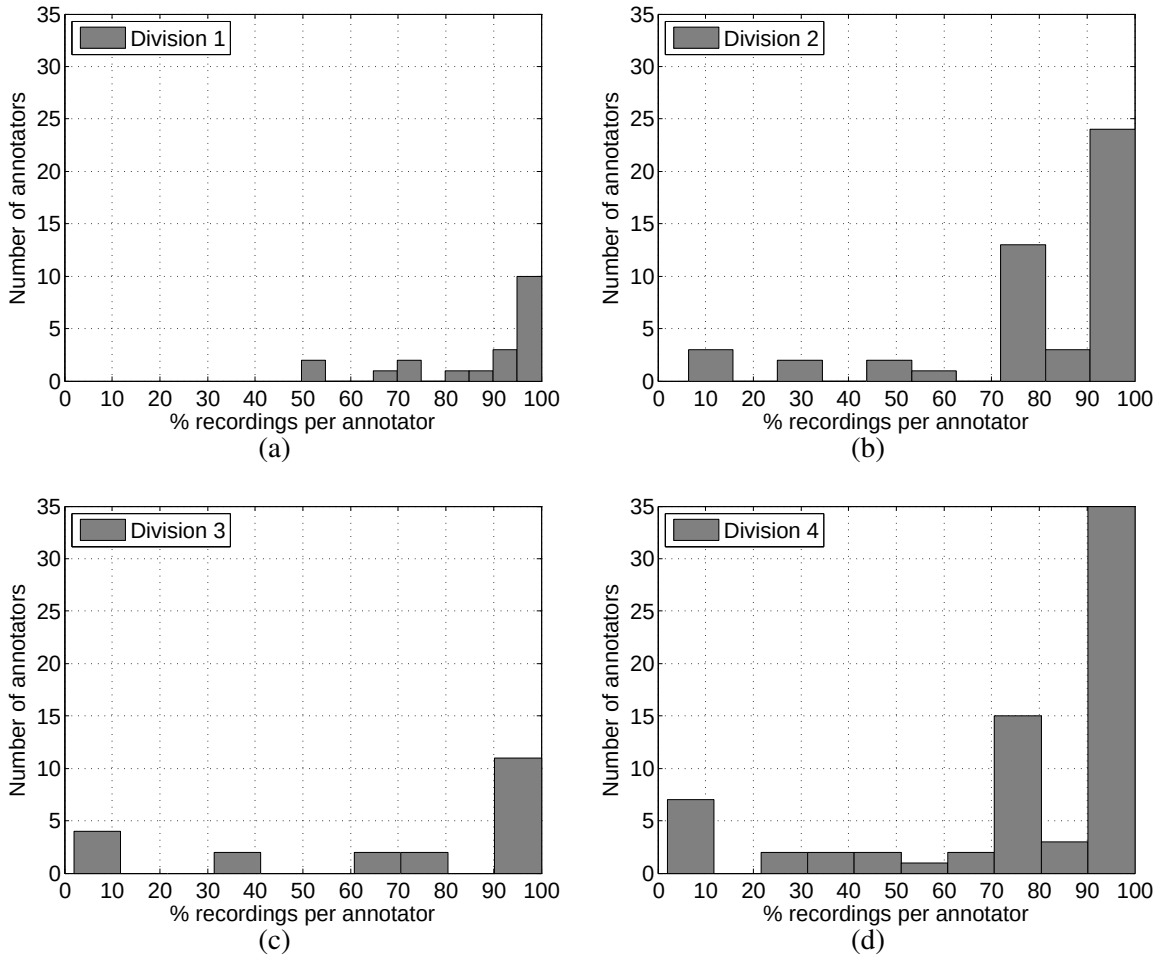


Fig. 3.3 Histograms of the percentage of QT recordings annotated per individual annotator for manual annotators in Division 1 and automated algorithms in Divisions 2, 3, and 4, where Division 4 is a combination of Divisions 2 and 3.

(see Fig. 3.4). There are at least 33.3% and 28.6% of the annotators labelled one recording in the automated entry and the manual entry respectively. Thus, we have a substantial "missing data" conditions that a fusion strategy must accommodate.

Although the QT annotations were provided in the PCinC QT dataset, the source code of the algorithms is not in the public domain. We therefore have no *a priori* information that the algorithms were independent. Furthermore, the reference QT intervals provided for the dataset were based on bootstrapping the median of 15 human annotators, which can be biased because humans tend to underestimate the QT interval [12]. A generative model is

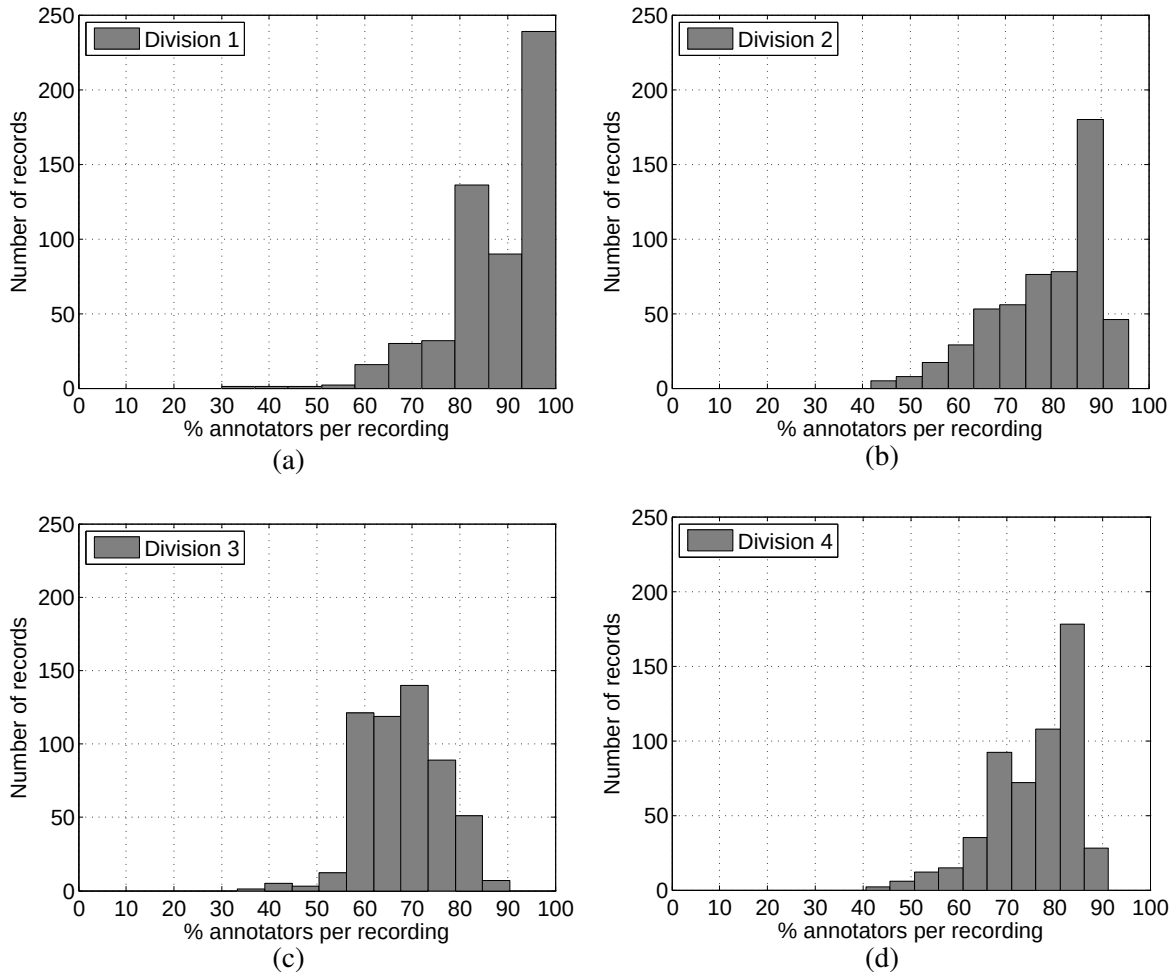


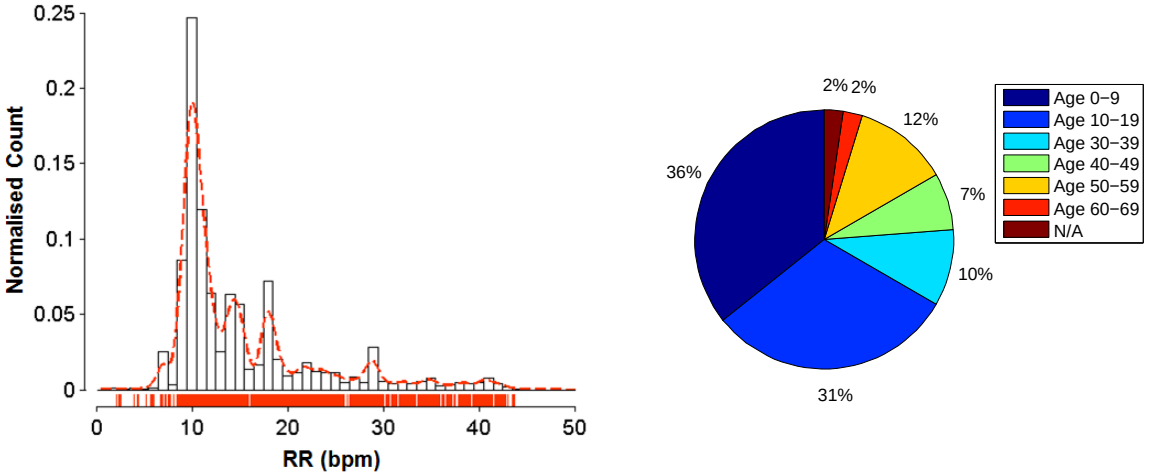
Fig. 3.4 Histograms of the percentage of annotators per individual recording for manual annotators in Division 1 and automated algorithms in Divisions 2, 3, and 4.

therefore proposed in this thesis due to the need to provide an unbiased estimate of ground truth of the QT intervals.

3.3 The Capnobase Respiratory Rate Dataset

3.3.1 Dataset Description

The Capnobase dataset [144] was used for evaluating the proposed fusion methodology described in this thesis. The records in the dataset were randomly selected by its creators from a larger collection of physiological signals collected during elective surgery and routine



(a) (b)

Fig. 3.5 (a) The reference histogram of the Capnobase database for 42 subjects, each with 150 respiratory rate values estimated over a 32s moving window (overlapped by 29s) for an 8min recording. An empirical distribution is shown superimposed on top as indicated in red dash line. (b) The pie chart of the categorised age groups and their corresponding percentage of recordings in the Capnobase database. Note that one subject has missing age information, denoted as N/A, and there is no subject with age 20 to 29.

anaesthesia. The dataset consists of PPG recordings and gold-standard invasive capnometry data (Fs = 300 Hz), from 59 children (median age: 8.7, range: 0.8 - 16.5 years) and 35 adults (median age: 52.4, range: 26.2 - 75.6 years). The dataset described in [81], which includes 42 subjects of 8 minutes reliable recordings (336 minutes in total) of spontaneous breathing or controlled ventilation was considered in the thesis. The capnometric waveform was used as the reference gold standard recording for RR estimates, as detailed in [81]. Each breath in the capnogram has been manually labelled by a research assistant, and these labels were used to derive the reference respiratory values (as described in [81]): one RR value was obtained for a 32s moving window, overlapped by 29s. This resulted in a total of 150 RRs per recording, and each recording corresponds to one subject. The distribution of the reference RR values from 42 subjects, and the percentage of recordings corresponding to different age groups, are shown in Fig. 3.5. We can see from the figure that RR mostly takes values in the range 0-20 bpm and that most of the subjects are young.

Extraction of the Respiratory Derived Modulation

RR was computed for 32s windows of PPG, with successive windows having 29s overlap to match the windows used for construction of the reference from the capnography waveform. To extract the three respiratory-induced modulations (AM, BW, and FM), PPG beat detection was performed using a segmentation algorithm [145]. The latter produces a series of maximum and minimum intensities for each pulse. As shown previously in Fig. 1.2, the series of maximum intensities of the PPG pulses was used for extracting the BW timeseries. The (max-min) amplitude of successive beats was used to extract the AM timeseries. The intervals between successive beats was used to extract the FM timeseries. RR was estimated using two different spectral approaches that have been used in the recent literature (see Fig. 3.6): Fourier analysis and autoregressive (AR) modelling. Spectral analysis requires evenly-sampled data, and so each timeseries (corresponding to BW, AM and FM) was first re-sampled at 4 Hz using linear interpolation. For the first method [81], the frequency spectra of the resulting respiratory signals were calculated. The frequency at which the maximum intensity of each spectrum is obtained within the frequency range of interest (corresponding to 3 to 60 bpm), was taken as the respiratory frequency (Fig. 3.6). For the latter [146, 86], an AR model (of order 7) was fitted to each timeseries [147]. The respiratory frequency was identified as that corresponding to the pole with the greatest magnitude within the plausible range of frequencies. It is noted that for each window, and for each algorithm, three RR estimates were determined. These RR estimates might be correlated because they were derived from the same analytical method but with non-linear transformation on the three individual modulations.

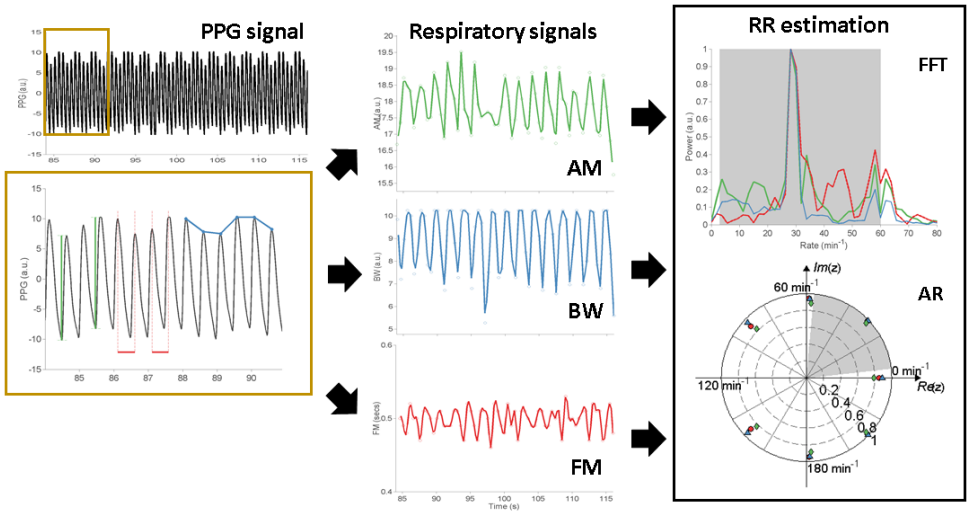


Fig. 3.6 Example of a 32s window of PPG used for RR estimation. The amplitude modulation (AM in green), baseline wander (BW in blue), and modulation of beat-to-beat intervals (FM in red) are extracted from the PPG. An autoregressive (AR) method can be used to find the dominant spectral component within the plausible range of RR (grey area), which is assumed to correspond to respiration. The power spectrum is calculated for each modulation using Fourier analysis such as Fast Fourier Transform (FFT), and the maximum power is selected within the physiologically-plausible RR range (grey area).

3.4 Simulated Datasets

Different simulated datasets were generated to test the proposed models throughout the thesis. The goal of each simulated experiment was to determine if a proposed model can recover the true bias and precision of each annotator that are known in the experiment, and which were used to generate the simulated data. As each simulated dataset was generated accordingly to a specific proposed model, it will be described latter in the thesis after the proposed model has been introduced.

Chapter 4

Non-Bayesian Continuous-Valued Label Aggregation

"By three methods we may learn wisdom:
first, by reflection which is noblest; second by imitation
which is easiest; and third by experience, which is bitterest."

– Attributed to Confucius

4.1 Introduction

This chapter describes a probabilistic framework for aggregating labels (from human experts, algorithms, or both) to provide a reliable ground truth, with no prior knowledge of the performance of individual annotators. As an alternative to median- or mean-voting strategies, we incorporate novel contextual features (signal quality and physiology), then use a probabilistic expectation-maximisation (EM) method [89] to aggregate labels from a panel of annotators. The EM method will form the basis for comparison with more complex models developed in later chapters, and is termed EM-R henceforth (from EM-Raykar, the lead author of that existing approach).

4.2 The EM-R Approach

The EM framework proposed by Raykar *et al.* [89] was applied in this chapter to assess its potential in: (i) combining annotations from manual experts to build a QT "ground truth"; (ii) fusing annotations from algorithms to assess the performance of combining automated algorithms' outputs for QT estimation. Note that we only consider the QT dataset in this chapter, to illustrate the use of the existing EM-R approach.

The Ground Truth Model

Suppose that there are N records of physiological time series data. The underlying ground truth (the *true* time or duration of an event or diameter of an object for example) for the ith record, z_i, can be assumed to be drawn from a univariate Gaussian distribution[1] with mean a_i and variance $1/b$. The probability density function (denoted as pdf) of z_i is defined as follows:

$$p(z_i \mid a_i, b) = \mathcal{N}(z_i \mid a_i, 1/b). \tag{4.1}$$

Furthermore, z_i can be defined as being dependent on some feature vector $\mathbf{x}_i$, using a linear regression model $z_i = \mathbf{x}_i^\intercal \mathbf{w} + \varepsilon$, where a can be expressed as a linear regression function $f(\mathbf{w}, \mathbf{x})$ with an intercept w_0: $\mathbf{w}$ are the coefficients of the regression that also includes w_0. $\mathbf{x}_i$ is a column feature vector for the ith record containing d features (i.e., d-dimensional design matrix, $\mathbf{X} = \left[\mathbf{x}_1^\intercal, ..., \mathbf{x}_N^\intercal\right]$). To cater for the modelling of w_0, a scalar value of one was added in the feature matrix (i.e., $\mathbf{x}_i := [\mathbf{1}, \mathbf{x_i}]$)). ε is a zero-mean Gaussian noise with precision b. The pdf of z_i may then be defined as:

$$p(z_i \mid \mathbf{x}_i, \mathbf{w}, b) = \mathcal{N}\left(z_i \mid \mathbf{x}_i^\intercal \mathbf{w}, 1/b\right). \tag{4.2}$$

[1] A univariate Gaussian distribution can be defined as $\mathcal{N}(z \mid \mu, \sigma^2) = \left(2\pi\sigma^2\right)^{(-1/2)} \exp\left\{-(z-\mu)^2/2\sigma^2\right\}$, where μ is the mean and σ^2 is the variance of the distribution.

If one further assumes that the ground truth can be drawn independently from the N records, the conditional pdf of $\mathbf{z}$ is given by:

$$p(\mathbf{z} \mid \mathbf{x}_i, \mathbf{w}, b) = \prod_{i=1}^{N} \mathcal{N}\left(\mathbf{z} \mid \mathbf{x}_i^{\intercal}\mathbf{w}, 1/b\right). \tag{4.3}$$

Model for Annotators

We assume that there are N records labelled by R annotators. Let $\mathbf{D} = \left[\mathbf{x}_i^{\intercal}, \mathbf{y}_i\right]_{i=1}^{N}$, where $\mathbf{x}_i$ is a column feature vector for the ith observation, y_i^j corresponds to the annotation provided by the jth annotator for the ith record ($j = 1, \cdots, R$). We also assume that y_i^j is a noisy version of z_i, with a Gaussian distribution $\mathcal{N}\left(y_i^j \mid z_i, (\sigma^j)^2\right)$. Here σ^j is the standard deviation in annotation of the jth annotator (around the underlying ground truth). The pdf of y_i^j can be written as:

$$p\left(y_i^j \mid z_i, \tau^j\right) = \mathcal{N}\left(y_i^j \mid z_i, 1/\tau^j\right), \tag{4.4}$$

where σ^j is replaced with $1/\tau^j$, the precision of the jth annotator, defined as the inverse-variance of annotator j. Note that τ^j is considered to be a constant; i.e., annotators are assumed to be consistently precise (with precision τ^j) over all records $i = 1, \cdots, N$. Note that the term τ^j does not specify the bias of the jth annotator, only the spread around the underlying ground truth.

The Combined Model

The distribution of y_i^j can be estimated as the following posterior which combines equation (4.4) and (4.2):

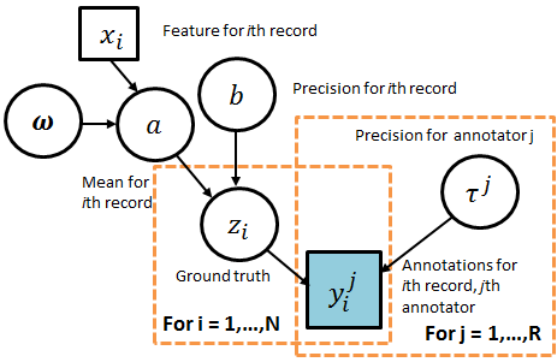


Fig. 4.1 Graphical representation of the EM-R model: y_i^j corresponds to the annotation provided by the jth annotator for the ith record, and it is modelled by the z_i (the unknown underlying ground truth), and the τ^j (precision). Furthermore, z_i is drawn from a Gaussian distribution with mean a and variance b^{-1}, where a is a function of some feature vector $\mathbf{x}_i$.

$$
\begin{aligned}
p\left(y_i^j \mid \mathbf{x}_i, \mathbf{w}, b\right) &= \int_{z_i} p\left(y_i^j, z_i \mid \mathbf{w}, \mathbf{x}_i, \tau^j, b\right) \mathrm{d}z_i \\
&= \int_{z_i} p\left(y_i^j \mid z_i, \tau^j\right) p\left(z_i \mid \mathbf{x}_i, \mathbf{w}, b\right) \mathrm{d}z_i \\
&= \int_{z_i} \mathcal{N}\left(y_i^j \mid z_i, 1/\tau^j\right) \mathcal{N}\left(z_i \mid \mathbf{x}_i^\intercal \mathbf{w}, 1/b\right) \mathrm{d}z_i \\
&= \int_{z_i} \frac{1}{\sqrt{2\pi/\tau^j}} \mathrm{e}^{-\left(y_i^j - z_i\right)^2 / \left(\frac{2}{\tau^j}\right)} \frac{1}{\sqrt{2\pi/b}} \mathrm{e}^{-\left(z_i - \mathbf{x}_i^\intercal \mathbf{w}\right)^2 / \left(\frac{2}{b}\right)} \mathrm{d}z_i \\
&= \frac{1}{\sqrt{2\pi(\frac{1}{b} + \frac{1}{\tau^j})}} e^{-\left(y_i^j - \mathbf{x}_i^\intercal \mathbf{w}\right)^2 / 2\left(\frac{1}{b} + \frac{1}{\tau^j}\right)} \\
&= \mathcal{N}(y_i^j \mid \mathbf{x}_i^\intercal \mathbf{w}, 1/\lambda^j), \qquad (4.5)
\end{aligned}
$$

where $\boldsymbol{\lambda}$ is the precision for all annotators with $\boldsymbol{\lambda} = \{\lambda^{j=1}, \cdots, \lambda^{j=R}\}$ and $1/\lambda^j = 1/\tau^j + 1/b$. The graphical representation of the model is shown in Fig. 4.1.

4.2.1 The Expectation-Maximisation Algorithm

Given the feature matrix and the labels $\mathbf{D}$, the weights vector $\mathbf{w}$ can be determined by maximising the log-likelihood of the data. Furthermore, in order to solve for the unknown hidden true annotation z_i for the ith record, the annotations for each record may be combined

as a weighted sum using their respective precision values, $\boldsymbol{\lambda}$. Under the assumption that records are independent, and $y_i^1, \cdots, y_i^R$ are conditionally independent given the feature $\mathbf{x}_i$ (i.e., each annotator works independently to provide QT annotations)[2], the likelihood of the parameter $\boldsymbol{\theta} = \{\mathbf{w}, \boldsymbol{\lambda}\}$ for a given $\mathbf{D}$ can be formulated as:

$$p(\mathbf{D} \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} p\left(y_i^1, \cdots, y_i^R \mid \mathbf{x}_i, \boldsymbol{\theta}\right), \tag{4.6}$$

$$= \prod_{i=1}^{N} \prod_{j=1}^{R} \mathcal{N}\left(y_i^j \mid \mathbf{x}_i^\intercal \mathbf{w}, 1/\lambda^j\right). \tag{4.7}$$

The likelihood of parameters $\boldsymbol{\theta}$ for a given input set $\mathbf{D}$ can be written using Bayes' theorem as follows:

$$p(\boldsymbol{\theta} \mid \mathbf{D}) \propto p(\mathbf{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})$$

$$= \prod_{i=1}^{N} \prod_{j=1}^{R} \mathcal{N}\left(y_i^j \mid \mathbf{x}_i^\intercal \mathbf{w}, 1/\lambda^j\right), \tag{4.8}$$

where $p(\boldsymbol{\theta})$ is assumed to be a uniform prior distribution on the parameters (i.e., $p(\boldsymbol{\theta}) = 1$). Parameter values $\boldsymbol{\theta}$ may be estimated using a maximum likelihood approach, which maximises the log-likelihood of the parameters, i.e., $\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\{\log p(\boldsymbol{\theta} \mid \mathbf{D})\}$. The log-likelihood can now be written as:

$$\log p(\boldsymbol{\theta} \mid \mathbf{D}) = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{R} \left[\log\left(\frac{2\pi}{\lambda^j}\right) + \left(y_i^j - \mathbf{x}_i^\intercal \mathbf{w}\right)^2 \lambda^j \right]. \tag{4.9}$$

The Expectation-Maximisation algorithm proposed by Dempster *et al.* [148] can be used to solve the complete data (i.e., $\{\mathbf{D}, \mathbf{z}\}$) log-likelihood in a two-step iterative process (see Appendix A.1 for details):

[2]This may not be necessarily true, especially in cases of labels from automated algorithms. As methods for ECG segmentation are well-studied, many algorithms may well be partially dependent variations of the same approach. Nevertheless, this choice was made to simplify the model and subsequent derivation of the likelihood. Relaxation of this strong assumption will be considered in later chapters.

(i) The E-step estimates the expected *true* annotations for all records, $\hat{\mathbf{z}}$, as a weighted sum of the provided annotations from all annotators and their precisions:

$$\hat{\mathbf{z}} = \frac{\sum_{j=1}^{R} \lambda^j \mathbf{y}^j}{\sum_{j=1}^{R} \lambda^j}. \tag{4.10}$$

(ii) The M-step is based on the current estimation of $\hat{\mathbf{z}}$ and $\mathbf{D}$. The model parameters such as the precision and regression coefficient values, $\boldsymbol{\lambda}$ and $\mathbf{w}$, can be calculated by finding the zero gradient of the expectation of the complete data log-likelihood respectively:

$$\mathbf{w} = \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^\intercal \right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i \hat{z}_i. \tag{4.11}$$

$$\frac{1}{\lambda^j} = \frac{1}{N} \sum_{i=1}^{N} \left(y_i^j - \mathbf{x}_i^\intercal \mathbf{w} \right)^2. \tag{4.12}$$

Note: these are solved for $\mathbf{w}$ in a similar fashion to the approach in standard linear regression with features $\mathbf{x}$, except when it is trained using $\hat{\mathbf{z}}$ as the inferred ground truth.

When the features $\mathbf{x}$ are not available, the precision may be solved as the least-square difference between the annotated labels and the inferred ground truth $\hat{\mathbf{z}}$ as:

$$1/\lambda^j = \frac{1}{N} \sum_{i=1}^{N} \left(y_i^j - \hat{z}_i \right)^2. \tag{4.13}$$

Each annotator was assumed to have equal precision value at the initialisation, thus the initial estimate of $\hat{\mathbf{z}} = \frac{1}{R} \sum_{j=1}^{R} \mathbf{y}^j$. Then the E-step and M-step were iterated until convergence of the precision. As a result, EM-R establishes a weighted sum of annotations $\mathbf{y}$ estimating the inferred true annotations $\hat{\mathbf{z}}$, as well as providing the precision of each annotator λ^j.

4.2.2 Learning from Incomplete Data

In the case where there exist missing labels from annotators, only the available annotations should be considered when inferring the ground truth. The expected z_i can be re-written as:

$$\hat{z}_i = \frac{\sum_{j \in V_i} y_i^j \lambda^j}{\sum_{j \in V_i} \lambda^j}. \tag{4.14}$$

The equation (4.12) for precision of the jth annotator can be re-written as:

$$\frac{1}{\lambda^j} = \frac{1}{N_j} \sum_{i \in U_j} \left(y_i^j - \mathbf{x}_i^\intercal \mathbf{w} \right)^2. \tag{4.15}$$

and the equation (4.13) can be shown as:

$$\frac{1}{\lambda^j} = \frac{1}{N_j} \sum_{i \in U_j} (y_i^j - \hat{z}_i)^2. \tag{4.16}$$

where U_j is the set of records with annotations provided by the jth annotator, and V_i is the set of annotators that provided annotations for the ith record. N_j is the number of records annotated by the jth annotator.

4.3 Data Description and Methodology for Comparison

4.3.1 Data Description

The dataset considered here is the QT interval dataset described in Chapter 3. To recap, there were two categories of annotations: manual and automated (see Table 3.2). Eighty-nine entries were submitted, including revised submissions for a total of 38,621 annotations sourced from: 20 human annotators in Division 1, 48 automated algorithms (closed-source) in Division 2, and 21 algorithms (open-source) in Division 3. Division 4 was further created in this thesis to combine all automated algorithms from Divisions 2 and 3.

The competition score for each entry was calculated from the root-mean-square-error (RMSE) between the submitted and the reference QT intervals. The reference annotations were generated from Division 1's entries using a maximum of 15 participants by taking the "median self-centering approach" as detailed in [143].

4.3.2 Beat Detection and Feature Extraction

The QT interval can be modulated by heart rate [1] and by the electrophysiological substrate which affects the dispersion of ventricular repolarisation [149]. Berger *et al.* [150] in 1997 proposed relating the variability inherent in the estimation of the QT interval with heart rate variability. When heart rate is constant, the QT interval remains relatively constant for many subjects [151]. A sudden increase or decrease in the heart rate reduces or prolongs the duration of myocardial repolarisation, and consequently alters the QT interval. However, these changes in the QT interval are neither instantaneous nor linear, and appears as in the manner of hysteresis [152]. Changes in the QT interval can be divided into two stages: a short phase which occurs within a few heart beats [153], and a long phase which can be of up to two to three minutes duration depending on the type of cardiac disease [154]. When ECG recordings are shorter than 10s, a certain amount of heart rate variability is expected [151]. Patients with a more prolonged adaptation were related to an increased risk of arrhythmia [153]. Thus QT adaptation should be considered when investigating QT interval as it is able to provide discrimination between patients with high and low risk of arrhythmia.

To reduce large inter-beat variations, a short segment of each record was chosen to ensure that all annotators had approximately labelled the same region of a record with similar QT morphologies. We argue that, were a whole record to be chosen, it would bias the QT estimation towards the segment of the whole 2-min record with the most annotations. This would result in penalising annotators who have annotated data that lie away from the most heavily-labelled segment, regardless of whether the annotation was accurate or not. In

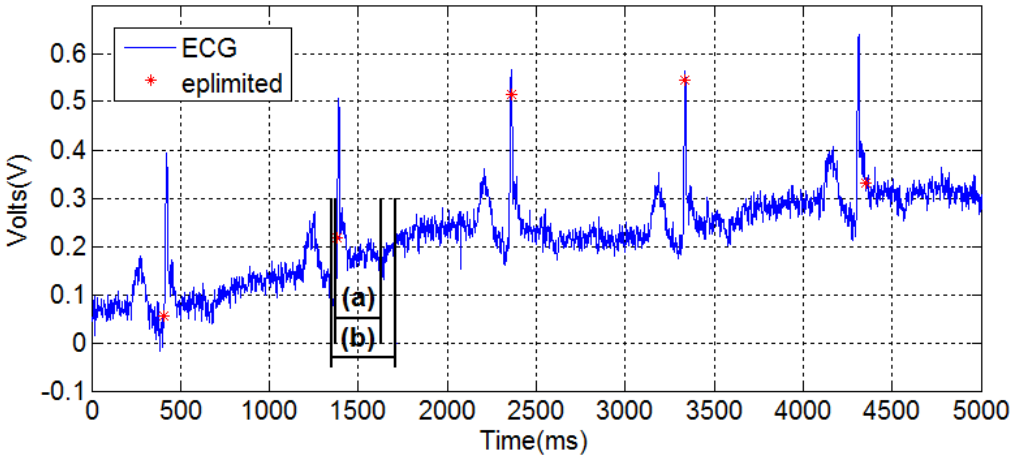


Fig. 4.2 An example of a 5-second noisy ECG segment of a patient with myocardial infarction. The QRS complexes (denoted by *red*$\star$) were detected using the *eplimited* algorithm [155]. Within the same ECG cycle, the QT interval (361 ms) was annotated by the best-performing human annotator (RMSE = 6.65 ms) as shown in (b), whereas (a) shows the QT interval (260 ms) annotated by one of the worst-performing human annotators (RMSE = 48.69 ms).

addition, it might be an unfair comparison for the annotators who might be labelling correct QT changes due to pathological changes in a different location of the record.

In this chapter, the first 5-second segment of each record was chosen since the majority of annotations in the dataset were within this region. The annotations that were outside this segment were considered to be missing information, and were thus discarded in the process of the QT estimation to avoid the bias described previously. An example of the first 5-second ECG segment with R-peaks detection is shown in Fig. 4.2, with R-peak detection performed using the *eplimited* [155] algorithm. In each record, the physiological context together with seven signal quality indices (SQIs) proposed in [156] were used as the features for the EM-R model.

The physiological context was incorporated by adding beat-specific heart rate (HR) as a feature. To account for inter-beat variations, each annotated beat was located and the median of the five preceding R-R intervals was used as the beat-specific heart rate feature value. If the annotated beat occurred at the beginning of the record, the first five beats were considered

to estimate the median R-R interval. The square root of this median R-R interval was used as a feature for an individual record to account for the non-linear nature of QT interval adaptation.

Similarly, the SQI features were calculated in a 5-second sliding window with four seconds of overlap between successive windows. The SQIs were calculated for each 5-second window centred around the annotated beat. As only the first 5-second segment was chosen in this dataset, the SQI features were estimated within this segment. An example of a noisy ECG signal is shown in Fig. 4.2 with large inter-beat variation (i.e., 100 ms) between two human annotators. It may be seen immediately from the figure that QT annotation is decidedly non-trivial. The SQI features provide extra information concerning the signal quality and hence inform the decision of EM-R in reducing the weight of labels from annotators who labelled a noisy ECG cycle (i.e., one with low SQI). The seven SQIs [156] are defined as:

(i) bSQI - The percentage of beats on which two different QRS detectors (the *wqrs* [157] algorithm and the *eplimited* algorithm) agree in terms of the R-peak position. The two detectors agree if *wqrs* locates the fiducial point (the onset of the QRS) 100 ms before or 25 ms after the fiducial point located by *eplimited*, averaged over all detections in the window [156]. The asymmetric window is due to *wqrs* detecting the Q-point, and *eplimited* detecting the R-point. Since the *wqrs* algorithm is far more sensitive to noise than the *eplimited* algorithm [158], disagreements in detection are assumed to indicate the presence of noise.

(ii) kSQI - The fourth moment (kurtosis) of the distribution of the signal. Kurtosis measures the relative “peakedness" of the probability distribution with respect to the Gaussian distribution. It is defined as $kSQI = E\left[\{X-\mu\}^4\right]/\sigma^4$, where X is the signal vector considered as the random variable, μ is the mean of X, σ is the standard deviation of X, and $E\left[\{X-\mu\}^4\right]$ is the expected value of the quantity of $\{X-\mu\}^4$. It is expected

that an ECG with good diagnostic quality to be non-Gaussian since there is assumed to be little random noise on the signal (and which is typically Gaussian in distribution). Random noise from sources such as muscle artefacts and baseline wander tend to have Gaussian distribution [12].

(iii) sSQI - The third moment (skewness) of the signal which describes its asymmetry and is defined as $sSQI = E\left[\{X-\mu\}^3\right]/\sigma^3$. A good diagnostic ECG signal is expected to be highly skewed due to the presence of the QRS complex.

(iv) pSQI - The relative power in the QRS complex, estimated as $\frac{\int_5^{15} P(f)df}{\int_5^{40} P(f)df}$, where P is the power and f is the frequency in Hz. It is expected that most of the power of a clean ECG to be in the 5-15 Hz band [12].

(v) basSQI - The relative power in the baseline as represented by $1-\frac{\int_0^{1} P(f)df}{\int_0^{40} P(f)df}$. Low basSQI indicates abnormal drift (≤ 1 Hz) in the baseline of an ECG signal.

(vi) qSQI - The proportion of QRS complexes detected (number of QRS complexes detected by the *eplimited* algorithm, over the number of QRS complexes detected by the *wqrs* algorithm for a given 5-second window).

(vii) fSQI - The proportion of the signal which is not a flat line, where a flat line is defined as being consecutive samples which differ by less than 0.1 mV. A value of 0 indicates the ECG signal is a flat line.

Thus for annotators $j = 1,...,R$, the features $\mathbf{x}$ of the ith record can be formulated as:

$$\mathbf{x}_i^{\mathsf{T}} = \left[\sqrt{\tilde{RR}}_i^{j=1}, \mathbf{SQI}_i^{j=1}, ..., \sqrt{\tilde{RR}}_i^{j=R}, \mathbf{SQI}_i^{j=R}\right]. \tag{4.17}$$

where $\mathbf{SQI} = [bSQI, kSQI, sSQI, pSQI, basSQI, qSQI, fSQI]$, and which was estimated over a 5-second window around the annotated beat as previously described. The $\tilde{RR}$ is the median of the five preceding R-R intervals of the annotated beat. Note that features for the ith

record were described as annotator-specific to represent the case of $x_i^j = NaN$ when missing annotation was provided by jth annotator.

4.3.3 Methodology of Comparison

The EM-R algorithm was applied to Division 1 (manual annotations), Division 2 (closed-source algorithms), Division 3 (open-source algorithms), and Division 4 (closed- and open-source algorithms) to synthesise all of the annotations into a single estimated ground truth. The RMSEs were calculated using the reference provided for the QT dataset, and compared with the best-performing scores that were published for that dataset. In addition, the EM-R algorithm was compared to the median- and mean-voting strategies in the scenario of selecting different numbers of annotators. The steps of the EM-R algorithm are summarised in Fig. 4.3 with option (A).

A sub-sampling method was performed as described in the figure: K annotators were randomly selected 100 times, and this was repeated with K varied from three to the maximum number of annotators in each division. The mean and standard error ($\mu \pm \sigma_\mu$ms) of the RMSE of EM-R, the mean- and median-voting approaches were calculated and compared. The mean absolute error (MAE) of the QT annotations was also calculated as it provides a reliable interpretation of the true distance between the estimated QT intervals and the reference (with a resolution of 1 ms). The Jarque-Bera test for normality was applied to the distributions and a significant difference from a normal distribution was found for all data ($p < 0.01$). A two-sided paired t-test ($p < 0.05$) was applied to the 100 sub-sampled RMSEs and MAEs which were drawn from a normal distribution. This included comparison for (i) the EM-R algorithm with and without features; (ii) the EM-R algorithm with only beat-specific heart rate and SQIs (bHRSQIs) features versus the EM-R algorithm with the beat-specific heart rate (bHR) feature; (iii) EM-R with the bHRSQIs features versus the mean

Require: Annotation $y_i^j \in \mathbb{R}$ ▷ for $j = 1, \cdots, R$ and $i = 1, \cdots, N$ from R annotators on N records.
repeat
  ▷ Select random K from R annotators with option (A) or (B)
    Extract $\mathbf{x}^j$ using eqn (4.17), for $j = 1, \cdots, K$ annotators.
    Initialise λ^j, for $j = 1, \cdots, K$ annotators.
    Initialise $\hat{z}_i = \frac{1}{K}\sum_{j=1}^{K} y_i^j$, $\forall i = 1, \cdots, N$.
  repeat
    ▷ E step
      Estimate $\hat{z}_i$ using eqn (4.10), $\forall i = 1, \cdots, N$.
    ▷ M step
      Estimate $\mathbf{w}$ from eqn (4.11).
      Update λ^j using eqn (4.12), $\forall j = 1, \cdots, K$.
  until convergence of $\boldsymbol{\lambda}$.
  Estimate RMSE using $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{z}_i - y_{r,i})^2}$, where $y_{r,i}$ is the reference provided for y_i.
until 100 runs are reached.
return $\mu \pm \sigma_\mu$ of the RMSE for K annotators.

Fig. 4.3 Pseudo code for the EM-R algorithm with features: option (A) selects K random annotators with missing annotations (i.e., who did not annotate all records), whereas option (B) selects K random annotators who provided K annotations for each record.

and the median strategies. The average percentage of records with full annotations was also measured when selecting a different number of annotators.

Furthermore, the "Meta-6" algorithm [37] was created in the PCinC Challenge to aggregate the annotations of the six best-performing automated algorithms which we do not consider here. The "Meta-6" algorithm used the outputs of the three best-performing algorithms in each of Divisions 2 and 3, and removed any records that were rejected by more than one of these six algorithms, as well as those for which the base algorithms disagreed the most. The median-voting approach was then used to fuse annotations for the remaining records. The QT annotations inferred by the "Meta-6" algorithm were also compared to those of the manual entries. In this chapter, the annotations provided for the "Meta-6" algorithm were obtained from the original authors [37]. Any annotations that were not in the first 5-second segment were discarded in the process of QT estimation to avoid inter-beat variations. A total

of 521 records were considered for comparing the performance of EM-R with the median and mean strategies when using the six best-performing automated algorithms.

In the Challenge, not all records were labelled by each annotator. A second method of comparison was introduced to assess the performance of EM-R for varying numbers of annotations. To that purpose, only annotators without missing annotations (i.e., who actually annotated all records) were considered. The steps of the EM-R process when selecting annotators without missing annotations are shown in Fig. 4.3 with option (B).

A final last analysis involved examining the agreement of the QT interval measurement between annotators and the reference provided for individual records. The best-performing participants from the Challenge were picked to represent manual and automated entries respectively. The QT intervals estimated by EM-R were generated by selecting all annotators since their combination resulted the least RMSE. The Bland-Altman plots of EM-R and the best individual participant against the reference QT intervals were produced for both the manual and automated divisions.

4.4 Results

The minimum RMSE values when estimating the QT interval of EM-R with and without features were calculated and compared with the results of the mean and median strategies, averaging over 100 sub-selections of annotators. The minimum RMSEs and the MAEs across all records are displayed in Table 4.1. The RMSE and MAE results of selecting three and nine annotators/algorithms are shown in Table 4.2.

The convergence of the EM algorithm is dependent on the number of records, the number of QT annotations, and the number of features used. It took approximately 1.37 seconds for the EM algorithm to converge (15 iterations) when considering 69 automated annotators for 548 records (for a total of 28,661 QT annotations) with the bHRSQIs features using MATLAB on a system with a 3.3GHz Intel Xeon CPU.

Table 4.1 RMSEs and MAEs for different voting strategies in the PCinC QT dataset.

Division	**RMSE (ms)**					
	Mean	**Median**	**EM-R with**			
			No features (bNF)	**Feature SQIs (bSQIs)**	**Feature HR (bHR)**	**Feature HR+SQIs (bHRSQIs)**
1 (20 humans)	8.72±0.33‡	4.71±0.25‡	6.87±0.23	6.65±0.23†	6.22±0.22†	6.04±0.23†*
2 (48 algorithms)	16.20±0.54‡	15.31±0.55‡	15.03±0.50	14.85±0.49	14.61±0.48†	14.58±0.48†
3 (21 algorithms)	30.52±1.49‡	18.99±0.71‡	18.87±0.71	17.80±0.65†	16.66±0.59†	16.58±0.58†
4 (69 algorithms)	17.67±0.56‡	14.44±0.52	14.74±0.50	14.24±0.47†	14.12±0.46†	13.97±0.46†
Division	**MAE (ms)**					
	Mean	**Median**	**EM-R with**			
			No features (bNF)	**Feature SQIs (bSQIs)**	**Feature HR (bHR)**	**Feature HR+SQIs (bHRSQIs)**
1 (20 humans)	6.24±0.24‡	3.06±0.15‡	4.87±0.19	4.71±0.18†	4.42±0.17†	4.26±0.16†*
2 (48 algorithms)	12.64±0.42‡	11.74±0.41	11.74±0.37	11.64±0.37	11.39±0.38†	11.36±0.37†
3 (21 algorithms)	23.00±0.77‡	14.10±0.59‡	14.16±0.53	13.36±0.50†	13.57±0.47†	12.50±0.48†
4 (69 algorithms)	14.23±0.38‡	11.21±0.38‡	11.45±0.37	11.06±0.36†	10.98±0.36†	10.76±0.35†*

The † (for columns 5 to 7 only) indicates the EM-R results were significantly different from those with no features ($p < 0.05$). The * (for column 7 only) indicates that results obtained using the bHRSQIs feature set were significantly different from those obtained with the bHR feature ($p < 0.05$). The ‡ (for columns 2 to 3 only) indicates results obtained using the bHRSQIs feature set were significantly different from the mean and median strategies ($p < 0.05$).

Table 4.2 RMSEs and MAEs of using three and nine manual/automated annotators for different voting strategies.

Division	**RMSE (ms)**					
	Three annotators/algorithms			**Nine annotators/algorithms**		
	Mean	**Median**	**EM-R with bHRSQIs**	**Mean**	**Median**	**EM-R with bHRSQIs**
1	19.16±1.62	17.99±1.41	16.07±1.45‡	12.23±0.712	9.07±0.75	8.72±0.48‡
2	31.17±4.62	30.56±4.76	28.59±4.86‡	21.28±1.84	19.40±1.86	18.35±1.81‡
3	71.40±9.26	68.62±9.51	58.74±10.36‡	42.53±3.01	29.31±2.659	25.36±1.96‡
4	48.50±8.50	46.23±8.91	39.59±9.75‡	27.32±2.205	21.05±2.15	20.20±3.30‡
Division	**MAE (ms)**					
	Three annotators/algorithms			**Nine annotators/algorithms**		
	Mean	**Median**	**EM-R with bHRSQIs**	**Mean**	**Median**	**EM-R with bHRSQIs**
1	12.55±0.58	11.17±0.58	10.51±0.49‡	8.49±0.39	5.59±0.31	6.07±0.26‡
2	20.39±1.01	19.68±1.01	18.58±0.92‡	15.43±0.57	13.83±0.51	13.44±0.53‡
3	43.46±2.47	40.31±2.36	34.17±2.05‡	29.98±1.24	20.29±0.87	18.31±0.73‡
4	27.68±1.70	25.51±1.65	21.98±1.42‡	19.35±0.81	14.85±0.66	14.48±0.60‡

The ‡ indicates the EM-R results with the bHRSQIs feature set were significantly different from the mean and median strategies ($p < 0.05$).

The EM-R RMSEs and MAEs with no features were significantly ($p < 0.05$) larger than when using EM-R with the single feature of bHRSQIs (Table 4.1). The RMSEs and MAE results of EM-R with the bHRSQIs features were significantly smaller than those of the mean voting strategy in all entries. When considering human annotators, EM-R with the bHRSQIs features had a significantly larger RMSE (1.33$\pm$0.34 ms) than the median voting strategy, and a similar result was obtained for MAE (1.20$\pm$0.22 ms). In automated entries, EM-R with the bHRSQIs features outperformed the median strategy, but only the MAE result was significant ($p < 0.05$). When considering all participants, EM-R with the bHRSQIs features achieved a MAE of 4.26$\pm$0.16 ms for human annotators and a MAE of 10.76$\pm$0.35 ms for automated algorithms.

Figure 4.4 displays the average performance of each method across 100 randomly selected subsets of annotators. In Fig. 4.4A (a), the EM-R RMSE was 6.04 ms when considering all human annotators in Division 1. It outperformed the mean voting strategy for all numbers of annotators but was worse than the median voting approach after nine annotators (median RMSE = 4.71 ms for 20 annotators). Using 15 out of 20 annotators (RMSE = 6.62$\pm$0.98 ms), EM-R achieved a similar error as the best score (RMSE = 6.65 ms) provided in the Challenge. In Division 2 (see Fig. 4.4A (b)), EM-R consistently outperformed the other two approaches (15.31 ms for the median voting, and 16.20 ms for the mean voting) and achieved the minimum RMSE on Division 2 of 14.58 ms when considering all 48 annotators. Seventeen annotators (RMSE = 15.68$\pm$1.83 ms) from Division 2 would be required for EM-R to generate a similar RMSE as the best-performing algorithm (RMSE = 16.36 ms). In Fig. 4.4A (c) for Division 3, EM-R continued to outperform the other two approaches and achieved a RMSE of 16.58 ms when considering 21 annotators. It also achieved a lower RMSE when compared with the best-performing annotator (RMSE = 17.46 ms) in this division. EM-R had an RMSE of 13.97 ms for Division 4, which was lower than the RMSEs from any entrant in the two divisions from which it was created.

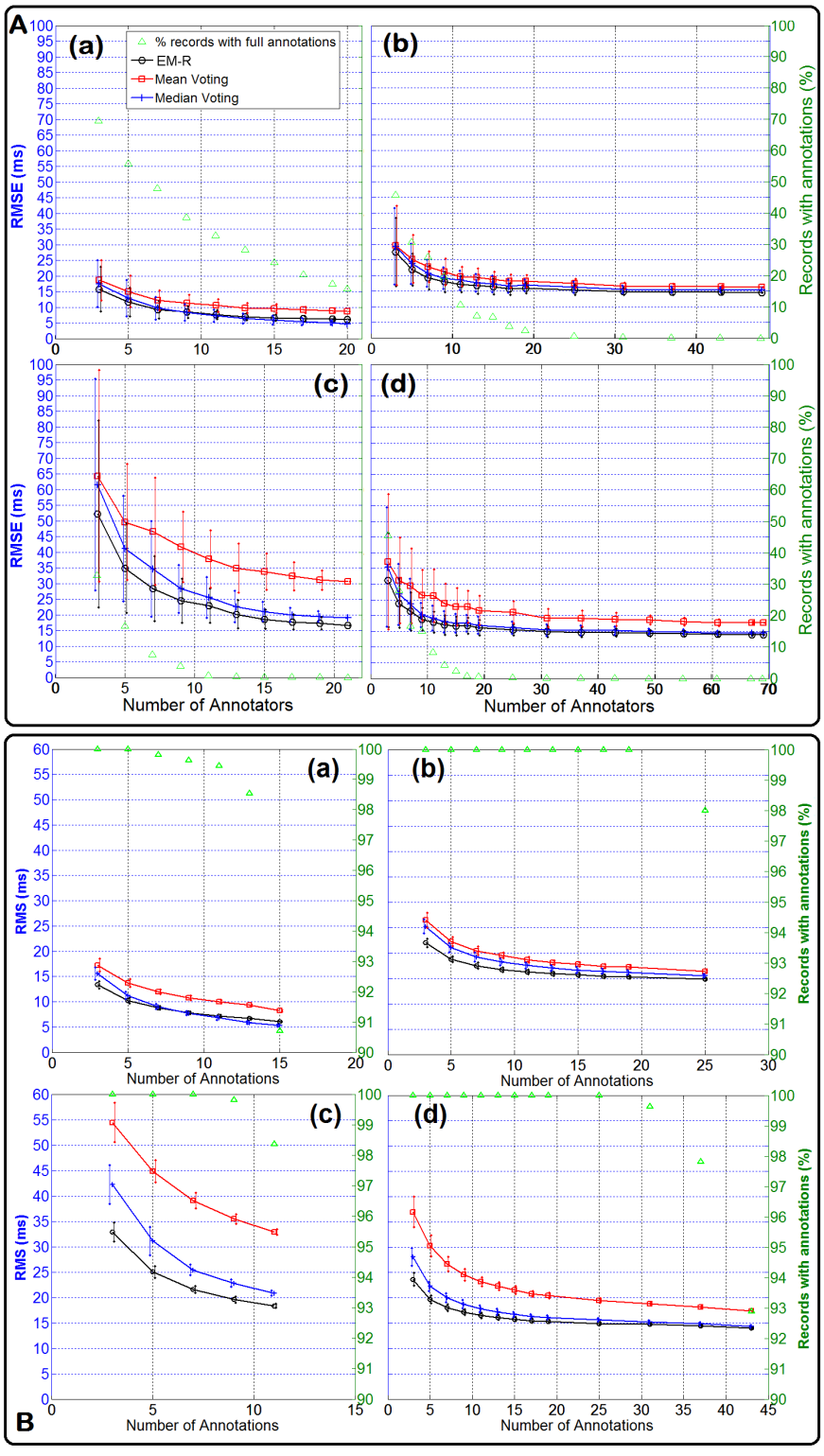
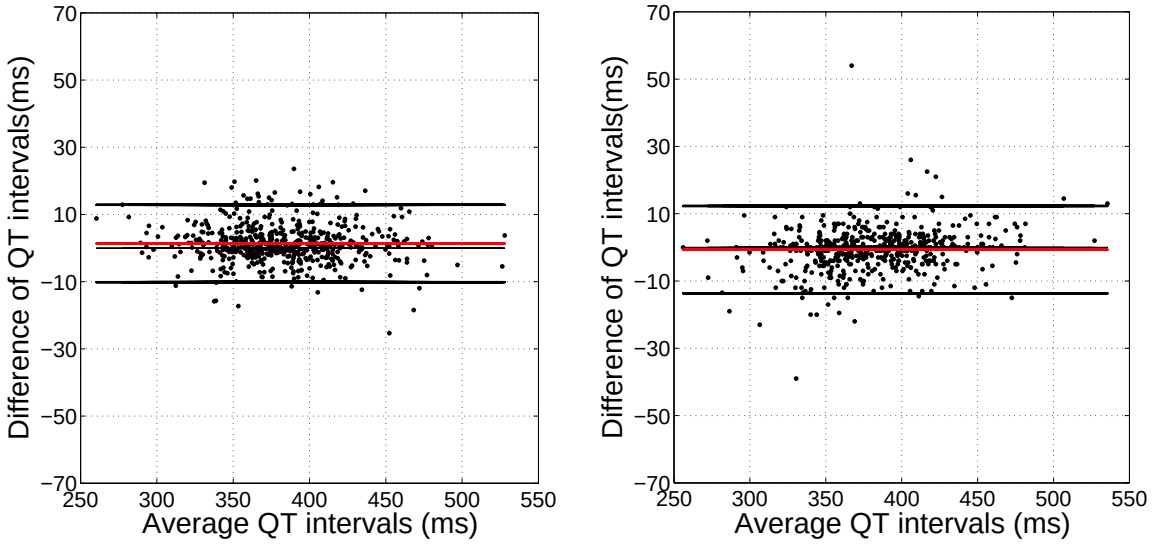


Fig. 4.4 Mean and standard error of the RMSE results of using EM-R with the bHRSQIs features, median and mean voting are shown for different divisions. The x-axis indicates selection of number of annotators with no replacement in Box A, whereas Box B shows selection of annotations for up to 90% of records. These plot were generated from 100 iterations of random samples of annotators. The left y-axis indicates the mean and standard error of RMSE in ms and the right y-axis indicates the average percentage of records with full annotations as shown as △ in (a) Division 1; (b) Division 2; (c) Division 3; and (d) Division 4.

Figure 4.4B compares the EM-R performance when only considering records fully annotated by selected annotators. EM-R achieved a minimum RMSE of 10.22±0.50 ms for Division 1 using all records as shown in Fig. 4.4B (a). It initially outperformed the other two strategies, but was slightly worse than the median voting approach when considering more than nine annotators with full annotations. After a slight reduction in the percentage of records being annotated (98.54% when selecting 13 annotations provided by 13 annotators), EM-R achieved a similar score (RMSE = 6.72±0.17 ms) as the Challenge's best-performing annotator in Division 1. In Division 2 with 100% records being annotated at 19 annotators (ie., 19 annotations as shown in Fig. 4.4B (b)), EM-R achieved a minimum RMSE of 15.26±0.24 ms, and this was lower than the other two strategies (16.12±0.51 ms for the median approach and 17.17±0.30 ms for the mean approach). In Division 3 (Fig. 4.4B (c)), the minimum EM-R RMSE of 19.68±0.55 ms was estimated with 99.82% records being annotated, and again it outperformed the other two strategies (22.72±0.50 ms for the median approach and 35.45±0.99 ms for the mean approach). With 96.17% records annotated by 12 annotators providing 12 annotations, EM-R achieved a RMSE of 17.35±0.31 ms, which was similar to the winning score in the Challenge. Finally in Division 4 (Fig. 4.4B (d)), EM-R achieved a minimum RMSE of 14.73±0.19 ms when selecting 31 annotators with full annotations for 99.54% records, which was an improvement over Division 2 and 3.

In the case of considering annotations submitted by the six best-performing algorithms selected by the "Meta-6" algorithm, EM-R with the bHRSQIs features achieved a RMSE of 9.946 ms, which was similar to the mean strategy (RMSE = 9.948 ms), but which outperformed the median strategy (RMSE = 10.151 ms).

The Bland-Altman plot of EM-R for manual annotators is shown in Fig. 4.5 (a), and the corresponding results of the best-performing manual annotator are shown in Fig. 4.5(b). The Bland-Altman plot of EM-R for automated algorithms is shown in Fig. 4.6(a), and the corresponding results of the best-performing automated algorithm are shown in Fig. 4.6



(a) (b)

Fig. 4.5 Bland-Altman plots of the QT differences in Division 1 (all manual labels) comparing the reference with the estimated QT intervals using (a) EM-R with the bHRSQIs features and (b) the best-performing annotator. The mean of the QT differences is shown in red and the 95% agreement limit is shown in black.

(b). The Bland-Altman plots for Division 1 (see Fig. 4.5) indicated that the QT intervals estimated from both EM-R and the best-performing annotator were in close agreement with the reference. The mean bias of EM-R was 1.36 ms with a 95% confidence interval of $\pm$11.52 ms, whereas the bias of the best-performing annotator was -0.68 ms with a 95% confidence interval of $\pm$12.98 ms. For Division 4, both plots indicated an over-estimation of the QT intervals when compared with the reference. The mean difference of EM-R was 7.96 ms with a 95% confidence interval of $\pm$22.43 ms, whereas the mean difference of the best-performing annotator was 5.83 ms with a 95% confidence interval of $\pm$28.22 ms.

4.5 Discussion

As compared to the median and mean approaches, EM-R provided a robust and reliable methodology to fuse labels into a single estimate of the underlying ground truth. As the number of annotators increased, the proportion of records that they each annotated decreased.

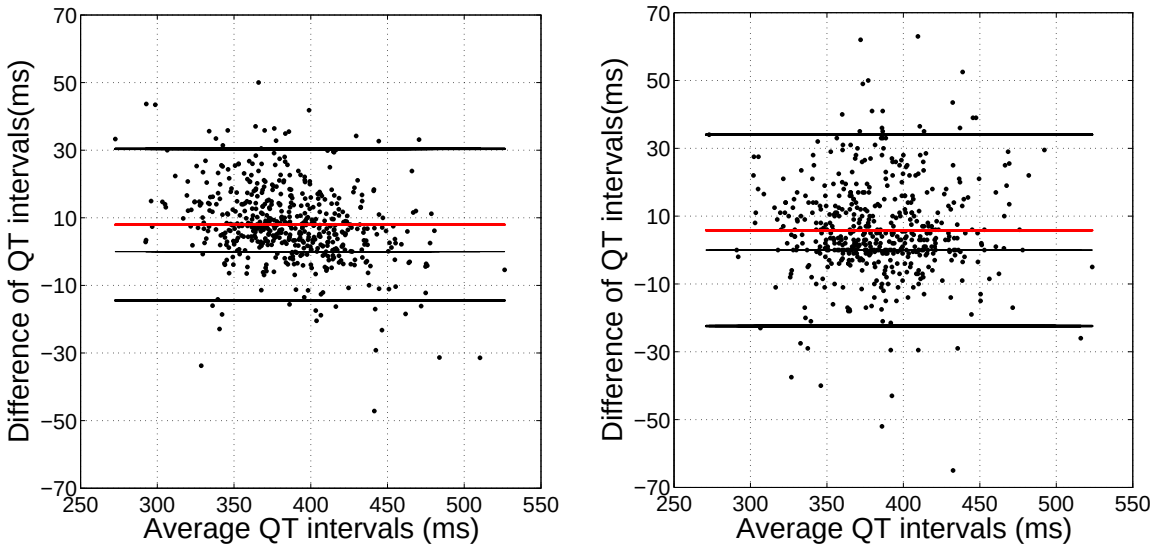


Fig. 4.6 Bland-Altman plots of the QT differences in Division 4 (all automated labels) comparing the reference with the estimated QT intervals using (a) EM-R with the bHRSQIs features and (b) the best-performing annotator. The mean of the QT differences is shown in red and the 95% agreement limit is shown in black.

Nevertheless, EM-R estimates of the QT intervals continued to improve, indicating that it was still able to use the additional information provided by the new annotators, as shown in Fig. 4.4A. Less intuitively, when ensuring that all records had been annotated (Fig. 4.4B), the RMSE begins to rapidly decrease as the proportion of records which are fully annotated decreases. This is due to the exclusion of these insufficiently-annotated records. It was hypothesised that the reduction of the number of records reduces the variability of the QT measurements, perhaps due to the fact that intervals which are simpler to label are annotated more frequently.

For manual annotations (i.e., Division 1), the RMSE of the labels obtained from EM-R was consistently smaller than that of the mean voting strategy but was slightly larger than that of the median voting strategy when using more than nine annotators. However, the reference QT intervals in this dataset were provided based on bootstrapping the median of 15 human annotators. This in some part accounts for the reduced performance of EM-R on Division 1 when compared to the median voting strategy using nine or more human annotators. It is

interesting to note the bias of the reference against the EM-R estimates: the QT intervals predicted by EM-R are longer than those of the human annotated reference for both Division 1 and Division 4. One possible explanation for this is a human bias to underestimate the QT interval [12]. The reference is based upon human annotations, and the errors provided in the automated entries were always larger than those in the manual entries. This may indicate that the larger errors for Divisions 2-4 are due to bias in the QT interval reference values. Nevertheless, the EM-R annotations were still consistently better than the mean and median approach for small numbers of annotators (e.g., three), the most common situation in medical labelling tasks. In this situation, EM-R provided substantial performance gains (10.7% for selecting three human annotators in Divison 1, and 14.4% for selecting three automatic algorithms in Division 4).

Table 4.1 evaluates multiple instantiations of EM-R which include and exclude contextual information regarding the record. The inclusion of heart rate significantly increases the accuracy of EM-R when compared to those with no features (bHR versus bNF, $p < 0.05$), indicating that constraining the QT interval to physiologically-reasonable values, as determined by the heart rate, leads to improved performance. EM-R with the bHRSQIs features was not always significantly different from that using the bHR feature alone. This was expected as the recordings were extracted from a database of high-quality 12-lead ECGs with very little noise contamination, allowing participants or algorithms to select a relatively clean sub-segment from which to estimate the QT interval. Hence, the effect of including SQIs would not be expected to be large. Note that the seven SQIs were highly correlated, limiting traditional approaches to coefficient significance testing for feature selection. Furthermore, certain SQIs are expected to be useful in only a subset of scenarios (e.g., basSQI being useful for humans who are sensitive to T-wave elevation, in contrast to automated methods, which are not so sensitive), and so the significance of each coefficient for the features $\mathbf{X}$ would be too variable for reasonable interpretation. To provide quantification of this assertion,

EM-R using a subset containing 80% of the data was re-ran. Since the features exhibit some collinearity (but not perfect collinearity), it was expected that the coefficients would be relatively unstable in values, but the overall model performance to be stable. Indeed, upon evaluating 100 repetitions of model development using a random subset of the data selected at random each time, it was observed that the coefficients varied between -1000 and 1000, while the RMSE remained stable (Division 1: 5.98 ± 0.10 ms, Division 2: 13.60 ± 0.11 ms, Division 3: 17.24 ± 0.21 ms, and Division 4: 13.22 ± 0.13 ms). Thus, no feature selection was performed and all seven SQIs were included.

Table 4.2 shows that the RMSE and MAE results for EM-R with the bHRSQIs features significantly ($p < 0.05$) improved over the mean and median voting strategies for all automated entries, even with a limited number of annotators available. For human annotators, EM-R with the bHRSQIs features outperformed the other two voting strategies when selecting three annotators. However, its MAE was slightly worse than the median voting strategy when using nine annotators.

It is common practice that three annotators are used to identify QT interval length, and that they may collaborate for labelling the QT interval. This type of direct collaboration often results in an incorrect weight for one annotator, and is time-consuming. In the situation where the annotators operate independently, employing EM-R provides improved performance over the baseline methods: (i) in the case of selecting three annotators randomly with possible missing annotations, we observed a 10.7% improvement over the next-best voting strategy for manual annotations (see Fig. 4.4A (a)), and 14.4% for automated annotations (see Fig. 4.4A (d)); (ii) in the case of selecting three annotators with full annotations, a 14.7% improvement over the next-best voting strategy for manual annotations (see Fig. 4.4B (a)), and 15.7% for automated annotations (see Fig. 4.4B (d)) was observed. Thus, combining labels from human annotators using EM-R could potentially provide an optimal "gold standard" in QT estimations even in the case when the ground truth is not available.

The Bland-Altman plots show that the QT intervals estimated using EM-R were in a closer agreement with the reference than those of the best-performing annotators in both human and automated divisions, with smaller QT differences and a narrower range of agreement limits (see Fig. 4.5 and Fig. 4.6). It was also observed that the means in the Bland-Altman plot (Fig. 4.6) were not close to zero. This implies that comparing the QT interval estimate of automated algorithms with the human reference always results in an overestimation and again is in agreement with previous studies [27]. Nevertheless, EM-R provided smaller QT differences from the reference as compared to the best-performing algorithm.

Comparing to the prior label aggregation model, "Meta-6" (RMSE = 10.15 ms), EM-R appears to have substandard performance. However, these results are misleading as the "Meta-6" algorithm excluded 27 records (4.93%) based on arbitrary disagreements between annotators, whereas EM-R considered all records that were available (except one which had no signal). Furthermore, in the case when using the annotations of the selected records provided by the exact six automated algorithms used in the "Meta-6" algorithm, EM-R achieved a smaller RMSE of 9.95 ms. This was similar to the mean voting strategy since the mean vote was the initial guess of the EM model. The development of the "Meta-6" algorithm required selecting annotators/algorithms based upon a known ground truth, making the approach infeasible for tasks where this ground truth is unavailable. Although the lowest EM-R RMSE (13.97$\pm$0.46 ms) in the automated division is larger than the best-performing human annotator in the Challenge (RMSE = 6.65 ms), there were only two other human annotators who achieved a score below 10 ms. Furthermore, as the annotations of automated algorithms in Division 4 are independently determined from the reference (as opposed to those in Division 1, from which the reference was derived), it is expected that automated algorithms would have worse performance when compared to human annotators.

4.6 Conclusion

The accuracy of estimating the QT interval for a single channel of ECG (lead II) using multiple annotators was analysed and compared in this chapter using different voting strategies. EM-R was shown to consistently outperform the median voting algorithms for fewer than nine human annotators and for any number of automated algorithms. For large numbers of annotators, EM-R estimates became approximately equal to the best-performing annotator, even though the identity of the best annotator was unknown. In addition, it outperformed the mean voting strategy in all circumstances. Therefore, EM-R has the potential to provide a more realistic reference when no "gold standard" exists. That is, since the probabilistic approach works by comparing inter-annotator variations, no absolute reference data is required and it can in fact be used to generate a reliable ground truth. The RMSE of 13.97 ± 0.46 ms when combining multiple algorithms (69 in total), is the lowest so far reported in the literature for non-human QT estimation. Using only 31 automated algorithms with 99.64% records annotated resulted in a RMSE of 14.73 ± 0.19 ms. EM-R with the bHRSQIs features not only addresses the issue of large inter-observer variability, but also provides an improvement in QT estimation as compared to the median and mean voting strategies when there are few annotators, which is the typical scenario in practice.

The key factor that improved EM-R performance was the addition of signal quality features. However, it was noted that the data used in this chapter are relatively low noise (which is the case for most QT studies), and it is therefore all the more surprising that signal quality provided a boost to EM-R performance. In general, most ECG data are far noisier that the data used here, and therefore EM-R would be expected to provide even more of an advantage in such situations. Future studies will aim to verify this conjecture.

The EM-R model of the annotators' behaviours currently does not factor in the possibility of a bias inherent to an individual annotator. Incorporating the bias of each annotator could

possibly allow for a better aggregation of the inferred ground truth, as annotators who are consistent but offset from the aggregated “truth” are heavily penalised in the current model.

Chapter 5

The Bayesian Continuous-Valued Label Aggregation Model

"The probable is what usually happens."

– Aristotle

5.1 Introduction

The previous chapter described a probabilistic Expectation-Maximisation method introduced by Raykar *et al.* (EM-R) [89] for labels aggregation to weight an algorithm or human based on its performance. It was shown that EM-R was unable to provide measure of a bias inherent to an individual annotator. To address this problem, a Bayesian Continuous-Valued Label Aggregator (BCLA) is proposed in this chapter to provide a reliable estimation of label aggregation while accurately inferring the precision and bias of each annotator. This chapter further proposes a maximum-a-posterior approach for solving the BCLA framework (denoted BCLA-MAP). We subsequently demonstrate its application to QT interval estimation. The use of the BCLA-MAP model will be compared to the results obtained from using the mean

and median aggregation approaches, along with the two benchmarking approaches described previously.

5.2 A Generative Model of Independent Annotators

5.2.1 The Ground Truth Model

The ground truth model proposed by Raykar *et al*. [89] was previously described in Chapter 4 section 4.2. We have extended this to incorporate a Gamma distribution[1] to constrain the precision of the ground truth as follows:

$$p\left(b \mid k_b, \vartheta_b\right) = \text{Gamma}\left(b \mid k_b, \vartheta_b\right), \tag{5.1}$$

where k_b is the shape parameter and ϑ_b is the scale parameter. The graphical representation of the ground truth is shown in Fig. 5.1.

5.2.2 The Annotator Model

In the EM-R model, it was assumed that y_i^j is a noisy version of z_i, with a Gaussian distribution $\mathcal{N}\left(y_i^j \mid z_i, (\sigma^j)^2\right)$. The same assumption is considered across all models in the thesis. The motivation for this comes from the central limit theorem: given the assumption that the annotations for a given annotator are independent and identically distributed, their residuals derived from the latent ground truth will converge to a Gaussian distribution. In the absence of prior knowledge, this assumption allows for a robust and generalisable model for the given data. Here σ^j is the standard deviation in annotation of the jth annotator and represents his variance in annotation around z_i. Furthermore, the bias of each annotator, defined as the

[1] A Gamma distribution can be defined as $\text{Gamma}(x \mid k, \vartheta) = \frac{1}{\Gamma(k)\vartheta^k} x^{k-1} \exp(-\frac{x}{\vartheta})$, where k is the shape of the distribution and ϑ is the scale of the distribution, $\Gamma(\cdot)$ is a gamma function. A Gamma distribution is commonly used to model positive continuous values [95] and it is therefore assumed that precision values are drawn from a Gamma distribution

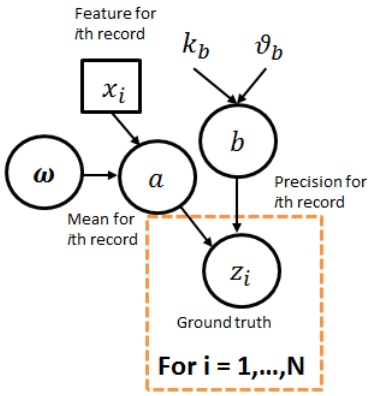


Fig. 5.1 Graphical representation of the ground truth model: z_i (the unknown underlying ground truth) corresponds to the *true* annotation for the ith record, from a total of N recordings. z_i is modelled by a Gaussian distribution with parameters mean a_i and variance $1/b$, where a can be a function of feature vector $\mathbf{x}_i$ as a linear regression function $f(\mathbf{w},\mathbf{x})$ with an intercept, and $\mathbf{w}$ being the coefficients of the regression. The precision value, b, is drawn from a Gamma distribution with parameters k_b, ϑ_b.

opposite of accuracy where it measures the average difference between the estimation and the *true* value, can be modelled as an additional term, denoted as ϕ^j [3]. The probability density function of estimating y_i^j can then be written as:

$$p\left(y_i^j \mid z_i, \phi^j, \lambda^j\right) = \mathcal{N}\left(y_i^j \mid z_i + \phi^j, 1/\lambda^j\right), \tag{5.2}$$

where $\left(\sigma^j\right)^2$ is replaced with $1/\lambda^j$. λ^j is the precision of the jth annotator, defined as the estimated inverse-variance of annotator j. Note that λ^j and ϕ^j are considered to be constants for the jth annotator, i.e., all annotators are assumed to have consistent performances throughout records. It is assumed that $y_i^1, \cdots, y_i^R$ are conditionally independent given the ground truth z_i, and under the assumption that records are independent, the conditional probability density function of $\mathbf{y}$ can be modelled as:

$$p(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\phi}, \boldsymbol{\lambda}) = \prod_{i=1}^{N} \prod_{j=1}^{R} \mathcal{N}\left(y_i^j \mid z_i + \phi^j, 1/\lambda^j\right). \tag{5.3}$$

This may not be necessarily true, especially in cases where the annotations are generated by algorithms, some of which may be partially dependent variations of the same approach. Nevertheless, this choice was made to drastically simplify the model and subsequent derivation of the likelihood. Furthermore, it is assumed that the probability density function of a given bias of annotator j, ϕ^j, drawn from a Gaussian distribution with mean μ_ϕ and variance $1/\alpha_\phi$ [130], is given by:

$$p\left(\phi^j \mid \mu_\phi, \alpha_\phi\right) = \mathcal{N}\left(\phi^j \mid \mu_\phi, 1/\alpha_\phi\right). \tag{5.4}$$

Although the biases of the annotators might be derived from other distributions, they are likely to be dataset dependent. In the absence of any knowledge of the underlying distribution of biases, they are assumed to be drawn from a Gaussian distribution. As described earlier that precision values can be modelled using a Gamma distribution, it is therefore assumed that precision values, such as λ^j and α_ϕ, were drawn from a Gamma distribution, with parameters k_λ, ϑ_λ, and k_α, ϑ_α respectively:

$$p\left(\lambda^j \mid k_\lambda, \vartheta_\lambda\right) = \text{Gamma}\left(\lambda^j \mid k_\lambda, \vartheta_\lambda\right). \tag{5.5}$$

$$p\left(\alpha_\phi \mid k_\alpha, \vartheta_\alpha\right) = \text{Gamma}\left(\alpha_\phi \mid k_\alpha, \vartheta_\alpha\right). \tag{5.6}$$

The graphical representation of the annotator model is shown in Fig. 5.2.

5.2.3 The Bayesian Continuous-Valued Label Aggregation (BCLA) Model

The BCLA model was created to combine the ground truth and annotator models (see section 5.2.1 and section 5.2.2 respectively). It comprises two key contributions: (i) BCLA provides an unsupervised estimation of the continuous-valued annotations that are valuable for time-series related data, as well as duration of events for physiological data; (ii) it introduces a unified framework for combining continuous-valued annotations to infer the underlying

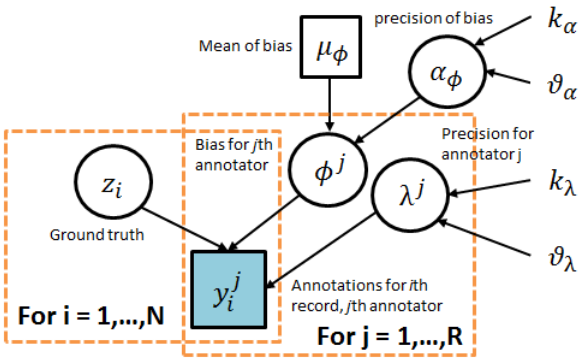


Fig. 5.2 Graphical representation of the annotator model: y_i^j corresponds to the annotation provided by the jth annotator for the ith record, and is modelled by the z_i (the unknown underlying ground truth), the ϕ^j (bias), and the λ^j (precision). Furthermore, ϕ^j is modelled from a Gaussian distribution with mean μ_ϕ and variance $1/\alpha_\phi$. The λ^j, and α_ϕ are drawn from a Gamma distribution with parameters k_λ, ϑ_λ, and k_α, ϑ_α, respectively.

ground truth, while jointly modelling annotators' bias and precision values. The graphical form of BCLA is presented in Fig. 5.3.

5.2.4 Simulated QT Dataset with Independent Annotators

As a demonstration of the application of BCLA in the context of ECG QT interval annotations, we here consider a simulated dataset of QT intervals. A total of 548 simulated records were generated, each with 20 independent annotators, thus providing a total of 10,960 annotations (see Fig. 5.4). The simulated dataset assumed that annotators have precision values, $\boldsymbol{\lambda} \sim \text{Gamma}(4, 0.0003)$, with the assumption that the annotations provided by the best performing annotator are ± 5 ms from the ground truth. Annotators' biases, $\boldsymbol{\phi} \sim \mathcal{N}(10, 25)$, a Gaussian distribution with 10 ms mean and a standard deviation $\alpha_\phi^{-1/2} = 25$ ms, assuming that the automated annotations tend to over-estimate manual annotations, as described in previous studies and discussed in the previous chapter [159–161]. The *true* annotation for each record, $z_i \sim \mathcal{N}(400, 40)$ [162, 163, 12], a Gaussian distribution with a mean $a_i = 400$ ms with a standard deviation $b^{-1/2} = 40$ ms. No particular features were considered in this

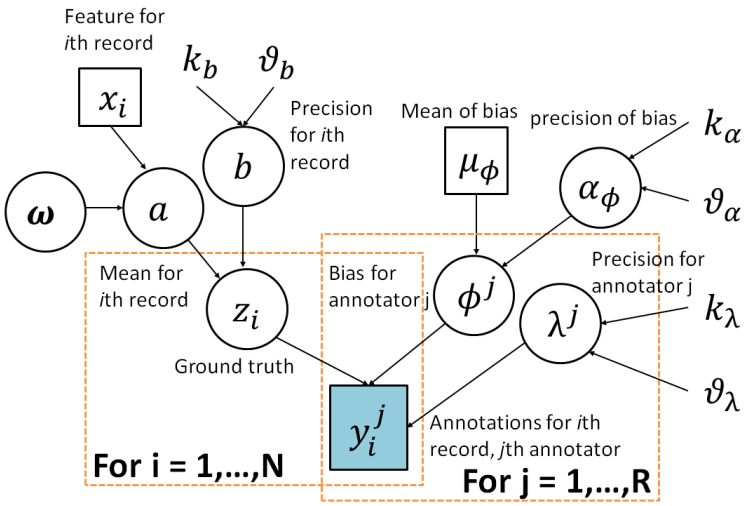
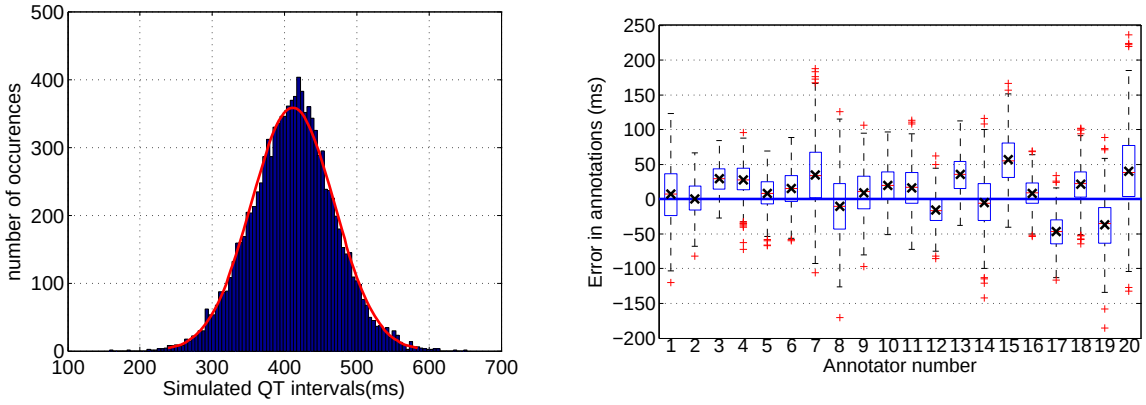


Fig. 5.3 Graphical representation of the BCLA model: y_i^j corresponds to the annotation provided by the jth annotator for the ith record, and is modelled by the z_i (the unknown underlying ground truth), the ϕ^j (bias), and the λ^j (precision). Furthermore, z_i is drawn from a Gaussian distribution with parameters mean a_i and variance $1/b$, where a can be a function of feature vector $\mathbf{x}_i$ as a linear regression function $f(\mathbf{w},\mathbf{x})$ with an intercept, and $\mathbf{w}$ being the coefficients of the regression. ϕ^j is modelled from a Gaussian distribution with mean μ_ϕ and variance $1/\alpha_\phi$. The b, λ^j, and α_ϕ are drawn from a Gamma distribution (denoted as Gamma) with parameters k_b, ϑ_b, k_λ, ϑ_λ, and k_α, ϑ_α respectively.

case (i.e., $x_i = 1$) for the purpose of illustrating the general use of the model. Furthermore, an intercept term in $f(\mathbf{w},\mathbf{x})$, w_0, was modelled in the feature (i.e., $\mathbf{x}_i := [1, x_i]$)). In addition, it was assumed that $\alpha_\phi \sim \text{Gamma}(3, 0.0005)$, ensuring the mean standard deviation where the biases drawn from is 25 ms. The $b \sim \text{Gamma}(3, 0.0002)$, ensuring the mean standard deviation where the *true* annotations drawn from is 40 ms.

The generated 10,960 annotations were then provided to the model to evaluate its accuracy in estimating the *true* annotation in an unsupervised manner, as well as predicting the bias and precision of each simulated annotator. The goal of this synthetic experiment is to determine if the BCLA model can recover the true bias and precision of each annotator, that are known in this experiment, and which were used to generate the synthetic data.



(a) (b)

Fig. 5.4 (a) the histogram of the simulated QT interval annotations for 548 records, with 20 annotations each (blue) provided by 20 simulated annotators. A fitted Gaussian distribution is superimposed (red solid line). (b) box plot of the error between the generated and *true* annotations for each of the 20 simulated annotators. The black '×' indicates the bias of each annotator. The span of each box represents the interquartile range over all annotations for each annotator.

5.3 BCLA-MAP

Suppose that there are N records of physiological time series data labelled by R annotators, as before. Let $\mathbf{D} = \left[\mathbf{x}_i^\intercal, y_i^{j=1}, \cdots, y_i^{j=R}\right]_{i=1}^{N}$, where $\mathbf{x}_i$ is a column feature vector for the ith record containing d features (i.e., the design matrix, $\mathbf{X} = \left[\mathbf{x}_\mathbf{1}^\intercal, ..., \mathbf{x}_\mathbf{N}^\intercal\right]$), y_i^j corresponds to the annotation provided by the jth annotator for the ith record, and z_i represents the unknown underlying ground truth as in previous sections (see Fig. 5.3). Under the assumption that records are independent, we remind ourselves that the likelihood of the parameter $\boldsymbol{\theta} = \{\mathbf{w}, \boldsymbol{\lambda}, \boldsymbol{\phi}, \alpha_\phi, b, z_i\}$ for a given dataset $\mathbf{D}$ can be formulated as:

$$p\left(\mathbf{D} \mid \boldsymbol{\theta}\right) = \prod_{i=1}^{N} p\left(y_i^1, \cdots, y_i^R \mid \mathbf{x}_i, \boldsymbol{\theta}\right). \tag{5.7}$$

The likelihood of the parameter $\boldsymbol{\theta}$ for a given dataset $\mathbf{D}$ can be written using Bayes' theorem as:

$$
\begin{aligned}
p(\boldsymbol{\theta} \mid \mathbf{D}) &\propto p(\mathbf{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}) \\
&= \text{Gamma}(\alpha_\phi \mid k_\alpha, \vartheta_\alpha)\text{Gamma}(b \mid k_b, \vartheta_b) \times \\
&\quad \left[\prod_{j=1}^{R} \mathcal{N}\left(\phi^j \mid \mu_\phi, 1/\alpha_\phi\right) \text{Gamma}(\lambda^j \mid k_\lambda, \vartheta_\lambda)\right] \times \\
&\quad \left[\prod_{i=1}^{N} \mathcal{N}\left(z_i \mid \mathbf{x}_i^\intercal \mathbf{w}, 1/b\right) \prod_{j=1}^{R} \mathcal{N}\left(y_i^j \mid z_i + \phi^j, 1/\lambda^j\right)\right]. \qquad (5.8)
\end{aligned}
$$

The estimation of $\boldsymbol{\theta}$ can be solved using a MAP approach, which maximises the log-likelihood of the parameters, i.e., $\underset{\boldsymbol{\theta}}{\text{argmax}}\{\log p(\boldsymbol{\theta} \mid \mathbf{D})\}$. The log-likelihood can be rewritten as:

$$
\begin{aligned}
\log p(\boldsymbol{\theta} \mid \mathbf{D}) = &-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{R}\left[\log\left(\frac{2\pi}{\lambda^j}\right) + \left(y_i^j - \phi^j - z_i\right)^2 \lambda^j\right] \\
&-\frac{1}{2}\sum_{j=1}^{R}\left[\log\left(\frac{2\pi}{\alpha_\phi}\right) + \left(\phi^j - \mu_\phi\right)^2 \alpha_\phi\right] \\
&-\frac{1}{2}\sum_{i=1}^{N}\left[\log\left(\frac{2\pi}{b}\right) + \left(z_i - \mathbf{x}_i^\intercal \mathbf{w}\right)^2 b\right] \\
&+\left[(k_\lambda - 1)\log\lambda^j - \log\left(\Gamma(k_\lambda)\vartheta_\lambda^{(k_\lambda)}\right) - \frac{\lambda^j}{\vartheta_\lambda}\right] \\
&+\left[(k_\alpha - 1)\log\alpha_\phi - \log\left(\Gamma(k_\alpha)\vartheta_\alpha^{(k_\alpha)}\right) - \frac{\alpha_\phi}{\vartheta_\alpha}\right] \\
&+\left[(k_b - 1)\log b - \log\left(\Gamma(k_b)\vartheta_b^{(k_b)}\right) - \frac{b}{\vartheta_b}\right]. \qquad (5.9)
\end{aligned}
$$

Parameters $\boldsymbol{\theta}$ can be derived by estimating the gradient of the log-likelihood respectively:

$$
\begin{aligned}
\frac{\mathrm{d}\log p(\boldsymbol{\theta} \mid \mathbf{D})}{\mathrm{d}\lambda^j} &= -\frac{1}{2}\sum_{i=1}^{N}\left[\left(y_i^j - \phi^j - z_i\right)^2 - \frac{1}{\lambda^j}\right] + \frac{k_\lambda - 1}{\lambda^j} - \frac{1}{\vartheta_\lambda}. \\
\frac{\mathrm{d}\log p(\boldsymbol{\theta} \mid \mathbf{D})}{\mathrm{d}\mathbf{w}} &= \frac{1}{b}\sum_{i=1}^{N}\left(z_i\mathbf{x}_i - \mathbf{x}_i^\intercal \mathbf{w}\mathbf{x}_i\right). \\
\frac{\mathrm{d}\log p(\boldsymbol{\theta} \mid \mathbf{D})}{\mathrm{d}\phi^j} &= \sum_{i=1}^{N}\lambda^j\left(y_i^j - \phi^j - z_i\right) - \phi^j\alpha_\phi + \mu_\phi\alpha_\phi.
\end{aligned}
$$

$$\frac{\mathrm{d}\log p(\boldsymbol{\theta} \mid \mathbf{D})}{\mathrm{d}\alpha_\phi} = \frac{R}{2\alpha_\phi} - \frac{1}{2}\sum_{j=1}^{R}\left(\phi^j - \mu_\phi\right)^2 + \frac{k_\alpha - 1}{\alpha_\phi} - \frac{1}{\vartheta_\alpha}.$$

$$\frac{\mathrm{d}\log p(\boldsymbol{\theta} \mid \mathbf{D})}{\mathrm{d}b} = \frac{N}{2b} - \frac{1}{2}\sum_{i=1}^{N}\left(z_i - \mathbf{x}_i^\intercal \mathbf{w}\right)^2 + \frac{k_b - 1}{b} - \frac{1}{\vartheta_b}.$$

By equating derivatives to zero, the parameters in $\boldsymbol{\theta}$ can be derived as:

$$\frac{1}{\lambda^j} = \frac{1}{N + 2(k_\lambda - 1)}\left[\sum_{i=1}^{N}\left(y_i^j - \phi^j - z_i\right)^2 + \frac{2}{\vartheta_\lambda}\right]. \tag{5.10}$$

$$\mathbf{w} = \left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^\intercal\right)^{-1}\sum_{i=1}^{N}\mathbf{x}_i z_i. \tag{5.11}$$

$$\phi^j = \frac{1}{N + \frac{\alpha_\phi}{\lambda^j}}\left[\sum_{i=1}^{N}\left(y_i^j - z_i\right) + \mu_\phi\left(\frac{\alpha_\phi}{\lambda^j}\right)\right]. \tag{5.12}$$

$$\frac{1}{\alpha_\phi} = \frac{1}{R + 2(k_\alpha - 1)}\left[\sum_{j=1}^{R}\left(\phi^j - \mu_\phi\right)^2 + \frac{2}{\vartheta_\alpha}\right]. \tag{5.13}$$

$$\frac{1}{b} = \frac{1}{N + 2(k_b - 1)}\left[\sum_{i=1}^{N}\left(z_i - \mathbf{x}_i^\intercal \mathbf{w}\right)^2 + \frac{2}{\vartheta_b}\right]. \tag{5.14}$$

This parameter estimation can be performed using the EM algorithm in a two-step iterative process (see Appendix A.2 for details):

(i) The E-step estimates the expected *true* annotations for the ith record, $\hat{z}_i$, in our MAP formulation as being a weighted sum of the provided annotations, and which can be estimated as:

$$\hat{z}_i = \frac{\sum_{j=1}^{R}\left[\left(y_i^j - \phi^j\right)\lambda^j\right] + \left(\mathbf{x}_i^\intercal \mathbf{w}\right)b}{\sum_{j=1}^{R}\lambda^j + b}. \tag{5.15}$$

(ii) The M-step is based on the current estimation of $\hat{\mathbf{z}}$ and the dataset $\mathbf{D}$. The model parameters, $\mathbf{w}$, $\boldsymbol{\phi}$, α_ϕ, b, and $\boldsymbol{\lambda}$ can be updated using equations (5.11), (5.12), (5.13), (5.14), and (5.10) accordingly in a sequential order until convergence, which is now described.

5.3.1 Convergence Criteria for BCLA-MAP using Extreme Value Theorem

Extreme Value Theorem

Extreme value theorem (EVT) is generally used to describe the modelling of the distribution of extreme values (being either maxima or minima): if a function is continuous and contained in a closed interval, then it has a maximum and a minimum value. According to Fisher-Tippett theorem, given that there are m independent, identically distributed random values (i.e., $\mathbf{x} = [x_{i=1}, \cdots, x_{i=m}]$) that are observed from a function $F(x)$, $x_{max} = \max(\mathbf{x})$ can be modelled using a family of extreme value distributions such as Gumbel, Fréchet, and Weibull distributions [164].

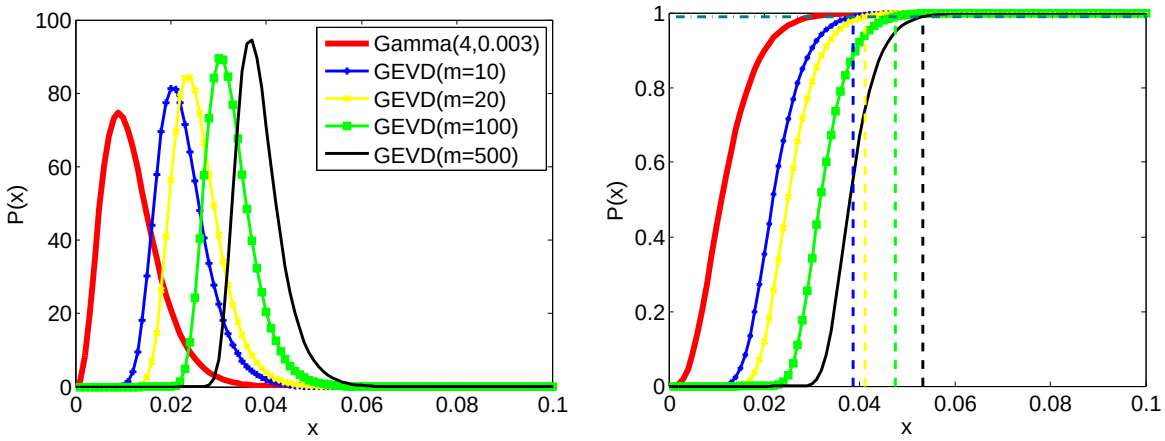
EVT for the BCLA-MAP Model

When using the EM algorithm for a MAP-based model, one may encounter a convergence problem, particularly when estimating a large number of parameters. The estimation of the precision λ^j may lead to values that tend to infinity because the model favours the annotator with the highest precision in each EM update step, while maximising the likelihood. Instead of incorporating an additional parameter for a regularisation penalty that increases with the complexity of the model, the generalised extreme value distribution (GEVD)[2] can be used to model the maxima of the precision distribution, denoted as λ_m, in order to restrict the upper bound of the precision values and guarantee convergence of the MAP algorithm. The probability density function of the GEVD for λ_m is:

$$p(\lambda_m \mid k, \mu, \vartheta) = \exp\left\{-\left[1 + k\frac{(\lambda_m - \mu)}{\vartheta}\right]^{-\frac{1}{k}}\right\} \frac{1}{\vartheta}\left[1 + k\frac{(\lambda_m - \mu)}{\vartheta}\right]^{\left(-1-\frac{1}{k}\right)}, \quad (5.16)$$

where k is a shape parameter, ϑ is a scale parameter, and μ is a location parameter. These

[2]GEVD combines Gumbel, Fréchet, and Weibull distributions into a single form.



(a) Probability density function (b) Cumulative density function

Fig. 5.5 An example of estimating the $\lambda_m \sim \text{Gamma}(4, 0.003)$: (a) shows the fitted GEVD corresponding to drawing m = 10, 20,100, and 500 samples from the Gamma distribution; (b) demonstrates the cumulative density function of the GEVDs with values (as dash vertical line) corresponding to the $99th$ percentile (as dotted horizontal line) for different m values.

parameters can be derived by fitting a GEVD to the maximum values drawn randomly from the *prior* distribution of the precision, $\text{Gamma}(\lambda \mid k_\lambda, \vartheta_\lambda)$. An upper bound of the maximum precision value can then be obtained by estimating $F(\lambda_m) = 0.99$ probability on the inverse cumulative distribution function $F(\lambda_m)$ of the GEVD. Fig. 5.5 demonstrates an example of drawing the maxima of the precision distribution using different sample size (denoted as m) values. Fig. 5.5 (a) shows the majority of the probability density described by the GEVDs is shifted toward higher values on the x-axis as m increases: it is expected to have higher value of λ_m as more samples are drawn from the Gamma distribution (see Fig. 5.5 (b)). The ideal sample size however is application-dependent [165]: As the threshold values are monotonically increasing with increasing m, the GEVD can include ever more extreme values which might be outliers in a given dataset. In the context of measuring ECG QT/QTc prolongation due to drugs, such an effect is only pertinent when the difference in QT/QTc exceeds ± 5 ms of the mean QT/QTc at the 95% confidence interval [1]. Thereby following this intuition, the upper bound of the precision is chosen to be approximately $\lambda_m = 0.04$. This corresponds to the fact that the most accurate estimation of QT/QTc would have an error rate

below ± 5 ms of the ground truth. The *prior* distribution of the precision, Gamma$(\lambda \mid k_\lambda, \vartheta_\lambda)$, can also be chosen accordingly to ensure that its GEVD fitted maxima to be approximately 0.04.

5.3.2 Learning from Incomplete Data using the BCLA-MAP Model

In the case where there exist missing labels from annotators, only the available annotations should be considered for inferring the ground truth. The expected z_i can be re-written as:

$$z_i = \frac{\sum_{j \in V_i} \lambda^j \left(y_i^j - \phi^j \right) + \left(\mathbf{x}_i^\intercal \mathbf{w} \right) b}{\sum_{j \in V_i} \lambda^j + b} \tag{5.17}$$

The precision of the jth annotator is as follows:

$$\frac{1}{\lambda^j} = \frac{1}{N_j + 2(k_\lambda - 1)} \left[\sum_{i \in U_j} \left(y_i^j - \phi^j - z_i \right)^2 + \frac{2}{\vartheta_\lambda} \right]. \tag{5.18}$$

The bias value for the jth annotator can now be written as:

$$\phi^j = \frac{\sum_{i \in U_j} \left(y_i^j - z_i \right) + \mu_\phi \left(\frac{\alpha_\phi}{\lambda^j} \right)}{N_j + \frac{\alpha_\phi}{\lambda^j}}. \tag{5.19}$$

where U_j is the set of records with annotations provided by the jth annotator, and V_i is the set of annotators that provided annotations for the ith record. N_j is the number of records annotated by the jth annotator.

5.4 The scalar Simultaneous Truth and Performance Level Estimation (sSTAPLE)

The sSTAPLE model was proposed by Warfield *et al.* [3] to resolve issues in fusing multiple (potentially inaccurate) segmentations from multiple annotators to result in a consensus, for problems in the domain of medical imaging. sSTAPLE estimates the quality of the annotator-generated segmentation by calculating the bias and variance of the distance between the annotator's segmentation boundary and a reference boundary (where a boundary refers to the separation between regions corresponding to different tissue types). An EM algorithm was used to infer the reference boundary while estimating the bias and variance of each annotator in an iterative process. A good-quality annotator will have a smaller average distance from the true boundary (i.e., low bias) and a high precision (i.e., small variance of distance from the true boundary). sSTAPLE is implemented in this thesis to serve as a benchmarking algorithm for measuring the biases of annotators.

5.4.1 Theory of sSTAPLE

Suppose that there are N voxels in an image, and that a segmentation was generated by the jth annotator, where $j = 1,...R$ as before. Let $\mathbf{D} = \left[y_i^{j=1}, \cdots, y_i^{j=R}\right]_{i=1}^{N}$, y_i^j corresponds to the score provided by the jth annotator for the ith voxel, and z_i represents the unknown underlying true score. The graphical representation of sSTAPLE is shown in Fig. 5.6.

In this model, it is assumed that y_i^j was a noisy version of z_i, i.e., $y_i^j = z_i + \phi^j + \varepsilon_i^j$, where the bias of jth annotator can be modelled as ϕ^j, the error is assumed to have an uncorrelated normal distribution, i.e., $\varepsilon_i^j \sim N\left(0, \sigma^{j2}\right)$. Re-arranging the equation, we have $\mathcal{N}\left(y_i^j \mid z_i + \phi^j, \left(\sigma^j\right)^2\right)$, where $\left(\sigma^j\right)^2$ is the inverse of the precision λ^j as before. Here σ^j is the standard deviation in annotation of the jth annotator and represents his variance in annotation around z_i. sSTAPLE assumes that each annotator independently assigns scores,

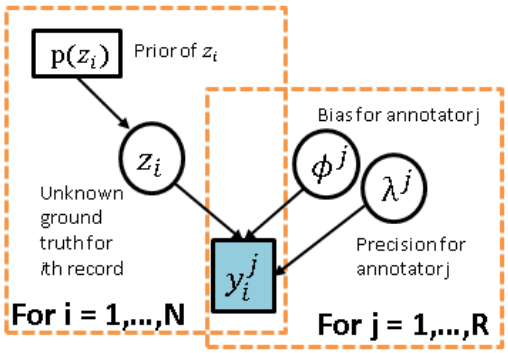


Fig. 5.6 Graphical representation of the sSTAPLE model: y_i^j corresponds to the score provided by the jth annotator for the ith voxel, and it is modelled by the z_i (the unknown underlying true score), the ϕ^j (bias), and the λ^j (precision). Furthermore, z_i has a prior, $p(z_i)$ which has a multivariate uniform distribution as proposed in [3].

and that the error in assigning a score, after accounting for annotator-specific bias and annotator-specific variance, has a standard normal distribution. Furthermore, sSTAPLE assumes that there is no spatial correlation among true scores between voxels. Therefore, the joint distribution of the scores of all the voxels conditional upon the true scores and annotators' parameters is:

$$p\left(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\phi}, \boldsymbol{\sigma}\right) = \prod_{i=1}^{N} \prod_{j=1}^{R} \mathcal{N}\left(y_i^j \mid z_i + \phi^j, \left(\sigma^j\right)^2\right). \tag{5.20}$$

Note that σ^j and ϕ^j are considered to be constants for the jth annotator for all voxels. Since the ground truth, z_i, is unknown, its conditional probability density function is given by the annotators' scores, and by their bias and variance values. The posterior probability density function of $\mathbf{z}$ is defined via Bayes' theorem as follows:

$$p\left(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\phi}, \sigma\right) = \frac{1}{Z} \prod_{i=1}^{N} p(z_i) \prod_{j=1}^{R} \mathcal{N}\left(y_i^j \mid z_i + \phi^j, \left(\sigma^j\right)^2\right), \tag{5.21}$$

where Z is the marginal likelihood (i.e., evidence), and $p(\mathbf{z})$ is the prior of $\mathbf{z}$ drawn from a multivariate uniform distribution. For each ith voxel, the conditional probability density

function of z_i becomes:

$$p\left(z_i \mid y_i^j, \phi^j, \sigma^j\right) = \frac{1}{Z_i} p(z_i) \prod_{j=1}^{R} \frac{1}{\sigma^j \sqrt{2\pi}} - \exp\left\{ \frac{\left(y_i^j - \phi^j - z_i\right)^2}{2\left(\sigma^j\right)^2} \right\}. \tag{5.22}$$

Let $\mu_{z_i}^{\ j} = y_i^j - \phi^j$ be the posterior of the true score, which is a product of normal distributions with mean μ_{zi} and variance σ_z. Assuming that $p(z_i)$ is uniform, the variance and mean of the truth score at ith voxel can be defined by completing the square:

$$\frac{1}{{\sigma_z}^2} = \sum_{j=1}^{R} \frac{1}{\left(\sigma^j\right)^2}. \tag{5.23}$$

$$\mu_{zi} = \frac{\sum_{j=1}^{R} \frac{y_i^j - \phi^j}{\left(\sigma^j\right)^2}}{\frac{1}{{\sigma_z}^2}}. \tag{5.24}$$

The estimation of the likelihood of the parameter $\boldsymbol{\theta} = \{\boldsymbol{\sigma}, \boldsymbol{\phi}\}$ for the complete dataset $\{\mathbf{D}, \mathbf{z}\}$ solved using the expectation of the complete data log-likelihood as:

$$\begin{aligned} &\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \mathbb{E}\left[\log p\left(\boldsymbol{\theta} \mid \mathbf{D}, \mathbf{z}\right)\right] \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{j=1}^{R} -\log \sigma^j - \frac{1}{2\left(\sigma^j\right)^2} \left[\left(y_i^j - \phi^j\right)^2 - 2\left(y_i^j - \phi^j\right) \mu_{zi} + {\sigma_z}^2 + \left(\mu_{zi}\right)^2 \right]. \end{aligned} \tag{5.25}$$

Parameters $\boldsymbol{\theta}$ can be derived by estimating the gradient of the equation (5.25) respectively:

$$\begin{aligned} \frac{\mathrm{d} \log p\left(\boldsymbol{\theta} \mid \mathbf{D}, \mathbf{z}\right)}{\mathrm{d}\sigma^j} &= -N + \frac{1}{\left(\sigma^j\right)^2} \sum_{i=1}^{N} \left[\left(y_i^j - \phi^j - z_i\right)^2 + {\sigma_z}^2 \right]. \\ \frac{\mathrm{d} \log p\left(\boldsymbol{\theta} \mid \mathbf{D}, \mathbf{z}\right)}{\mathrm{d}\phi^j} &= \sum_{i=1}^{N} y_i^j - \phi^j - z_i. \end{aligned}$$

By equating derivatives to zero, the parameters in $\boldsymbol{\theta}$ can be written as follows:

$$\sigma^{j2} = \frac{1}{N}\sum_{i=1}^{N}\left[\left(y_i^j - \phi^j - z_i\right)^2 + {\sigma_z}^2\right]. \tag{5.26}$$

$$\phi^j = \frac{1}{N}\sum_{i=1}^{N}\left(y_i^j - z_i\right). \tag{5.27}$$

The latent true score z_i for each voxel can be obtained from its posterior of equation (5.21). As the prior of $\mathbf{z}$ is a multivariate uniform distribution in this case, the maximum log-posterior of z_i occurs at its mean (see μ_{z_i} in equation (5.24)). The expectation of the complete data log-likelihood can be solved using the EM algorithm in a two-step iterative process:
(i) The E-step estimates the expected *true* score for all voxels, $\hat{\mathbf{z}}$, as a weighted sum of the provided annotations, and can be estimated using equation (5.21) by equating the gradient to zero:

$$\hat{z}_i = \frac{\sum_{j=1}^{R}\frac{y_i^j - \phi^j}{(\sigma^j)^2}}{\frac{1}{{\sigma_z}^2}}. \tag{5.28}$$

(ii) The M-step is based on the current estimation of $\hat{\mathbf{z}}$ and given the dataset $\mathbf{D}$. The model parameters, $\boldsymbol{\sigma}$ and $\boldsymbol{\phi}$, can be updated using equations (5.26) and (5.27) accordingly in a sequential order until convergence.

5.4.2 Learning from Incomplete Data using sSTAPLE

In the case where there exist missing scores from annotators, let $(\sigma^j)^2$ to be replaced with $1/\lambda^j$. Equation (5.26) with imbalanced dataset can be re-written as:

$$\frac{1}{\lambda^j} = \frac{1}{N_j}\sum_{i\in U_j}\left[\left(y_i^j - \phi^j - z_i\right)^2 + {\sigma_z}^2\right], \tag{5.29}$$

where U_j is the set of voxels with scores provided by the jth annotator. N_j is the number of voxels annotated by the jth annotator. Similarly, the bias of jth annotator can be re-written as:

$$\phi^j = \frac{1}{N_j}\sum_{i\in U_j} y_i^j - z_i. \tag{5.30}$$

The latent true score z_i can be inferred using the available scores provided by annotators as:

$$z_i = \frac{\sum_{j \in V_i} \left(y_i^j - \phi^j \right) \lambda^j}{\sum_{j \in V_i} \lambda^j}, \tag{5.31}$$

where V_i is the set of annotators that provided scores for the ith voxel.

5.5 Data Description and Methodology for Comparison

5.5.1 Data Description

Simulated QT Dataset

To test the reliability of BCLA as a generative model, it was applied to the simulated QT dataset as described in Chapter 5 section 5.2.4: a total of 548 simulated records were generated, each has 20 independent annotators, thus providing a total of 10,960 annotations (see Fig. 5.4 in Chapter 5). Note that the mean (i.e., a) of the underlying ground truth was updated using a linear regression function, $f(\mathbf{w}, \mathbf{x})$, where the coefficients, $\mathbf{w}$, were estimated using equation (5.11), and no features were considered (i.e., $x_i = 1$).

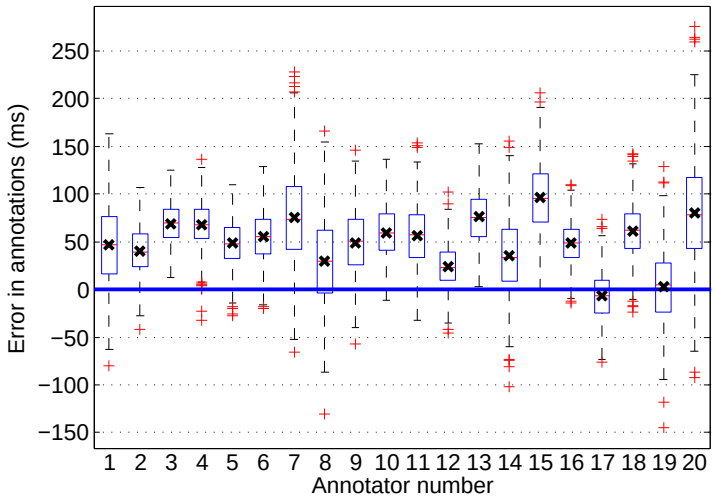
Modelling of the biases The bias term ϕ in the sSTAPLE model is defined as the average of the difference between the given scores and the inferred true scores among all voxels (see equation 5.27). However, it does not model the bias value of each annotator explicitly. In the context of inferring the truth scores using the EM algorithm, Xing *et al.* [130] has argued that sSTAPLE yielded equal likelihood for any bias values, hence cannot be used fully to evaluate annotators' performance. Although the biases of the annotators might be derived from a distribution that is likely to be dataset-dependent, in the absence of any knowledge of the underlying distribution of biases, they are assumed to be drawn from a Gaussian distribution. Furthermore, studies have shown that automated algorithms tend to over-estimate the QT interval in electrocardiogram when compared to manual reference as

described previously [159–161]. There are also other studies that demonstrated a consistent under- or over-estimation of respiratory rate from both manual and automated annotators [53–57, 59]. Therefore it is not ideal to assume the mean of the biases (i.e., μ_ϕ) to be zero, and by incorporating the mean value as an *a priori* input into the framework of BCLA, it allows the model to adapt adequately to specific datasets.

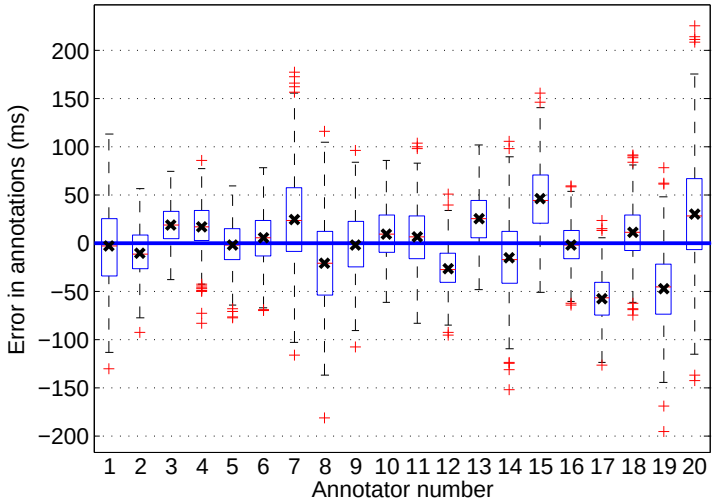
To investigate the effect of modelling the biases, the synthetic dataset of 10,960 annotations was presented to the BCLA-MAP and sSTAPLE models to evaluate their accuracies in estimating the *true* annotation in an unsupervised manner. We investigate their use of predicting the bias and precision of each annotator for three scenarios with varying mean-bias values: 1) μ_ϕ = 50 ms; 2) μ_ϕ = 0 ms; 3) μ_ϕ = -10 ms. Boxplots for results obtained for the simulated dataset are shown in Fig. 5.7 for varying $\mu_\phi = 50$ ms, 0 ms, and -10 ms: the span of each box remains the same for various values of μ_ϕ, however, the error between the generated and *true* annotations in each box increases or decreases depending on the value of μ_ϕ.

QT Dataset

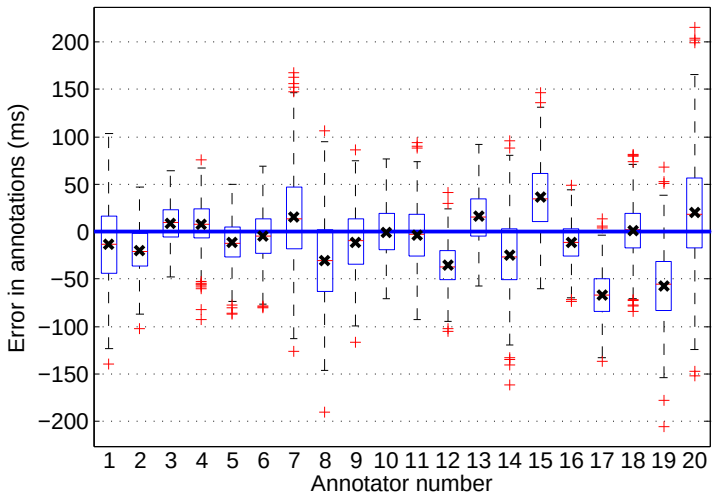
We will also present results of applying our methods to the QT dataset, described in Chapter 3. The set of manual entry (i.e., Division 1) was used to generate the reference annotations, and so we therefore focused on the analysis of the sets of automated labels (i.e., Divisions 2, 3, and 4). In terms of parameter setting (see Table 5.1), annotator-specific precision values, we chose $\boldsymbol{\lambda} \sim \text{Gamma}(k_\lambda, \vartheta_\lambda)$, with the assumption that the annotations provided by the best-performing algorithm are $\pm$5 ms from the reference. Annotators' biases were set via $\boldsymbol{\phi} \sim \mathcal{N}\left(\mu_\phi, \alpha_\phi^{-1/2}\right)$, with $\mu_\phi = 10$ ms and $\alpha_\phi \sim \text{Gamma}(k_\alpha, \vartheta_\alpha)$, assuming that the automated annotations tend to over-estimate manual annotations as described previously. The *true* QT interval for each record is $z_i \sim \mathcal{N}(a, b^{-1/2})$, where $b \sim \text{Gamma}(k_b, \vartheta_b)$ [162, 163, 12]. Instead of assuming the mean a of the underlying ground truth to be a fixed scalar,



(a) μ_ϕ = 50 ms



(b) μ_ϕ = 0 ms



(c) μ_ϕ = -10 ms

Fig. 5.7 Box plots of the error between the generated and *true* annotations for each of the 20 simulated annotators. The black '×' indicates the bias of each annotator with mean of the biases set to be μ_ϕ. The span of each box represents the interquartile range over all annotations for each annotator.

Table 5.1 The parameters of the BCLA model and their values for modelling the PCinC QT dataset.

Symbol	Definition	Value
k_b	shape of Gamma distribution for b	3 ‡
ϑ_b	scale of Gamma distribution for b	0.0002 ‡
μ_ϕ	mean of the bias distribution	10 †
k_α	shape of Gamma distribution for α_ϕ	3 †
ϑ_α	scale of Gamma distribution for α_ϕ	0.0005 †
k_λ	shape of Gamma distribution for $\boldsymbol{\lambda}$	4*
ϑ_λ	scale of Gamma distribution for $\boldsymbol{\lambda}$	0.003*

Note: b is the precision parameter for the model of the ground truth. α_ϕ is the precision parameter for the model of the bias. $\boldsymbol{\lambda}$ refers to annotators' precision values. The values denoted by * are determined with the assumption that the annotations provided by the best performing algorithm is ± 5 ms away from the reference. The values denoted with † are derived from [159–161]. The values with ‡ are derived from [12, 162, 163].

it was updated using a linear regression function, $f(\mathbf{w}, \mathbf{x})$, where the coefficients, $\mathbf{w}$, were estimated using equation (5.11). An intercept was included in $\mathbf{w}$ for modelling the overall offset predicted in f, and no particular features were considered for this example case (i.e., $x_i = 1$).

5.5.2 Methodology for Comparison

The precision values $\boldsymbol{\lambda}$ inferred by BCLA-MAP were compared with those estimated using the EM algorithm proposed by Raykar *et al.* [89] (denoted as EM-R), which serves as one of our benchmarking algorithms. As EM-R does not explicitly model the bias of each annotator, sSTAPLE serves as the second benchmarking algorithm for comparison. For the QT dataset, the mean and standard deviation ($\mu \pm \sigma_\mu$ ms) of 1,000 bootstrapped samples (i.e., random sampling with replacement) across records from the BCLA-MAP model were compared with the best algorithm (i.e., the "theoretical best" algorithm with the least RMSE which can only be determined with knowledge of the reference labels), the two benchmarks, and the traditional naïve mean and median voting approaches. The same method was applied to the simulated dataset for 100 bootstrapped samples. The mean absolute error (MAE) of

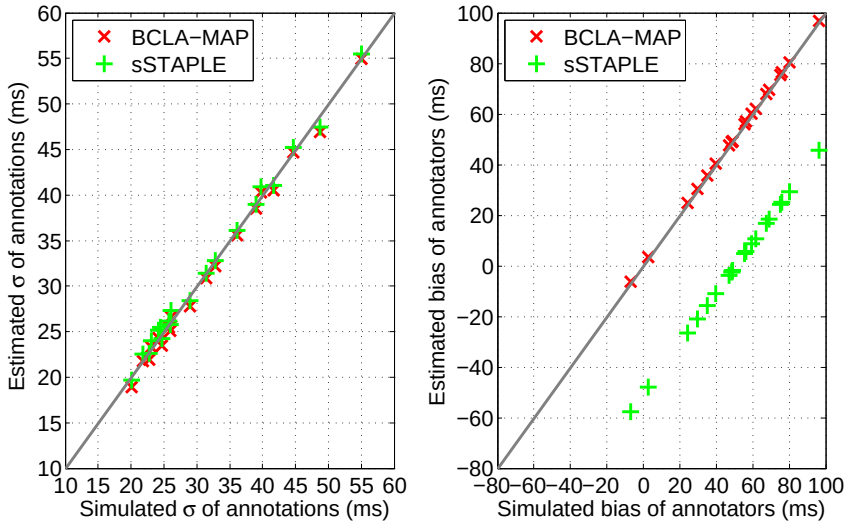
the annotations was calculated for both datasets, which provides a measure of the difference between the estimated and the reference annotations (with a resolution of 1 ms). A two-sided Wilcoxon rank-sum test ($p < 0.0001$) was applied to the bootstrapped RMSEs and MAEs, to provide a comparison between the various methods. In assessing the performance of BCLA-MAP as a function of the number of annotators, a random number of annotators was selected 1,000 times. This was repeated with the number of the annotators varied from three to the maximum number in the division. The minimum number of annotators was chosen to be three to allow for obtaining results from the median voting approach.

5.6 Results

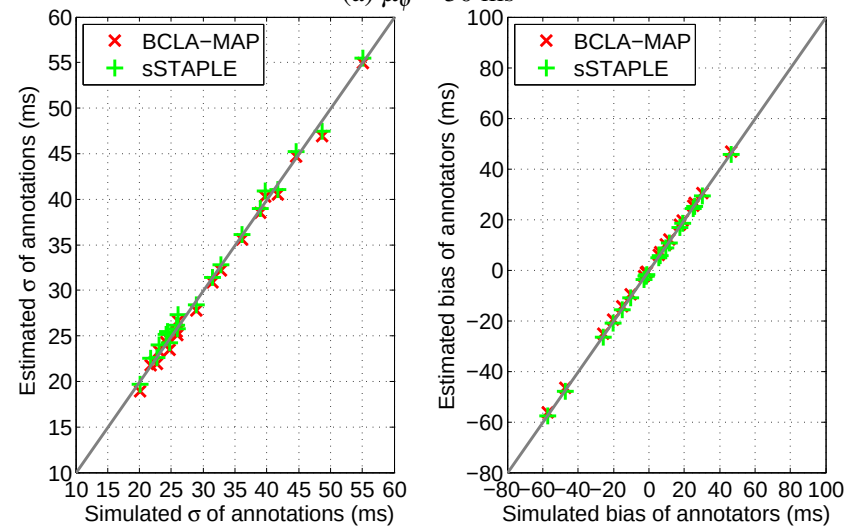
The number of iterations required for BCLA-MAP to converge is dependent on the number of records and the number of annotations. To illustrate the practical utility of the proposed model, it took 7.55 seconds for parameter estimation to perform 5,000 iterations when considering a total of 20,712 annotations (Division 2 in the PCinC QT dataset) using MATLAB R2011a on a 3.3GHz Intel Xeon CPU. Approximately 2,500 iterations were required to stabilise the values of all the parameters. In comparison, both EM-R and sSTAPLE took similar amounts of time to perform 5,000 iterations on the same processor using the same amount of annotations.

5.6.1 Modelling of the Biases

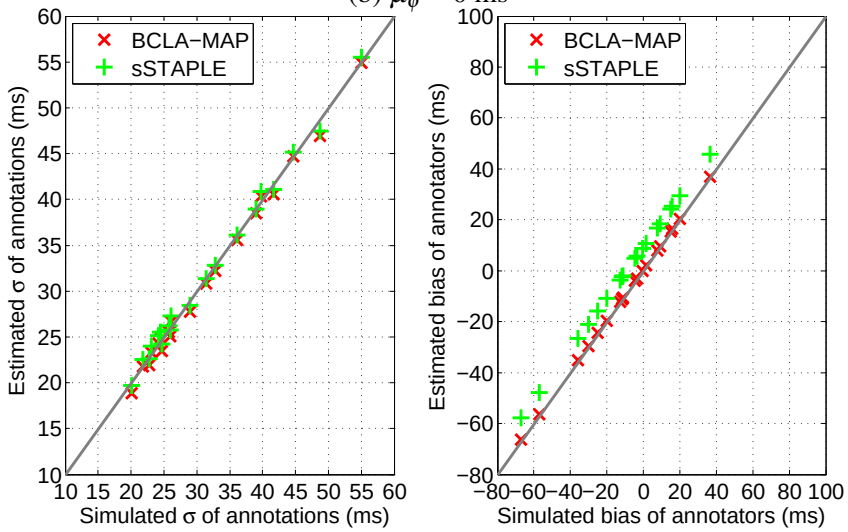
The estimated bias and precision values for each scenario are plotted against the simulated true values as shown in Fig. 5.8. The results show that sSTAPLE significantly over/under-estimates the bias values when their mean, μ_ϕ, was non-zero in the simulations. This phenomenon can be explained by considering that the EM algorithm is guaranteed to converge to a local optimum by maximising the likelihood, and any bias values in the sSTAPLE model could indicate such an optimum. By introducing a prior on the biases as described in the



(a) μ_ϕ = 50 ms



(b) μ_ϕ = 0 ms



(c) μ_ϕ = -10 ms

Fig. 5.8 A comparison of the simulated and inferred σ in (a) and bias in (b) of each annotator in the simulated dataset for different mean bias values. The precision λ can be found via $\sigma^{-1/2}$. The diagonal line indicates a perfect match between simulated and estimated results. Note that sSTAPLE significantly over/under-estimates the bias values when their mean μ_ϕ was non-zero in the simulations.

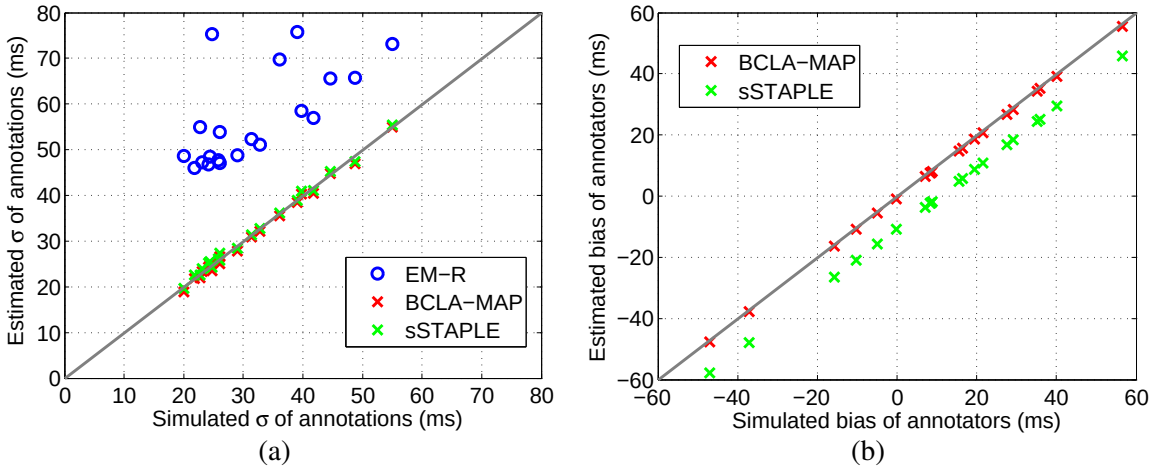


Fig. 5.9 A comparison of the simulated and inferred σ in (a) and bias in (b) of each annotator in the simulated dataset. The diagonal line indicates a perfect match between simulated and estimated results. Note that EM-R significantly over-estimates the σ values and sSTAPLE significantly under-estimates the bias values in all simulations.

BCLA-MAP model, the most probable μ_ϕ can be assigned to improve on the estimation of the inferred ground truth.

5.6.2 Simulated Dataset

Fig. 5.9 (a) shows an example of the inferred results estimated using EM-R, sSTAPLE, and BCLA-MAP. As the EM-R algorithm modelled jointly the precision of each annotator and the noise of the underlying ground truth, its estimated σ cannot represent the real precision of each annotator. Furthermore, the EM-R algorithm does not consider the bias of each annotator, and it is observed that its estimated values of σ were well above the line of identity, indicating a consistent over-estimation. By way of contrast, the BCLA-MAP and sSTAPLE models inferred values for σ that lie closely to the line of identity in the plot, indicating that both models can provide a reliable estimation of the *true* precision in the simulated results. In addition to precision, BCLA-MAP modelled the bias of each annotator accurately, which is superior to those estimated using sSTAPLE. The results are shown in Fig. 5.9 (b): the estimated biases from BCLA-MAP are very close to the *true* biases, whereas the sSTAPLE

Table 5.2 RMSEs and MAEs of the inferred labels using different approaches in the simulated dataset.

	Best Annotator	**Median**	**Mean**	**EM-R**	**sSTAPLE**	**BCLA-MAP**
RMSE (ms)	23.79±0.63*	14.84±0.38*	13.11±0.31*	14.21±0.36	12.45±0.32‡	6.27±0.19*†
MAE (ms)	18.99±0.58*	12.60±0.36	11.26±0.30*	12.64±0.36	10.94±0.33‡	4.97±0.16*†

Results significantly different from others ($p < 0.0001$) as shown in † for BCLA-MAP, ‡ for sSTAPLE, and * (columns 2 to 4, and columns 6 to 7 only) for EM-R using the two-tailed Wilcoxon rank-sum test.

under-estimated all the biases values. Although not all the estimated precisions and biases of each annotator were identical to the simulated values, the BCLA-MAP model inferred annotations without any prior knowledge of which annotator was the best and did so in an unsupervised manner.

In order to compare the accuracy of the inferred labels using the BCLA-MAP model, the simulated 548 annotations were bootstrapped 100 times with replacement. Each time, RMSE and MAE were calculated and compared to the best annotator, mean, EM-R, sSTAPLE, and median voting strategies. The results are shown in Table 5.2, which demonstrate that BCLA-MAP significantly outperformed the mean, median, EM-R, sSTAPLE, and best annotator when compared with the simulated *true* annotations.

5.6.3 QT Dataset

Fig. 5.10 (a)-(g) show the inferred precision and bias results estimated using EM-R, sSTAPLE, and BCLA-MAP for different divisions in the PCinC QT dataset. As mentioned previously, the EM-R algorithm does not directly model the precision of each annotator; its estimated σ of each annotator produces an offset from the values provided by the reference annotations. In contrast, BCLA-MAP and sSTAPLE inferred σ results that lie much closer to the line of identity, in Fig. 5.10 (a), (c), and (e); this indicates that the BCLA-MAP and sSTAPLE models can provide a reliable estimation of the *true* precision of each annotator. In terms of the bias estimation (see Fig. 5.10 (b), (d), and (f)), which is considered in the sSTAPLE

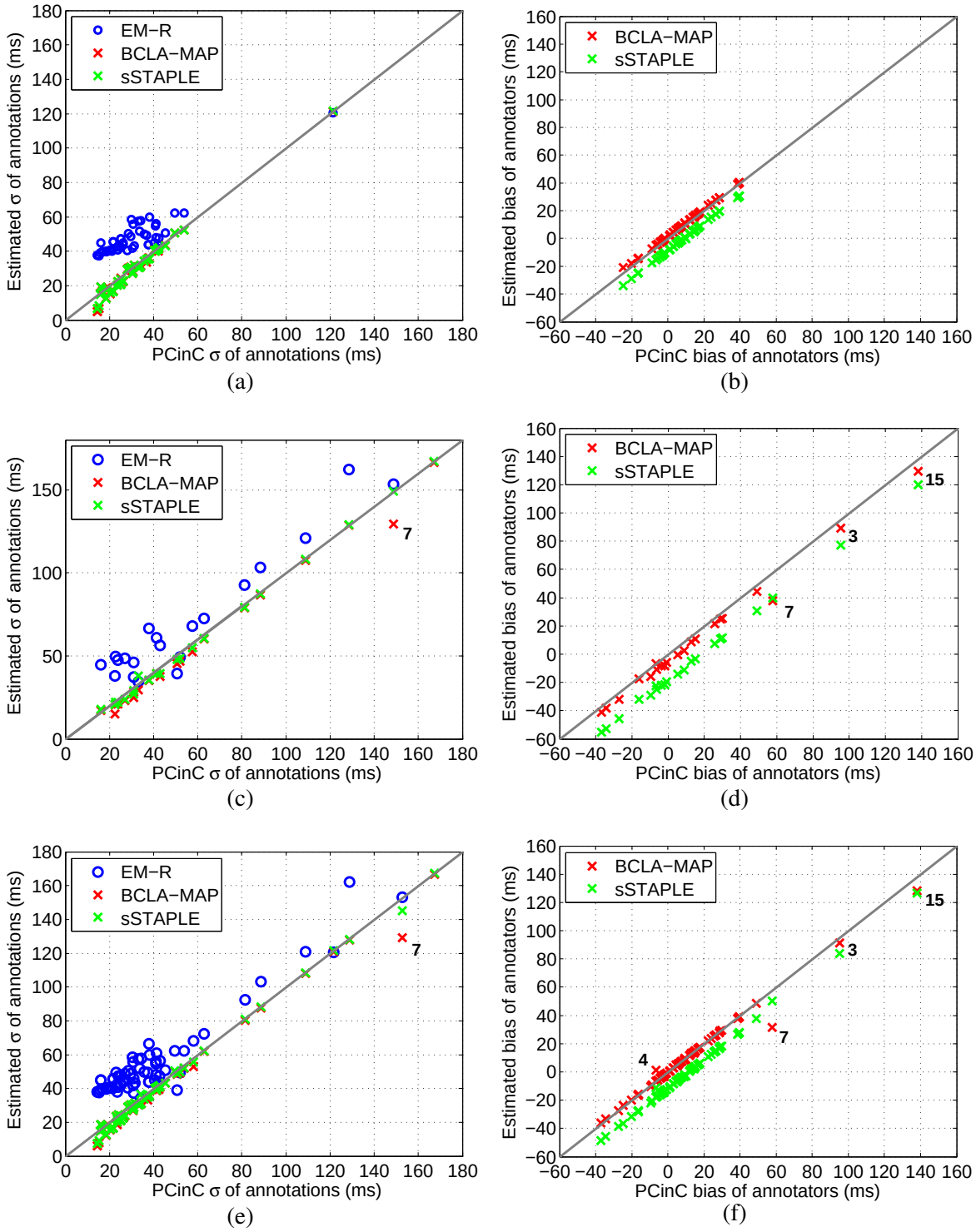


Fig. 5.10 A comparison of the PCinC QT reference and inferred σ and bias of each annotator for Division 2 in (a) and (b), Division 3 in (c) and (d), and Division 4 in (e) and (f) respectively. The leading diagonal line of each plot indicates a perfect matched between the Challenge reference and the estimated results. Note the annotators 3, 4, 7, and 15 are labelled in the corresponding plots.

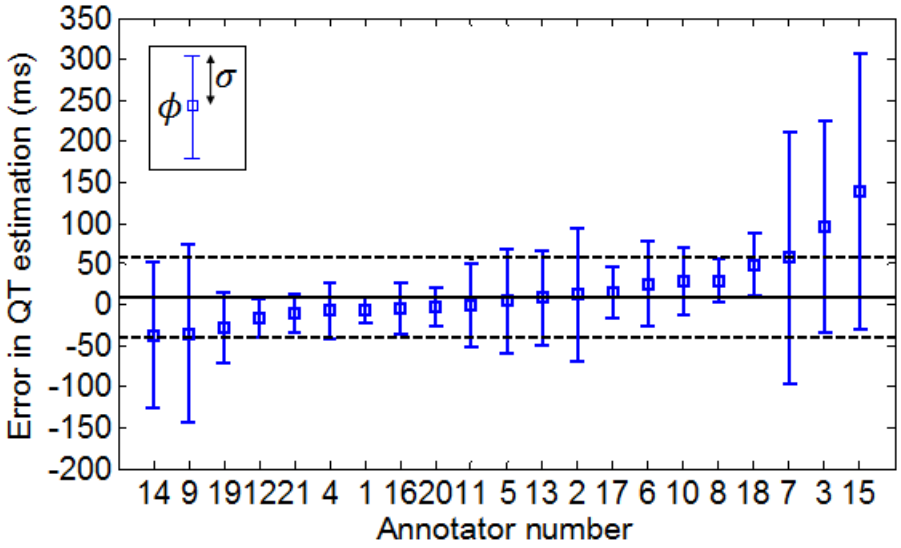


Fig. 5.11 The mean (i.e., bias), ϕ, and σ (i.e., standard deviation) of the difference in annotations for Division 3. The annotators were ranked based on their bias values. The solid line indicates the mean of the biases whereas the dotted lines indicate 1.96σ of the mean assumed in BCLA-MAP.

model, it does not model the mean of the biases, hence consistently produced an underestimation of bias values. In comparison, BCLA-MAP modelled the bias of each annotator accurately (see Fig. 5.10 (b), (d), and (f)). Although automated annotators 3 and 15 were predicted by BCLA-MAP to have lower bias values than those provided by the reference, they are considered to be outliers due to the assumption made in the model: annotators' biases were drawn from a Gaussian distribution with 10 ms mean and 25 ms standard deviation. As Fig. 5.11 shows, the biases of annotators 3 and 15 lie outside the 95% of the area (i.e., $\pm 1.96\sigma$ of the mean under the normal distribution) predicted by BCLA-MAP. In the case of annotator 7, its precision was underestimated (see Fig. 5.10 (c) and (e)), which also affected BCLA-MAP's estimation of its bias value. It was observed that only 3.47% of records were annotated by annotator 7, making it harder for BCLA-MAP to provide a reliable estimation of precision and bias values for that annotator. It is a similar case for annotator 4, where only 2.74% of annotated records were provided, and which likewise affects the BCLA estimation of the bias value for that annotator.

Table 5.3 RMSEs and MAEs of inferred labels using different voting approaches in the PCinC QT dataset.

RMSE (ms)						
Division	**Best Annotator**	**Median**	**Mean**	**EM-R**	**sSTAPLE**	**BCLA-MAP**
2(48)	15.36±0.66*†‡	15.29±0.56*†‡	16.17±0.57*†‡	15.02±0.52†	15.20±0.99†	12.65±0.64*‡
3(21)	16.75±1.81*†‡	19.13±0.83*†‡	30.68±1.46*†‡	18.89±0.83†‡	22.33±1.08*†	14.19±0.87*‡
4(69)	15.12±1.22*†‡	14.44±0.52*†‡	17.66±0.57*†‡	14.75±0.54†‡	16.32±0.61*†	11.89±0.66*‡
MAE (ms)						
Division	**Best Annotator**	**Median**	**Mean**	**EM-R**	**sSTAPLE**	**BCLA-MAP**
2(48)	10.80±0.57*†‡	11.75±0.42†	12.64±0.44*†‡	11.80±0.43†	11.64±0.65†	9.34±0.43*‡
3(21)	10.62±1.14*‡	14.05±0.55†‡	22.99±0.83*†‡	14.10±0.61†‡	19.15±1.78*†	10.60±0.69*‡
4(69)	10.73±0.86*†‡	11.23±0.39*†‡	14.22±0.45*†‡	11.50±0.43†‡	13.23±0.49*†	8.60±0.44*‡

Results significantly different from others ($p < 0.0001$) as shown in † (columns 2 to 6 only) for the BCLA-MAP model, ‡ (columns 2 to 5, and 7 only) for sSTAPLE, and * (columns 2 to 4, and 6 to 7 only) for EM-R using the two-tailed Wilcoxon rank-sum test. Note that the "Best" annotator is defined as the single annotator with the least RMSE.

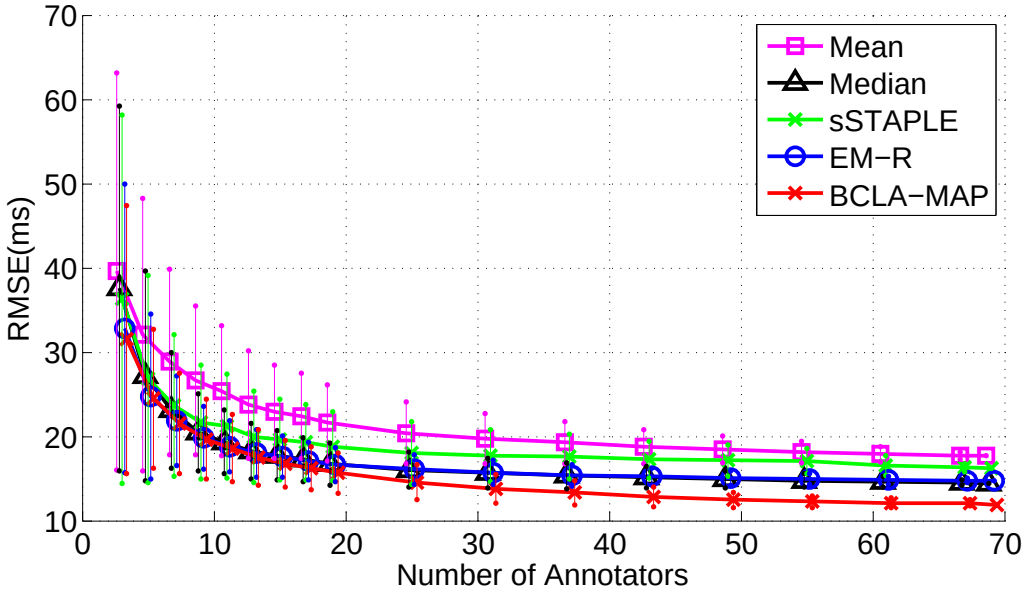
In the evaluation of the inferred labels, the 548 records were bootstrapped 1,000 times, the RMSEs and MAEs of the BCLA-MAP model were generated and compared to the best annotator, mean, EM-R, sSTAPLE, and median voting approaches for the given reference. The results are displayed in Table 5.3: for Division 2 using 48 algorithms, BCLA-MAP achieved a RMSE of 12.65$\pm$0.64 ms, which significantly outperformed other approaches and provides an improvement of 15.78% over the next-best approach (EM-R with RMSE of 15.02$\pm$0.52 ms); in the closed-source entry Division 3 using 21 algorithms, BCLA-MAP again exhibited a superior performance over the other methods with a RMSE of 14.19$\pm$0.87 ms, and a 15.28% improved error rate over the next-best method (RMSE of 16.75$\pm$1.81 ms). When considering all automated entries (Division 4), BCLA-MAP provided an even more accurate performance than with the other two datasets (Divisions 2 and 3), as well as over other methods tested, with a RMSE of 11.89$\pm$0.66 ms. Note that as the PCinC QT dataset contains missing annotations, BCLA-MAP might produce a different error when different recordings are selected, even though the differences should become insignificant as the frequency of bootstrapping increases. Nevertheless, the BCLA-MAP model always outperformed the other voting strategies in our experiments.

A further evaluation of the accuracies in terms of RMSE were made as a function of the number of annotators (see Fig. 5.12). The results were generated by sub-sampling annotators 1,000 times. EM-R, as a benchmarking algorithm, outperformed mean and median approaches initially, but then underperformed when compared to the median approach after 43 algorithms are used. The performance of sSTAPLE was worse, and only outperformed the mean voting approach. The BCLA-MAP model outperformed the other methods being tested with any number of annotators considered. In practice, it is rare to have more than three to five independent algorithms for estimating a label or predicting an event. In the case where only three automated algorithms were randomly selected, BCLA-MAP had on average 3.99%, 13.20%, 16.11%, and 20.41% improvement over the EM-R, sSTAPLE,

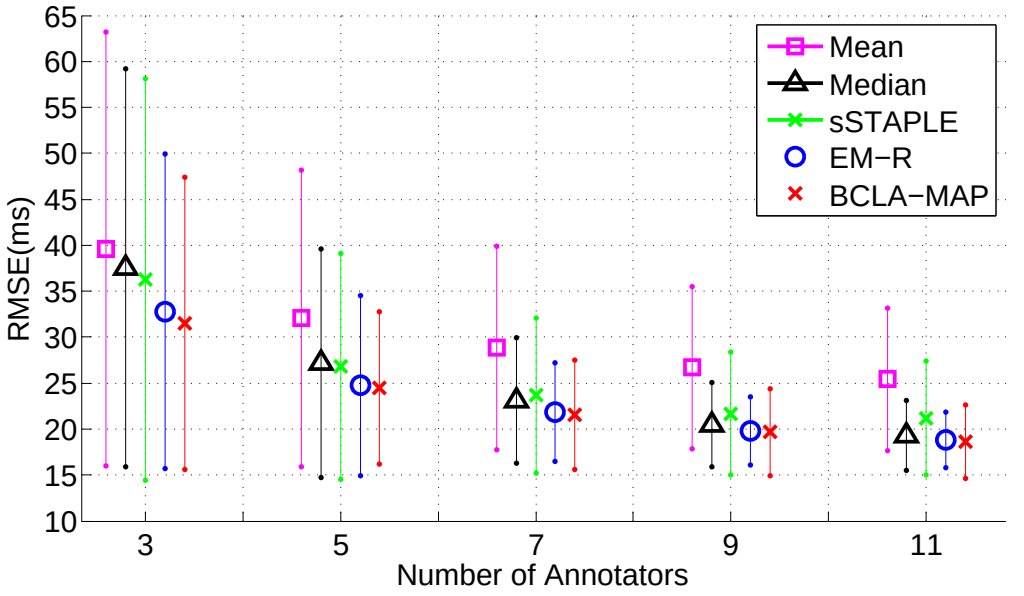
median and mean voting approaches, respectively. A further analysis was conducted to compare the difference in RMSE of the inferred ground truth between BCLA-MAP and the EM-R algorithm for Division 4 (see Fig. 5.13). The results in the figure show that the mean and median of the RMSEs of BCLA-MAP are always smaller than those of EM-R out of the 1,000 runs. The frequency that BCLA-MAP outperformed EM-R (i.e., with smaller RMSE) is shown as a percentage in the lower part of Fig. 5.13 for selecting different numbers of annotators. Although the lowest BCLA-MAP RMSE (11.89$\pm$0.66 ms) in the automated entry is larger than the best-performing human annotator in the Challenge (RMSE = 6.65 ms), there were only two other human annotators who achieved a score below 10 ms. Furthermore, as the annotations of automated algorithms were independently determined from the reference, whereas the reference includes the best human annotators, it is unsurprising that a combination of the automated algorithms would have worse performance. In comparison to the best-performing algorithms selected in the PCinC Challenge (see Table 3.2 in Chapter 3), BCLA-MAP has an improvement of 22.68%, 18.73%, and 27.32% RMSE for Division 2, 3, and 4, respectively.

5.7 Summary

This chapter has proposed a generative Bayesian Continuous-Valued Label Aggregation framework incorporating the ground truth and annotator models to be used in this thesis. Furthermore, a maximum-a-posteriori approach was proposed for the Bayesian Continuous-Valued Label Aggregator (i.e., BCLA-MAP), to infer the ground truth of continuous-valued labels where accurate and consistent expert annotations are not available. As a proof-of-concept, BCLA-MAP was applied to the QT interval estimation from the ECG using labels from the 2006 PhysioNet/Computing in Cardiology Challenge database, and it was compared to the mean, median, EM-R, and sSTAPLE methods. While accurately predicting each labelling algorithms's bias and precision, the root-mean-square error of BCLA-MAP outper-



(a) Random sample of three to 69 annotators.



(b) RMSE results when using 11 annotators or fewer.

Fig. 5.12 Mean and standard deviation of the RMSE results of using different voting approaches are shown as a function of the number of automated annotators. The plot was generated by randomly sampling the annotators 1,000 times.

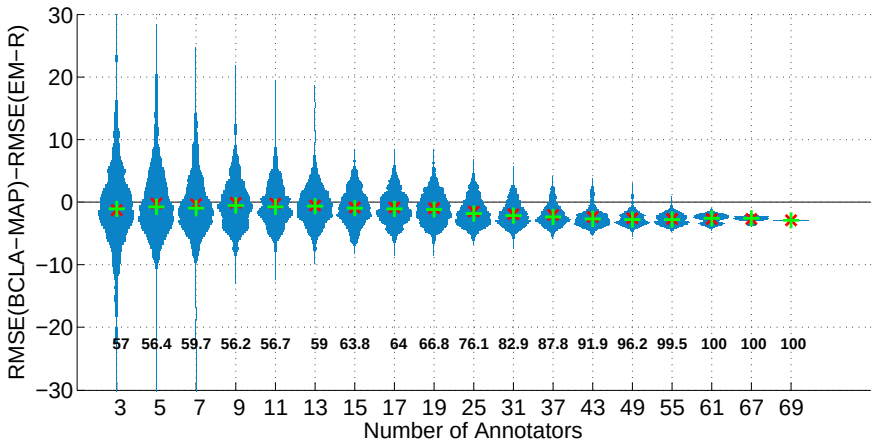


Fig. 5.13 The difference in the RMSE results between BCLA-MAP and EM-R as a function of the number of automated annotators. A probability density estimate of the sub-sampled 1,000 RMSE differences is calculated for selecting number of annotators from three to 69. The mean and the median of the density estimate are labelled as $\times$ in red and $+$ in green respectively. The differences in RMSE that are less than zero were also computed as percentages as shown in the bottom of the plot. They indicate the percentage of times that BCLA-MAP outperformed the EM-R algorithm from a total of 1,000 sub-sampled RMSEs.

formed the best Challenge entry, as well as other voting strategies. BCLA-MAP operates in an unsupervised Bayesian learning framework; no reference data were used to train the model parameters, and separate training and validation test sets were not required. Combining more experienced annotators could therefore be expected to provide a better estimation of the inferred ground truth. Importantly, BCLA-MAP does guarantee a performance better than the best annotator without any prior knowledge of who or what is the best annotator. Finally, combining annotations derived from reliable experts using the BCLA model could potentially lead to improved training for supervised labelling approaches.

Chapter 6 will describe a fully-Bayesian approach to compute parameters in the BCLA model and attempt to further improve the accuracy in estimating the ground truth while jointly modelling the precision and bias values of individual annotators. Although the current chapter only considered the QT dataset as we mainly focused on the demonstration of the BCLA-MAP model, both the Capnobase RR and the QT datasets will be considered in the next chapter to validate aforementioned approaches.

Chapter 6

A Fully-Bayesian BCLA Framework

"Far better an approximate answer to the right question, which is often vague, than the exact answer to the wrong question, which can always be made precise."

– John Tukey

6.1 Introduction

This chapter proposes a fully-Bayesian approach to the BCLA model using Gibbs sampling, denoted as BCLA-Gibbs. We compare BCLA-Gibbs to the performance of all previously considered algorithms. The analysis was applied to the simulated QT dataset, and the PCinC QT dataset. To further challenge the performance of the BCLA framework in a relative small number of recordings as well as limited number of annotators available, the BCLA model was applied to fuse annotations in the Capnobase RR dataset for 42 subjects individually.

6.2 Gibbs Sampling

Markov chain Monte Carlo (MCMC) sampling is the most widely-used method of approximate inference, when direct sampling is difficult [166]. The general concept behind MCMC

is to approximate a posterior distribution by sampling from a large class of distributions following the process of *Monte Carlo* integration over a number of *Markov chains*. Monte Carlo integration refers to numerical integration over the posterior distribution of the model parameters given the data. The *Markov chain* describes the process of sampling a new value from the posterior distribution, given its previous value; this forms a Markov chain of sample draws from the posterior. A simple MCMC algorithm is the Gibbs sampler [167], in which the Markov chain is created by iteratively sampling each parameter of interest, while keeping the most recently sampled values of the other parameters fixed.

6.2.1 Theory

Suppose there is a target distribution over parameters, $\boldsymbol{\theta}$, which can be further partitioned into L subsets of parameters (i.e., $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_L\}$). Given an observed dataset $\mathbf{D}$, the likelihood, $p(\boldsymbol{\theta} \mid \mathbf{D})$, can be solved using the Gibbs sampler only if: there exists an analytic expression for the conditional probability distribution of each parameter in the joint distribution of the likelihood, given all other parameters. The lth parameter can be drawn from its conditional distribution and expressed as:

$$\theta_l \sim p(\theta_l \mid \boldsymbol{\theta}_{-l}, \mathbf{D}), \tag{6.1}$$

where $\boldsymbol{\theta}_{-l}$ denotes all parameters in $\boldsymbol{\theta}$ except l. The conditional probability $p(\theta_l \mid \boldsymbol{\theta}_{-l}, \mathbf{D})$ must therefore be known prior to sampling θ_l, while keeping all other parameters fixed.

The Sampling Process of the Gibbs Sampler

The Gibbs sampling approach can be considered as a stochastic analogue to the expectation-maximisation algorithm, where the expectation and maximisation steps are replaced by random sampling. The random samples of the parameters in the likelihood function $p(\boldsymbol{\theta} \mid \mathbf{D})$ are drawn from the posterior distribution, while each parameter is updated in a sequential manner, conditional on all other parameters at each sampling iteration. In the hth sequence of

sampling, the lth parameter is drawn from its conditional distribution and can be as expressed as:

$$\theta_l^h \sim p\left(\theta_l \mid \boldsymbol{\theta}_{-l}^{(h-1)}, \mathbf{D}\right). \tag{6.2}$$

The implementation of a Gibbs sampler can be defined as follows:

Algorithm 1 The Gibbs Sampler

1: $\boldsymbol{\theta}^0 \leftarrow \mathbf{x}$
2: **for** each sequence $h = 0$ to M **do**
3: set $\hat{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}^h$
4: **for** $l = 1$ to L **do**
5: Draw $X \sim p\left(\theta_l \mid \hat{\boldsymbol{\theta}}_{-l}, \mathbf{D}\right)$
6: Set $\hat{\theta}_l \leftarrow X$
7: **end for**
8: Set $\boldsymbol{\theta}^{(h+1)} \leftarrow \hat{\boldsymbol{\theta}}$
9: **end for**

Note that $\mathbf{x}$ is an initialisation vector, and X is the updated value drawn from the posterior conditional distribution for parameter θ_l.

The Expectation of the Gibbs Sampler

Similar to other MCMC algorithms, the Gibbs sampler is a stochastic process, where a "burn-in period" is defined as discarding undesired samples to (1) avoid correlation among initial successive samples and (2) make sure the random draws of the posterior are closer to the stationary distribution[1] of the Markov chain and less dependent on the starting distribution of the model. The amount of burn-in required depends on the complexity of the model, which affects the convergence rate and subsequently the number of samples required to reach the stationary distribution.

[1] N.B.: it is a distribution that remains invariant with time regardless of the starting state.

After a burn-in period, an expectation of the remaining samples can be calculated to provide a posterior estimation of the parameter, such as:

$$\mathbb{E}[f(x)]_p = \int_{i=m+1}^{i=M} f\left(x^{(i)}\right) p_i dx, \tag{6.3}$$

where $f(x)_p$ is the posterior distribution of the function $f(x)$ of interest, and $f\left(x^{(i)}\right)$ is the ith sample drawn from p, the posterior distribution. M refers to the total number of samples, m is the burn-in period, and p_i refers to the likelihood of the ith sample. Under the ergodic theorem of Markov chains [166], the expectation approximates the sample average:

$$\mathbb{E}[f(x)]_p \simeq \frac{1}{M-m} \sum_{i=m+1}^{M} f\left(x^{(i)}\right). \tag{6.4}$$

6.2.2 Convergence Measures

As supported by its underlying theory, MCMC algorithms guarantee convergence with a relatively large (potentially infinite) number of iterations. However, the number of burn-ins is difficult to determine *a priori*, and depends on the starting point of the process. The following methods are used to provide a general measure of convergence in MCMC:

(i) Autocorrelation can be used to measure the correlation among different sample sequences or draws for the parameter(s) of interest. The lag k autocorrelation, ρ_k, is defined as the correlation between every draw and its kth lag [168] as:

$$\rho_k = \frac{\sum_{i=k+1}^{M} \left[f\left(x^{(i)}\right) - \bar{f}\right] \left[f\left(x^{(i-k)}\right) - \bar{f}\right]}{\sum_{i=1}^{M} \left[f\left(x^{(i)}\right) - \bar{f}\right]^2}, \tag{6.5}$$

where $\bar{f}$ is the mean value of $f(x)$ for M samples. The value of ρ_k lies in the range [-1, 1], where 1 indicates a perfect correlation and -1 indicating perfect inverse correlation. It should be expected that the kth lag autocorrelation becomes smaller (i.e., approaches 0) as k increases. If the autocorrelation is relative high (i.e., close to 1 or -1) with larger value of k,

it implies a high degree of correlation in the sample draws and slow mixing (i.e., the Markov chain is taking a long time to move around the parameter space), hence it will take longer to converge.

(ii) The Geweke diagnostic [169]: this measures convergence by extracting the first 10% and the last 50% of a sample of draws, then compares the means using a Z-test of equality. The Z score is measured by calculating the ratio of the difference between the two means and the asymptotic standard error of their difference, where the variance of each set of the sample draws is determined by the spectral density estimation at zero, and which also takes into account any autocorrelation. The hypothesis is that if two sets of sample draws are not significantly different, they are assumed to be drawn from the same standard normal distribution as the number of the sample draws approaches infinity (i.e., the Markov chain has converged).

(iii) The Raftery-Lewis diagnostic [170, 171]: this evaluates the number of sample draws and and the number of burn-in iterations required to reach the accuracy of the desired posterior quantile q within an accuracy of $\pm r$ with probability P. The minimum sample size, M_{min}, is defined as:

$$M_{min} = \left[\Phi^{-1} \left(\frac{(P+1)}{2} \right) \frac{\sqrt{q(1-q)}}{r} \right]^2, \tag{6.6}$$

where $\Phi^{-1}(\cdot)$ is the inverse of the normal cumulative distribution function. Note that the minimum sample size is estimated with the assumption that there is no correlation between consecutive samples, which might not be the case in practice. Furthermore, an estimate of the dependence factor, I, measures the proportional increase in the number of samples required to reach convergence due to the dependence between the sample draws. A value of I that is larger than five indicates strong autocorrelation among sample draws which may be due to a poor choice of starting point, high posterior correlations, or poor mixing. This diagnostic also provides the thinning interval, K, which is defined as every Kth sample to be considered for approximating the expectation of the samples, to behave as if they are

independent. Thinning reduces the autocorrelation between samples and can be useful for memory-storage requirements.

(iv) The Gelman-Rubin diagnostic [172]: this proposes a method for measuring convergence while running sample draws with multiple starting points (i.e., different seeds of a random generator). The concept is that if all runs with different starting points approximate the same value of the parameter of interest then the Markov chain may be deemed to have converged. This method takes into consideration the mean and variance of a parameter of interest within a run and across different runs. A potential scale reduction factor is calculated to diagnose convergence: it is defined as the square root of the variances of the values for all the runs divided by the average of the variances within the separate runs. Gelman [173] suggests that the scale reduction factor should be less than 1.1 or 1.2 if the run has converged. As the diagnostic is based on means and variances, it is an ideal choice to measure convergence of the sample draws where the posterior is Gaussian-distributed.

The above convergence diagnostics are not an exhaustive list, but rather some which are popular in literature and which are simple to implement as their measures of convergence are quantitative. Cowles and Carlin [174] stated in their comparative review of MCMC convergence diagnostics that there is no method available to guarantee the detection of convergence, and recommend the use of a combination of strategies to monitor convergence.

6.3 BCLA-Gibbs

Our starting point, as before, is that there are N records of physiological data labelled by R annotators. Let $\mathbf{D} = \left[\mathbf{x}_i^\intercal, y_i^{j=1}, \cdots, y_i^{j=R}\right]_{i=1}^{N}$, where $\mathbf{x}_i$ is a column feature vector for the ith record containing d features, y_i^j corresponds to the annotation provided by the jth annotator for the ith record, and z_i represents the unknown underlying ground truth. As described in previous chapters, the y_i^j was assumed to be a noisy version of z_i, denoted as $\mathcal{N}\left(y_i^j \mid z_i + \phi^j, 1/\lambda^j\right)$, where ϕ^j is the bias of the jth annotator (see Chapter 5, section 5.2).

Following the derivation of the Gibbs sampler, each parameter in the BCLA likelihood is assumed to be independent, and can therefore be updated by sampling from its conditional posterior distribution with its hyperparameters (denoted by *) as follows (see Appendix B for derivations of the hyperparameter of the conditionally conjugate prior):

The conditional posterior for z_i as the mean parameter of a normal distribution given the precision is also a normal distribution:

$$z_i \sim \mathcal{N}\left(z_i \mid a_i^*, \frac{1}{b^*}\right), \tag{6.7}$$

where

$$a_i^* = \frac{\left(\mathbf{x}_i^\intercal \mathbf{w}\right) b + \sum_{j=1}^{R}\left[\left(y_i^j - \phi^j\right)\lambda^j\right]}{b + \sum_{j=1}^{R}\lambda^j}, \quad b^* = b + \sum_{j=1}^{R}\lambda^j.$$

Similarly, ϕ^j can be drawn from its conditional posterior distribution as:

$$\phi^j \sim \mathcal{N}\left(\phi^j \mid \mu_\phi^{j*}, \frac{1}{\alpha_\phi^{j*}}\right), \tag{6.8}$$

where

$$\mu_\phi^{j*} = \frac{\mu_\phi \alpha_\phi + \lambda^j \sum_{i=1}^{N}\left(y_i^j - z_i\right)}{\alpha_\phi + N\lambda^j}, \quad \alpha_\phi^{j*} = \alpha_\phi + N\lambda^j.$$

The conditional posterior distribution of the precision of the jth annotator, λ^j, with a given mean can be described as:

$$\lambda^j \sim \text{Gamma}\left(\lambda^j \mid k_\lambda^{j*}, \vartheta_\lambda^{j*}\right), \tag{6.9}$$

where

$$k_{\lambda}^{j*} = k_{\lambda} + \frac{N}{2}, \quad \frac{1}{\vartheta_{\lambda}^{j*}} = \frac{\sum_{i=1}^{N}\left(y_i^j - \phi^j - z_i\right)^2}{2} + \frac{1}{\vartheta_{\lambda}}.$$

The b and α_{ϕ} acts as a precision hyperparameter of z_i and ϕ^j respectively can also be drawn from a Gamma distribution as:

$$b \sim \text{Gamma}\left(b \mid k_b^*, \vartheta_b^*\right), \tag{6.10}$$

where

$$k_b^* = k_b + \frac{N}{2}, \quad \frac{1}{\vartheta_b^*} = \frac{\sum_{i=1}^{N}\left(z_i - \bar{z}\right)^2}{2} + \frac{1}{\vartheta_b}.$$

$$\alpha_{\phi} \sim \text{Gamma}\left(\alpha_{\phi} \mid k_{\alpha}^*, \vartheta_{\alpha}^*\right), \tag{6.11}$$

where

$$k_{\alpha}^* = k_{\alpha} + \frac{R}{2}, \quad \frac{1}{\vartheta_{\alpha}^*} = \frac{\sum_{j=1}^{R}\left(\phi^j - \bar{\phi}\right)^2}{2} + \frac{1}{\vartheta_{\alpha}}.$$

Note that N and R are the number of annotations and annotators respectively. Furthermore, $\bar{z}$ and $\bar{\phi}$ corresponds to the averaged value of $\mathbf{z}$ and $\boldsymbol{\phi}$ individually. After M sequences of the Gibbs sampler, the parameters will converge and their expectation values can be approximated after the burn-in period (e.g., $\geq \frac{M}{2}$).

6.3.1 Learning from Incomplete Data using BCLA-Gibbs

As described previously, N and R correspond to the maximum number of annotations and annotators respectively in the BCLA framework. In the case where there are missing

annotations from different annotators (i.e., not all annotators have labelled all recordings), the hyperparameters in the posterior distribution of BCLA-Gibbs can be re-written as follows:

$$a_i^* = \frac{\left(\mathbf{x}_i^\intercal \mathbf{w}\right) b + \sum_{j \in V_i} \left[\left(y_i^j - \phi^j\right) \lambda^j\right]}{b + \sum_{j \in V_i} \lambda^j}, \quad b_i^* = b + \sum_{j \in V_i} \lambda^j.$$

$$\mu_\phi^{j*} = \frac{\mu_\phi \alpha_\phi + \lambda^j \sum_{i \in U_j} \left(y_i^j - z_i\right)}{\alpha_\phi + \sum_{i \in U_j} \lambda^j}, \quad \alpha_\phi^{j*} = \alpha_\phi + \sum_{i \in U_j} \lambda^j.$$

$$k_\lambda^{j*} = k_\lambda + \frac{N_j}{2}, \quad \frac{1}{\vartheta_\lambda^{j*}} = \frac{\sum_{i \in U_j} \left(y_i^j - \phi^j - z_i\right)^2}{2} + \frac{1}{\vartheta_\lambda}.$$

$$k_\alpha^* = k_\alpha + \frac{R}{2}, \quad \frac{1}{\vartheta_\alpha^*} = \frac{\sum_{j=1}^R \left(\phi^j - \bar{\phi}\right)^2}{2} + \frac{1}{\vartheta_\alpha}.$$

$$k_b^* = k_b + \frac{N}{2}, \quad \frac{1}{\vartheta_b^*} = \frac{\sum_{i=1}^N \left(z_i - \bar{z}\right)^2}{2} + \frac{1}{\vartheta_b}.$$

Note that U_j is the set of records/samples with annotations provided by the jth annotator/rater, and V_i is the set of annotators that provided annotations for the ith record. N_j is the number of records annotated by the jth annotator. Furthermore, $\bar{z}$ and $\bar{\phi}$ are the averaged value of $\mathbf{z}$ and $\boldsymbol{\phi}$ respectively.

6.4 Data Description and Methodology of Comparison

6.4.1 Data Description

Simulated QT Dataset

The simulated QT dataset is detailed in Chapter 5 section 5.2.4: a total of 548 simulated records were generated, each has 20 independent annotators, thus providing a total of 10,960 annotations (see Fig. 5.4 in Chapter 5). Note that no features were considered (i.e., $x_i = 1$) in BCLA-Gibbs.

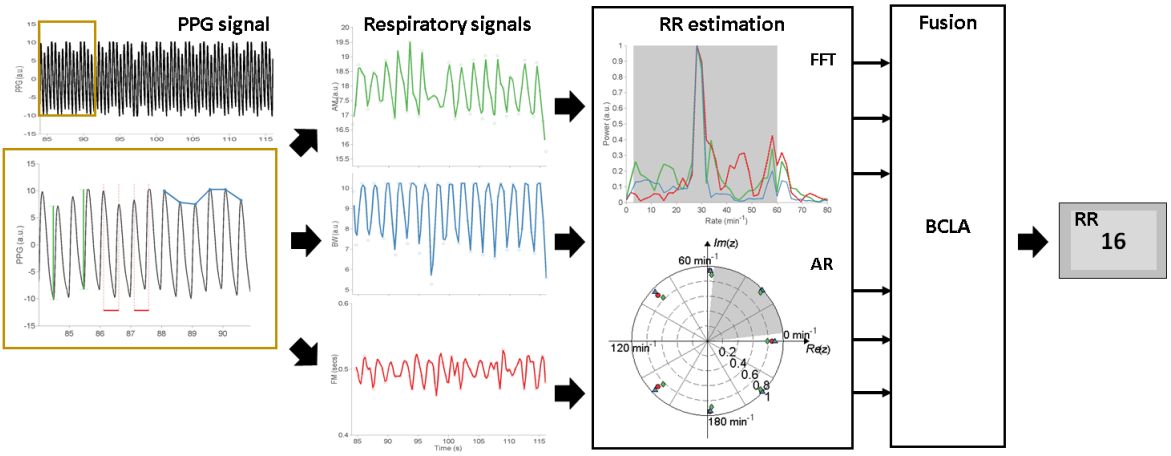


Fig. 6.1 An example of a 32-sec PPG window used for RR estimation and fusion using BCLA: AM (green), BW (blue), and FM (red) respiratory modulations are extracted; for the FFT-based method, the power spectrum is calculated for each modulation using FFT, and the maximum power is selected within the physiologically-plausible RR range (grey area); for the AR-based method, the poles for each modulation are determined using an AR model, and the dominant pole within the plausible range of RR (grey area) is selected. The proposed BCLA framework is then used to combine estimates, and provide a final "fused" value for that window.

PCinC QT Dataset

The annotations were extracted from the PCinC dataset for labelling QT intervals, and the description of the dataset can be found in Chapter 3 section 3.2.

Capnobase RR Dataset

Six respiratory rate (RR) estimates derived from three modulation signals of the PPG (i.e., amplitude, frequency, and intensity modulations), each using both the FFT- and AR-based modelling approaches, were considered as the annotations per window for an individual subject. Each subject had approximately 900 RR estimates over a course of eight minutes with RR being calculated over a 32-second window overlapped by 29 seconds (see Fig. 6.1). There were a total of 42 subjects (age ranged from 0.8 to 69) selected from the Capnobase dataset [144], which is detailed in Chapter 3, section 3.3.

6.4.2 Methodology of Comparison

Simulated and PCinC QT Datasets

The BCLA-Gibbs model was evaluated in the same manner as the BCLA-MAP model, which was described in Chapter 5.

Capnobase RR Dataset

The mean of the MAE and the standard error of the inferred RR estimates from the PPG across all subjects using the proposed fusion framework, BCLA, were compared to the reference RR values from capnography. The BCLA model was applied to the RR estimates extracted using the conventional FFT-based, AR-based, and all (FFT+AR) algorithms described in Chapter 3. Additionally, the performance of BCLA-MAP and BCLA-Gibbs were compared to that of "Smart Fusion" (i.e., the benchmarking algorithm in this field proposed by Karlen *et al.* [81]), and with all other approaches considered previously. To further analyse the performance of different voting algorithms with varied signal qualities, a comparison was made in which we compute the MAE results of (i) RR estimates for all windows; (ii) only RR estimates for those windows following the criteria considered by "Smart Fusion" (denoted *)[2]; (iii) only RR estimates for noisy windows (i.e., those which are rejected by "Smart Fusion") were considered.

6.5 Results

6.5.1 Convergence of the BCLA-Gibbs Model

Fig. 6.2 (a) shows an example of the posterior samples of the inferred QT interval of one recording from the PCinC QT dataset using the BCLA-Gibbs approach, and a histogram

[2]i.e., those windows where RR estimates with one standard deviation exceeding four bpm were discarded.

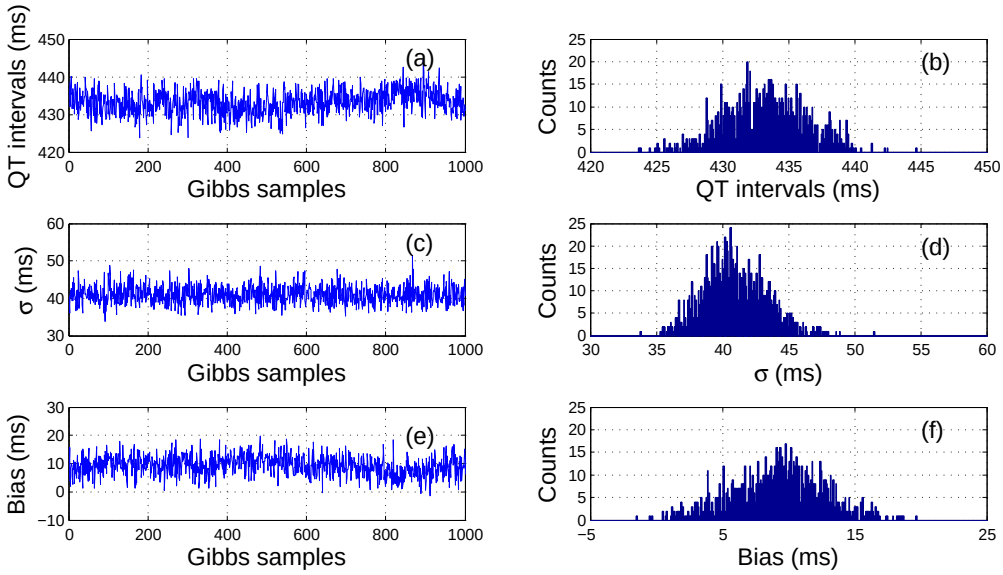


Fig. 6.2 Plot of the Gibbs samples of the parameters using the BCLA-Gibbs model: the inferred 1,000 samples (excluding the burn-in period) of the QT interval of one recording from the PCinC QT dataset and its histogram plot are shown in (a) and (b). The inferred samples of the σ of an annotaor and its histogram plot are shown in (c) and (d). The corresponding results of bias for the same annotator are shown in (e) and (f).

plot of these samples is shown in (b). Similar plots were produced for the inferred standard deviation in annotation (i.e., σ) and bias values of one annotator in (c)-(f). To measure the convergence of the sample draws from the BCLA-Gibbs model, various approaches were considered as discussed in section 6.2.2 and applied to the individual parameter of interest. The autocorrelation was estimated directly using the *autocorr* function provided in Matlab version 2011b. An example of the autocorrelation is provided in Fig. 6.3 (a). The correlogram of a lag 50 autocorrelation of each parameter indicates good mixing as samples with low correlation values. However, not all recordings exhibited similar correlation values, and there is no reliable metric available to define the ideal lag number to determine the number of sample draws required for convergence. Alternatively, other more sophisticated approaches such as the Geweke and Raftery-Lewis diagnostics are available in the Matlab Econometrics Toolbox [175]. The Raftery-Lewis diagnostic was used to define the conservative number of sample draws required to reach convergence. The default setting was selecting a posterior

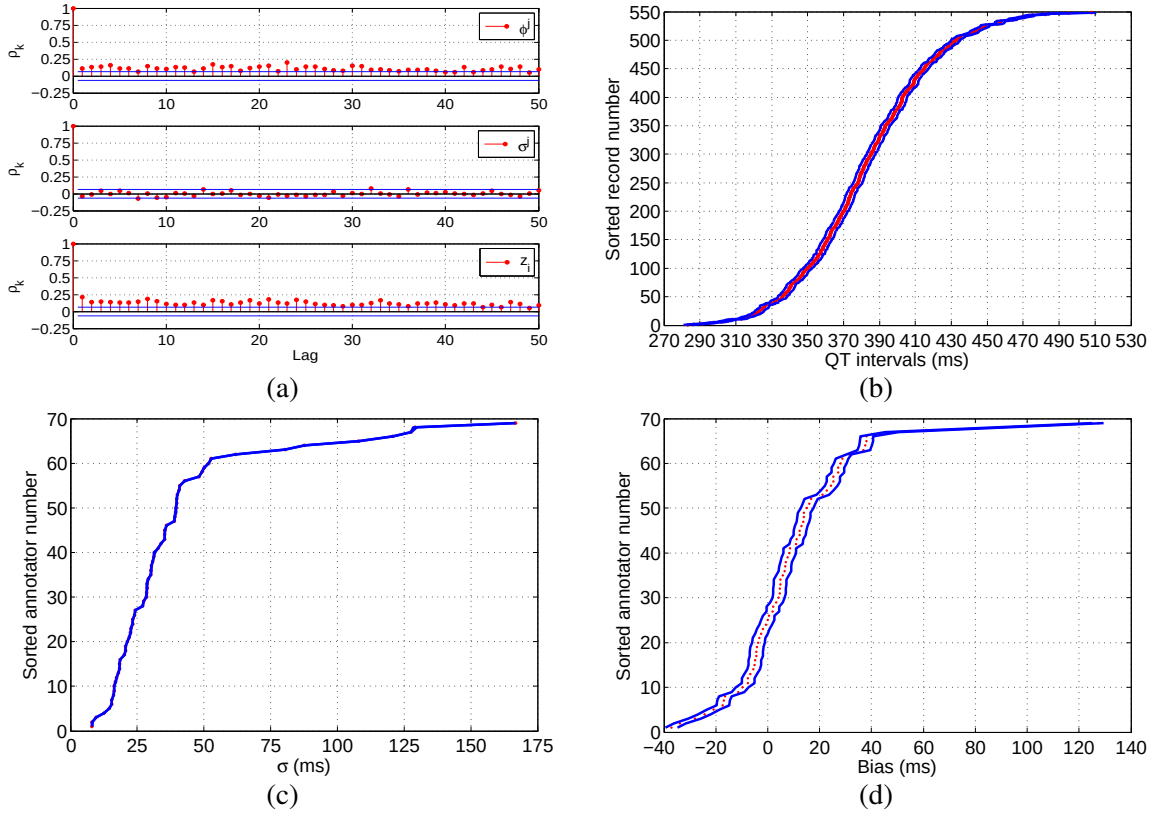


Fig. 6.3 (a) Correlogram of a lag 50 autocorrelation (i.e., ρ_k where $k = 50$) of 1,000 sample draws from the BCLA-Gibbs model using annotations from Division 4 in the PCinC QT dataset. The parameters of interest are the inferred ground truth for one recording (denoted as z_i for the ith record), the estimated standard deviation in annotations (denoted as σ^j), and the estimated bias (denoted as ϕ^j) of the jth annotator. Note the upper and lower bounds of each autocorrelation are plotted in blues with 95% confidence intervals. A comparison when using eight different seeds of a random generator is shown in (b)-(d) for the inferred ground truth of all recordings, the estimated σ and bias values of all annotators, respectively. Each red dot indicates the mean value across eight runs for each record/annotator, and its margin of error is plotted in blue lines for a 95% confidence interval .

quantile of 0.025 that has an accuracy of 0.01 with 95% probability. Note that such a method is highly sensitive to the quantile setting, and hence it was dataset-dependent. The Geweke diagnostic provides another measure of convergence by comparing the means of the first 10% and last 50% of the sample draws. The chi-squared test for equality of these means was also considered at a 5% significance level. With this diagnostic, it was difficult to choose the appropriate window of sample draws for comparison when burn-in was undefined.

It is also known that different seeds of a random generator will produce different sampling results in MCMC if the convergence is not reached. Fig. 6.3 (b)-(d) shows an example of the variation of the expected inferred ground truth across different recordings, as well as the expected σ and bias of each annotator for eight runs of over-dispersed seeds. The Gelman-Rubin diagnostic provided in the Computational Statistics Toolbox [176] (see *csgelrub* function) was applied to the Gibbs samples after the burn-in (i.e., half of the samples were discarded), and across different runs to test for convergence: the potential scale reduction factor was estimated to be 0.9413 ± 0.0048, 0.9268 ± 0.0252, and 0.9995 ± 0.0070 for the inferred ground truth of all recordings, bias and σ values of all annotators respectively, hence the sample draws were converged as their factors were less than the 1.1 suggested by the authors of that method. It was also observed that σ has negligible margin of error with a range of 0.05 ± 0.08 ms across annotators, whereas the bias and the inferred ground truth had a margin of error being 2.41 ± 0.19 ms across annotators and 2.46 ± 0.03 ms across records respectively. Note that the acceptable range of error of the QT interval (i.e., within ± 5 ms [1]) is much larger in clinical practice.

Although the different diagnostic tools mentioned above can be used to measure the convergence of the BCLA-Gibbs model, sometimes they disagree on whether the samples of an individual parameter has converged, particular when the number of burn-in iterations is difficult to define *a priori*. Note that none of the diagnostics is perfect, not even in the prediction of the amount of burn-in required [177, 178]. Furthermore, there exists a trade-off

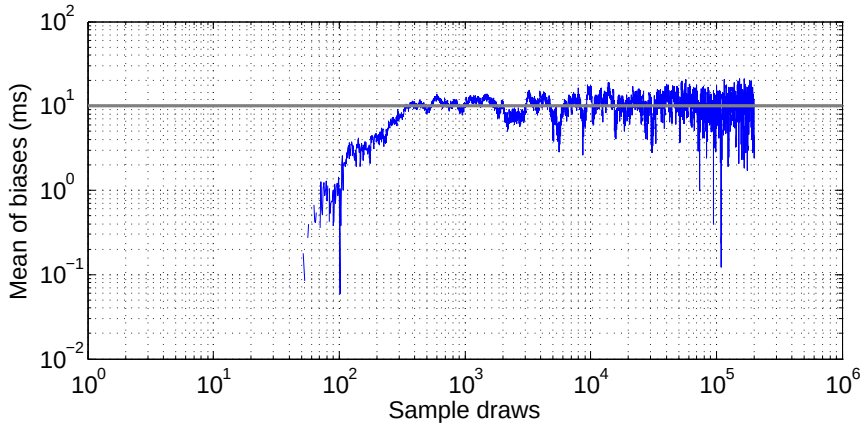


Fig. 6.4 Example of the sample draws of the posterior mean of the biases μ_ϕ^* in the BCLA-Gibbs model when considering annotations provided in Division 4 of the PCinC QT dataset. The results are plotted in logarithmic scale and the user defined mean of the biases (i.e., $\mu_\phi = 10$ ms) is indicated as the grey solid line.

between the amount of samples required to reach convergence and the time it takes to be applicable for a real-time implementation. Hence the methodology adopted in the thesis was a trial-and-error approach with different initialisation values and where convergence was evaluated by quantitative verification when possible. The steps considered are as follows:

(i) Taking relatively short sample draws of each parameter with multiple chains from over-dispersed seeds of a random generator, and the sample draws (with half of sample draws discarded as burn-in) are tested with the Gelman-Rubin diagnostic to ensure the potential scale reduction factors across records/annotators were approximately 1 but less than 1.1;

(ii) The mean of the bias value across annotators (i.e., μ_ϕ in Fig. 5.3) per posterior sample serves as a quantitative verification of convergence, assuming that samples should converge to the precision weighted mean plus the prior mean of the model. An example for μ_ϕ is shown in Fig. 6.4, for Division 4 of the PCinC QT dataset. The posterior sample draws were initialised at random, and gradually approached the target value (i.e., $\mu_\phi = 10$ ms). The "noise" around this value includes the additional bias weighted

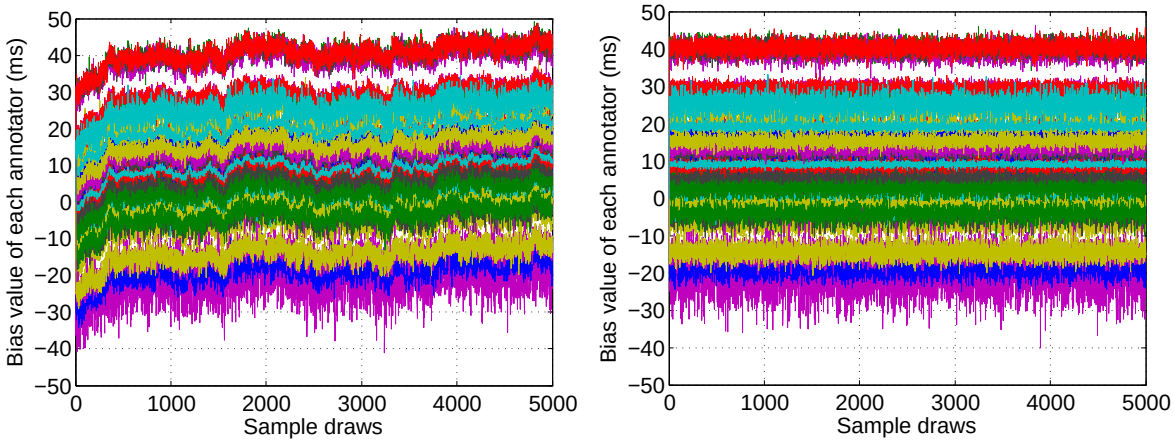


Fig. 6.5 Sample draws of the biases in the BCLA-Gibbs model from the PCinC QT Division 2 dataset. (a) assumes the mean of the biases μ_ϕ^* was allowed to change depending on the inferred values of the ground truth while (b) assumes μ_ϕ^* is fixed to be 10 ms.

by the precision value of each annotator (see equation (6.8)). The oscillation around the mean is caused by a strong coupling effect between the inferred ground truth and the bias values (i.e., they vary simultaneously) due to the many degrees of freedom in the model. Based on the assumption that the annotations are drawn from a Gaussian distribution in the BCLA model, the summation of the inferred ground truth and biases results in a convergent mean. To further investigate the coupling effect, the posterior mean of biases μ_ϕ^* in the BCLA model was fixed to a constant, and the estimated bias value of each annotator ϕ^j was evaluated. The resulting plot is shown in Fig. 6.5, wherein (a) shows the draws of the biases from the PCinC QT Division 2 dataset, and the mean of the biases was allowed to change depending on the inferred values of the ground truth while (b) assumes the mean is fixed. The results demonstrate that the bias values were strongly coupled with the ground truth if they are allowed to update freely in each Gibbs sequence. Nevertheless, when the posterior mean of the biases is unknown, it is not possible to speculate its value *a priori*, and the mean of the biases μ_ϕ is learned from the data.

Table 6.1 RMSEs and MAEs of the inferred labels using different strategies in the simulated dataset.

	Best Annotator	**Median**	**Mean**	**EM-R**	**sSTAPLE**	**BCLA-MAP**	**BCLA-Gibbs**
RMSE(ms)	23.79±0.63†* •	14.84±0.38†*•	13.11±0.31†*•	14.21±0.36†•	12.45±0.32†*‡•	6.27±0.19*	6.29±0.19*
MAE(ms)	18.99±0.58†* •	12.60±0.36†•	11.26±0.30†* •	12.64±0.36†•	10.94±0.33†*‡•	4.97±0.16*	5.00±0.16*

Results significantly different from others ($p < 0.0001$) (columns 2 to 6 only) as shown in † for the BCLA-MAP model, • (columns 2 to 7 only) for the BCLA-Gibbs model, ‡ for the sSTAPLE, and * (columns 2 to 4, and columns 6 to 8 only) for the EM-R using the two-tailed Wilcoxon rank-sum test. Note that the results of other approaches except BCLA-Gibbs (shaded in grey colour) were reproduced from Table 5.2 for the purpose of comparison.

The Markov chain that satisfies the aforementioned criteria is then chosen to compute the expectation of the individual parameter of interest.

6.5.2 Simulated Dataset

The simulated 548 annotations were bootstrapped 100 times with replacement, each time a RMSE and MAE of the expected inferred labels were generated using the BCLA-Gibbs model with 1,000 sample draws after burn-in. It was compared to the best annotator, BCLA-MAP, EM-R, sSTAPLE, and the mean and median voting strategies. The results are shown in Table 6.1. The RMSE and MAE results show that the BCLA-Gibbs inferred labels significantly outperformed the mean, median, EM-R, sSTAPLE, and best annotator when compared with the simulated *true* annotations. Furthermore, there is no significant difference in results between the BCLA-Gibbs and BCLA-MAP models.

Fig. 6.6 shows an example of the inferred results of bias and precision estimated using the EM-R, sSTAPLE, BCLA-MAP, and BCLA-Gibbs methods. As discussed in the previous chapter, the EM-R algorithm modelled jointly the precision of each annotator and the noise of the underlying ground truth, and thus its estimated σ cannot represent the real precision of each annotator. Furthermore, the EM-R algorithm does not consider the bias of each annotator. Again, although sSTAPLE can provide a reliable estimation of the *true* precision

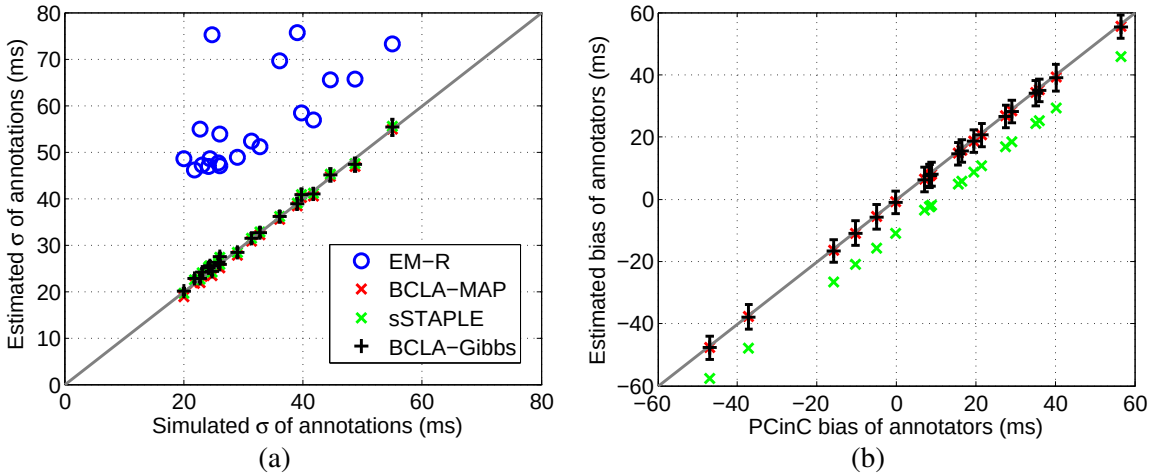


Fig. 6.6 A comparison of the simulated and inferred σ in (a) and bias in (b) of each annotator in the simulated dataset. The results of BCLA-Gibbs are plotted with one standard deviation from the mean (as '+'). Note that EM-R significantly over-estimates the σ values and that sSTAPLE significantly under-estimates the bias values in all simulations.

in the simulated results, it under-estimated all the biases values. In contrast, both BCLA-MAP and BCLA-Gibbs provided a reliable estimation of the *true* precision and bias values of each annotator in the simulated results. Note that both BCLA methods operated in an unsupervised manner without any prior knowledge of which is the best-performing annotator. In addition to BCLA-MAP, the BCLA-Gibbs method provides a confidence interval for each of its estimated values.

6.5.3 QT Dataset

For comparing performance, the EM-R, BCLA-MAP, and sSTAPLE methods took about three seconds for running 5,000 iterations with 10,960 annotations. In comparison, BCLA-Gibbs took approximately 60 seconds to run 2,000 sample draws, and half of the draws were discarded as burn-in to minimise correlation among samples.

The inferred precision and bias results estimated using the EM-R, sSTAPLE, BCLA-MAP, and BCLA-Gibbs methods are shown in Fig. 6.7 (a)-(g) for different divisions in the PCinC QT dataset. As mentioned previously, the EM-R algorithm does not directly model

the precision of each annotator; its estimated σ of each annotator produces an offset from the values provided by the reference annotations. Although the sSTAPLE models the bias of each annotator in addition to the EM-R, it assumed the mean of the biases to be zero in the model, hence consistently produced an under-estimation of bias values. In contrast, both the BCLA-MAP and BCLA-Gibbs methods produced similar results in estimation of the σ and bias values. Their estimation lie closely to the line of identity in comparison to the reference. Although annotator 7 was noticeably under-estimated (see Fig. 6.7 (e) and (f)) by more than 20 ms in σ and bias values when using both of the BCLA methods, it has only annotated 3.47% of records and its annotations varied substantially (see Fig. 5.11), hence making it harder for the BCLA approaches to provide a reliable estimation. Similar results for annotator 4, 3, and 15 are discussed previously in Chapter 5, section 5.6.3. Furthermore, BCLA-Gibbs has shown that each highlighted annotator exhibits relatively large variances in both of its estimated precision and bias values, with one standard deviation of their expectations (i.e., means), indicating a large uncertainty in its estimation. In comparison to BCLA-MAP, BCLA-Gibbs not only provides accurate estimates but also produces confidence in its estimation – this is a key advantage of fully-Bayesian inference.

The RMSE and MAE results are displayed in Table 6.2, and the BCLA methods outperformed all other voting approaches for all divisions. In comparison of BCLA-MAP and BCLA-Gibbs: for Division 2 using 48 open-source algorithms, BCLA-Gibbs achieved a RMSE of 12.11$\pm$0.68 ms, which significantly outperformed other approaches and provides an improvement of 4.27% over BCLA-MAP (RMSE of 12.65$\pm$0.64 ms); in the closed-source Division 3 using 21 algorithms, BCLA-MAP was superior to BCLA-Gibbs with a RMSE of 14.19$\pm$0.87 ms, and had a 4.83% improved error rate. When considering all automated entries (Division 4), BCLA-MAP was slightly worse than BCLA-Gibbs (RMSE of 11.89$\pm$0.66 ms versus 11.68$\pm$0.58 ms), and their MAEs were not significantly different.

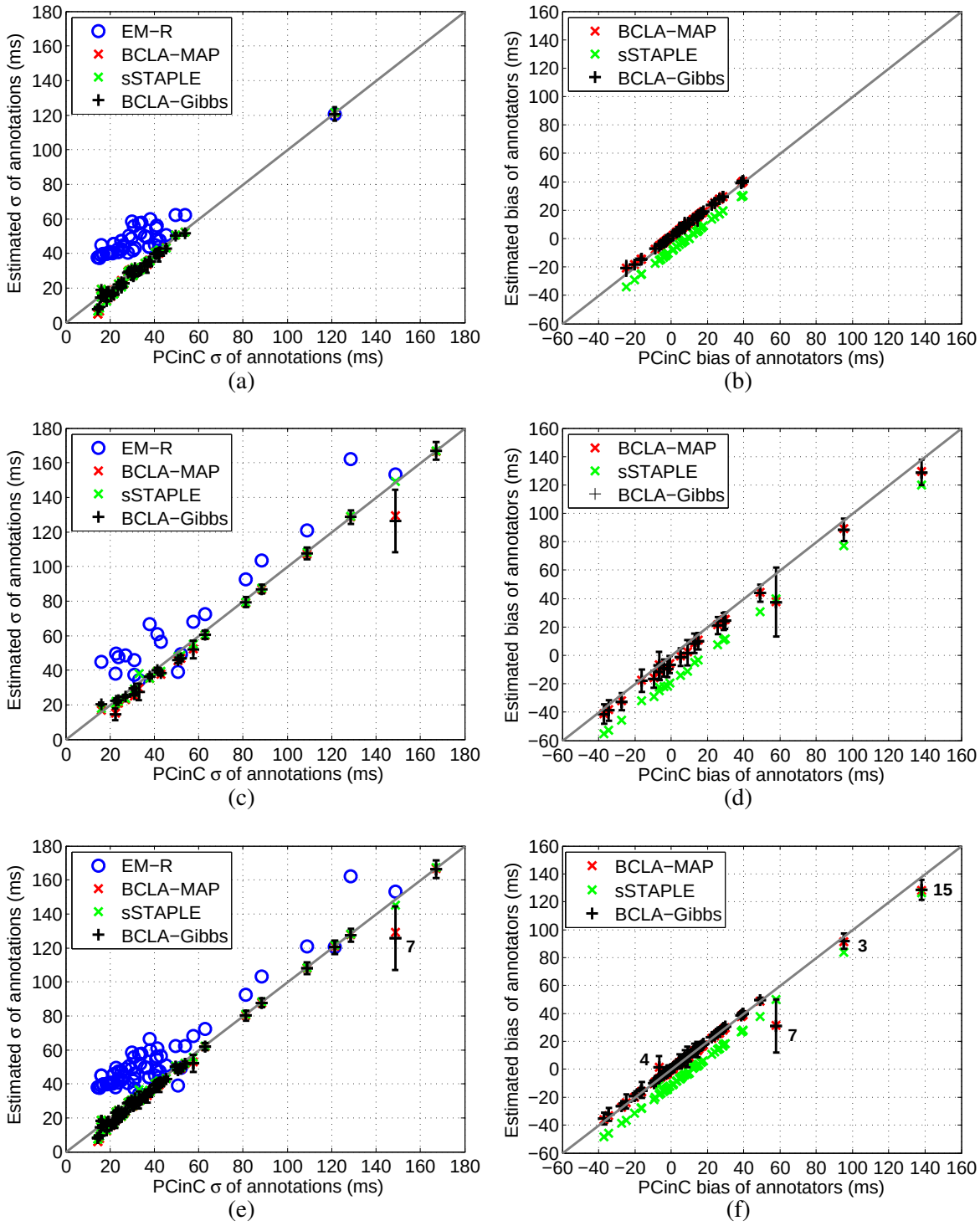


Fig. 6.7 A comparison of the PCinC QT dataset reference and inferred σ and bias of each annotator for Division 2 in (a) and (b), Division 3 in (c) and (d), and Division 4 in (e) and (f) respectively. The results of BCLA-Gibbs are plotted as with one standard deviation from the mean (as '+'). Note that annotators 4, 7, 3, and 15 are labelled in a selection of the plots alone.

Table 6.2 RMSEs and MAEs of inferred labels using different voting approaches in the PCinC QT dataset.

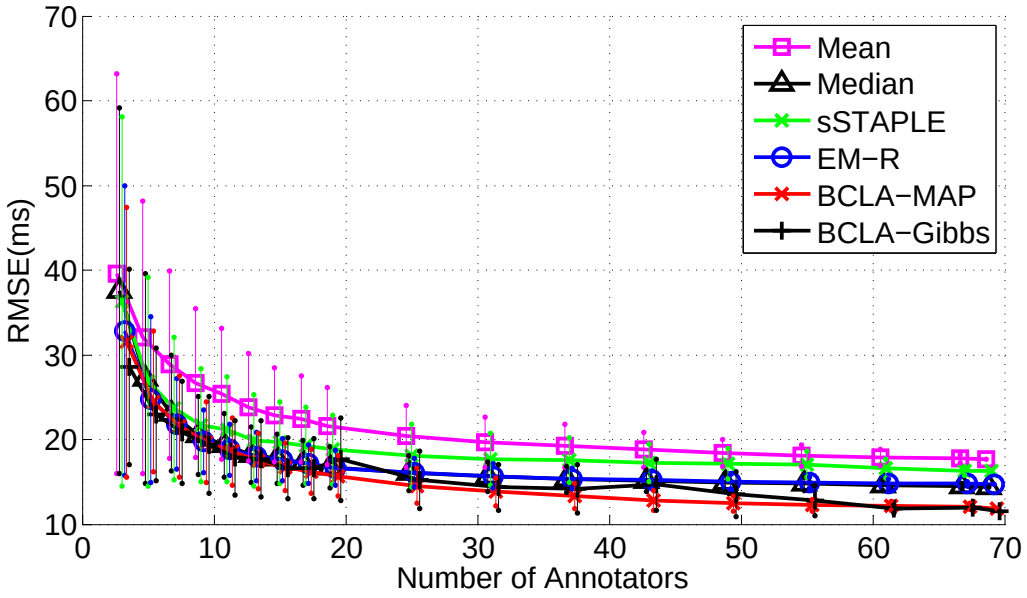
RMSE (ms)							
Division	**Best Annotator**	**Median**	**Mean**	**EM-R**	**sSTAPLE**	**BCLA-MAP**	**BCLA-Gibbs**
2(48)	15.36±0.66*†‡•	15.29±0.56*†‡•	16.17±0.57*†‡•	15.02±0.52†•	15.20±0.99†•	12.65±0.64*‡•	12.11±0.68*‡
3(21)	16.75±1.81*†‡•	19.13±0.83*†‡•	30.68±1.46*†‡•	18.89±0.83†‡•	22.33±1.08*†•	14.19±0.87*‡•	14.90±0.87*‡
4(69)	15.12±1.22*†‡•	14.44±0.52*†‡•	17.66±0.57*†‡•	14.75±0.54†‡•	16.32±0.61*†•	11.89±0.66*‡•	11.68±0.58*‡
MAE (ms)							
Division	**Best Annotator**	**Median**	**Mean**	**EM-R**	**sSTAPLE**	**BCLA-MAP**	**BCLA-Gibbs**
2(48)	10.80±0.57*†‡•	11.75±0.42†•	12.64±0.44*†‡•	11.80±0.43†•	11.64±0.65†•	9.34±0.43*‡•	8.87±0.43*‡
3(21)	10.62±1.14*‡•	14.05±0.55†‡•	22.99±0.83*†‡•	14.10±0.61†‡•	19.15±1.78*†•	10.60±0.69*‡•	11.18±0.70*‡
4(69)	10.73±0.86*†‡•	11.23±0.39*†‡•	14.22±0.45*†‡•	11.50±0.43†‡	13.23±0.49*†•	8.60±0.44*‡	8.64±0.48*‡

Results significantly different from others ($p < 0.0001$) as shown in † (columns 2 to 6 only) for the BCLA-MAP model, • (columns 2 to 7 only) for the BCLA-Gibbs model, ‡ (columns 2 to 5, and 7 to 8 only) for sSTAPLE, and * (columns 2 to 4, and 6 to 8 only) for the EM-R using the two-tailed Wilcoxon rank-sum test. Note that the "Best" annotator is defined as the annotator with the least RMSE. Note that the results of other approaches except BCLA-Gibbs (shaded in grey colour) were reproduced from Table 5.3 for the purpose of comparison.

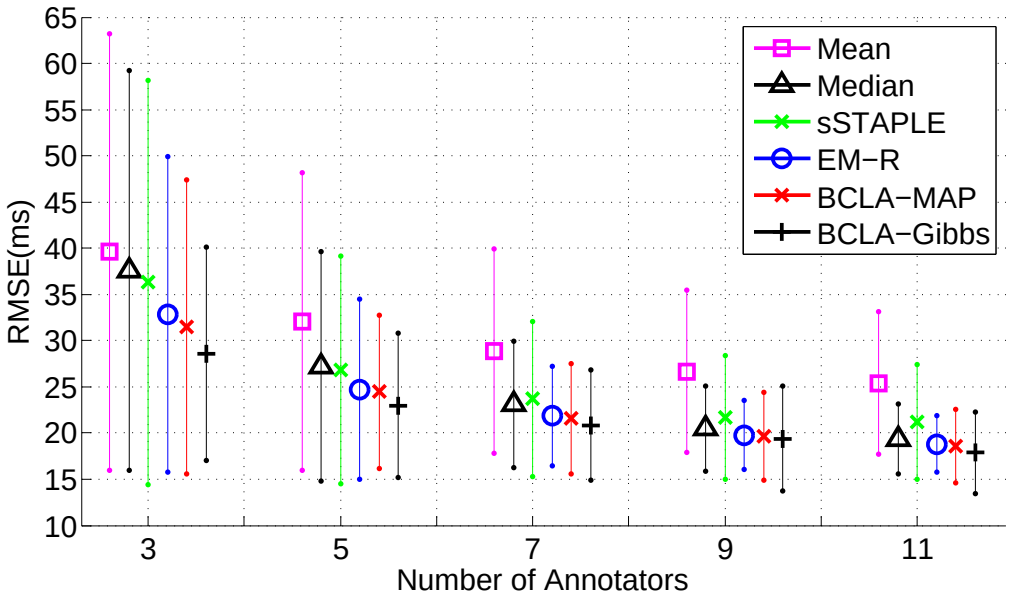
Fig. 6.8 shows a further evaluation of the accuracies in terms of RMSE as a function of the number of annotators. The results were generated by sub-sampling annotators with 1,000 times from three to 69 annotators in Division 4. BCLA-Gibbs outperformed the EM-R, sSTAPLE, mean, and median voting approaches for all number of annotators except when selecting 19 annotators. Note that as the PCinC QT dataset contains missing annotations, this might result in BCLA approaches producing a slightly different error value depending on the recordings that were selected. In comparison to BCLA-MAP, BCLA-Gibbs outperformed it initially when selecting smaller number of annotators for up to 11 annotators, or when there are large number of annotators such as using 61 or more annotators onwards. This has highlighted the advantage of using a fully-Bayesian approach over the point-based estimation of MAP, particularly when fewer number of annotators are available: the MAP approach is essentially an approximation of the most probable estimate of the parameter of interest in the likelihood weighted by the prior provided. Assuming the same prior distribution for different numbers of annotators might not be optimal, especially when annotations are sparse or annotators are scarce, and hence the MAP approach showed suboptimal performance when compared to that obtained via Gibbs sampling.

Focusing on the analysis of the performance between Gibbs sampling and MAP, their differences in RMSE for 1,000 sub-samples were plotted in Fig. 6.9. Out of the 1,000 runs bootstrapped from a total of 69 annotators, the results show that the mean and median of the RMSEs of the BCLA-Gibbs results were smaller than those of BCLA-MAP for up to 11 annotators, as well as the case when the number of annotators selected was 61 or more. In addition to mean and median RMSE comparison, the percentage of runs that BCLA-Gibbs outperformed BCLA-MAP (i.e., with smaller RMSE) is shown in the bottom of the Fig. 6.9 for selecting different numbers of annotators.

When comparing the RMSE performance of the best-performing algorithm in the PCinC Challenge, BCLA-Gibbs had on average 21.16%, 11.04%, 22.75% improvement for Di-



(a) Random sample of three to 69 annotators.



(b) Inset: A close-up of the RMSE results when using 11 annotators or less.

Fig. 6.8 The mean and standard deviation of the RMSE results of using different voting approaches are shown as a function of the number of automated annotators. The plot was generated by randomly sampling the annotators 1,000 times. Note that the RMSE results of other voting approaches except BCLA-Gibbs were reproduced from Fig. 5.12 for the purpose of comparison.

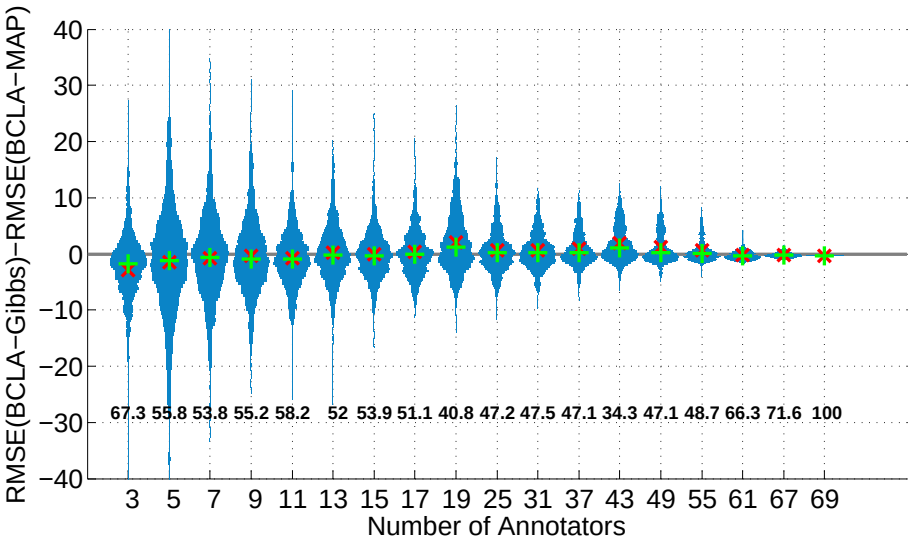


Fig. 6.9 The difference in the RMSE results between BCLA-Gibbs and BCLA-MAP as a function of the number of automated annotators. A probability density estimate of the sub-sampled 1,000 RMSE differences is calculated for selecting number of annotators from three to 69. The mean and the median of the density estimate are labelled as $\times$ in red and $+$ in green respectively. The differences in RMSE that are less than zero were also computed as percentages as shown in the bottom of the plot. They indicate the percentage of times that BCLA-Gibbs outperformed BCLA-MAP from a total of 1,000 sub-sampled RMSEs.

visions 2, 3, and 4, respectively. In the case where only three automated algorithms were selected at random, BCLA-Gibbs had on average 9.30%, 12.92%, 21.27%, 23.91%, and 27.81% improvement over the BCLA-MAP, EM-R, sSTAPLE, median, and mean voting approaches respectively.

6.5.4 Capnobase RR Dataset

The resulting values of the parameters and hyperparameters of the BCLA model for the Capnobase dataset are listed in Table 6.3. The mean of the biases μ_ϕ varied for individual subjects, and was calculated using the median of the RR estimates across all windows. The average time for fusing 900 estimates from six algorithms for 5,000 iterations using BCLA-MAP was 1.57 seconds, while BCLA-Gibbs took 38.80 seconds for 5,000 draws using MATLAB R2011a on a 3.3GHz Intel Xeon processor. Both approaches suggest that,

Table 6.3 The parameters of BCLA for modelling the Capnobase RR dataset

Symbol	Definition	Value
k_b	shape of Gamma distribution for b	3
ϑ_b	scale of Gamma distribution for b	0.006
μ_ϕ	mean of the bias distribution	variable †
k_α	shape of Gamma distribution for α_ϕ	5
ϑ_α	scale of Gamma distribution for α_ϕ	0.1
k_λ	shape of Gamma distribution for $\boldsymbol{\lambda}$	3 ‡
ϑ_λ	scale of Gamma distribution for $\boldsymbol{\lambda}$	0.02 ‡

N.B.: b is the precision for the estimate of the ground truth. α_ϕ is the precision for the estimate of the bias from ground truth. $\boldsymbol{\lambda}$ refers to signal-specific precisions. The values with ‡ are determined with the assumption that the RR estimates provided by the best modulation signal is ± 2 bpm away from the reference [179, 180]. The values with † are estimated from the median RR estimates provided by the algorithms.

in terms of computational time, the proposed BCLA framework is appropriate to be used for real-time production of RR estimates. The burn-in period varied depending on the number of algorithms considered: when RR estimates of the six algorithms (AR+FFT) were fused together, 5,000 sample draws were sufficient and the first 3,000 points were discarded as burn-in. The same number of draws were considered for fusing three algorithms derived from the FFT-based method. In comparison, the inference process for the three algorithms using the AR-based method was faster to converge and required smaller numbers of draws, resulting in 2,000 burn-in draws from 3,000. The criteria for convergence was to examine visually the convergence of the posterior mean of the biases μ_ϕ^* for each individual subject, and an example of the mean of biases (i.e., μ_ϕ) per sample draw is shown in Fig. 6.10. Note that the posterior samples of the mean of the biases approximates to the precision weighted μ_ϕ as there are limited observations in the RR dataset per subject. As noted previously, the convergence process is dataset-dependent, and is a function of the number of algorithms and annotations. It is therefore difficult to generalise concerning the required amount of the sample draws for convergence for each subject. The Gelman-Rubin diagnostic [172] was applied to the RR annotations (either from FFT or AR or both modelling approaches) for

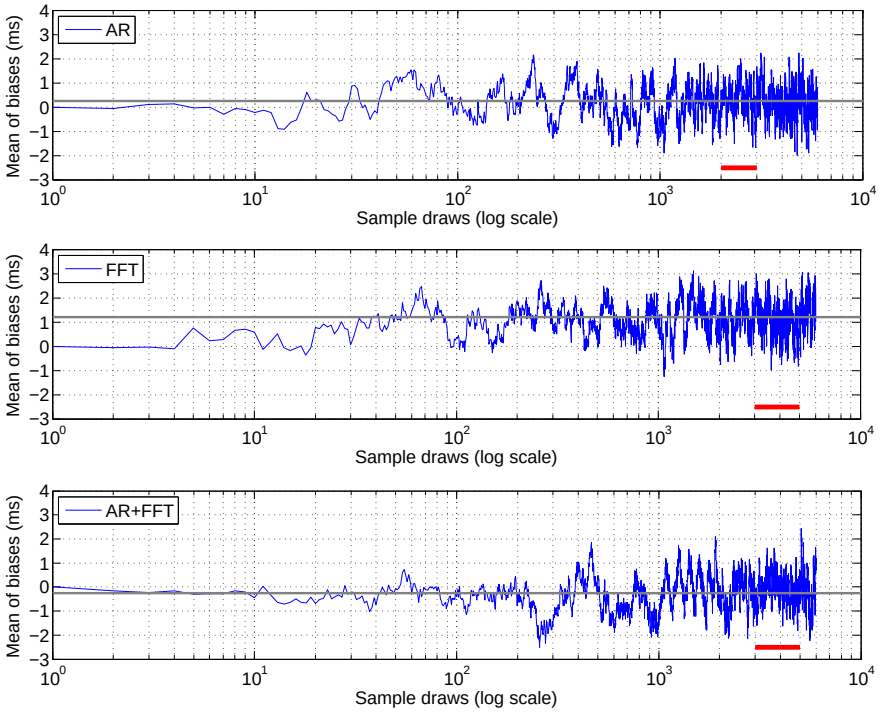


Fig. 6.10 Example of sample draws of the posterior mean of biases μ_ϕ^* using the BCLA-Gibbs model for one subject. The sample draws in the x-axis are shown on a logarithmic scale. The mean of the biases, μ_ϕ, is shown as the grey horizontal line for each sub-plot. The red bar in each sub-plot indicates the samples considered for estimating the expectation.

individual subjects to ensure that the number of posterior samples used to estimate expectation fulfils the requirement that their potential scale reduction factors across records/annotators were all within the limit (i.e., approximates to 1).

The mean and standard error of the MAE for each fusion strategy are shown in Table 6.4. The "Smart Fusion" is denoted as Mean* in the table as it is a variation of mean voting; it also discards some windows of RR estimates if they had a standard deviation greater than four bpm. An example of the percentage of windows used for each subject is shown in the lower plot of Fig. 6.11.

RR estimates fused via the BCLA approaches had lower MAEs than any of the voting approaches when using either the FFT-based, or the AR-based, or both combined (FFT+AR) methods. In the case of using the AR-based algorithm, BCLA-MAP achieved a MAE of 3.02±0.41 bpm and BCLA-Gibbs improved further with a MAE of 2.99±0.340 bpm. In the

Table 6.4 Comparison of MAE and standard error (SE) across 42 subjects in the Capnobase RR dataset.

Algorithms	**MAE ± SE (bpm)**							
(number)	**Mean**	**Mean***	**Best**	**Median**	**EM-R**	**sSTAPLE**	**BCLA-MAP**	**BCLA-Gibbs**
AR (3)	**3.58±0.41**	**3.26±0.41**	**3.31±0.41**	**3.25±0.41**	**3.12±0.42**	**3.51±0.42**	**3.02±0.41**	**2.99±0.40**
FFT (3)	**2.92±0.40**	**2.02±0.39**	**2.39±0.43**	**2.22±0.37**	**2.03±0.38**	**2.63±0.42**	**1.97±0.39**	**1.95±0.39**
AR+FFT (6)	**2.95±0.39**	**2.22±0.41**	**2.39±0.43**	**2.33±0.38**	**2.30±0.42**	**2.86±0.40**	**1.94±0.36**	**1.89±0.36**

Note: the “Smart Fusion” [81] is denoted as Mean* in the table as it is a variation of the mean voting, and it discards some windows of RR estimates if they had a standard deviation greater than four bpm.

case of using the same FFT-based algorithm as “Smart Fusion” [81], the BCLA-MAP method achieved a MAE of 1.97±0.39 bpm and used all available windows. BCLA-Gibbs provided a slightly improved result with a MAE of 1.95±0.39 bpm. Both results were smaller than the MAE of those of the “Smart Fusion” method (i.e., Mean* in Table 6.4 with a MAE of 2.02±0.39 bpm), even when it disregarded data (and therefore did not produce estimates) from 36.5% of windows. A key advantage for BCLA in the domain of noisy signals (such as estimating respiratory rate) is that it does not discard data, and produces robust results across all windows.

The MAE values for the AR-based algorithms were much worse than those of FFT-based algorithms, and yet the BCLA approaches were able to distinguish the quality of each algorithm’s output based on their noise variance as well as their accuracy. This was true even in cases where there was a mixture of “good" and “poor" algorithms (e.g., FFT+AR). Modulation signals with high accuracy and low noise variance were favoured by the BCLA approaches, which demonstrates the ability of the BCLA framework to produce high-accuracy fused estimates in an unsupervised manner from a set of candidates, some of which may be individually poor estimates.

The difference in MAE for each subject (i.e., the MAE of the “Smart Fusion” subtracted from the MAE of the BCLA approaches) is shown in the upper and middle plots of Fig. 6.11.

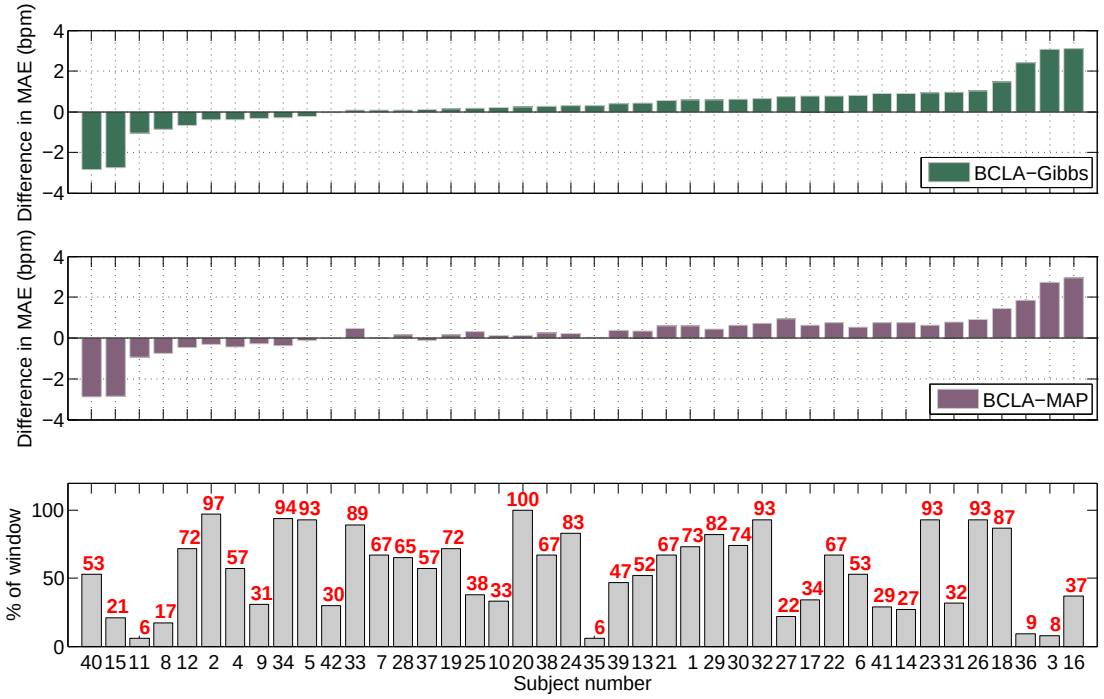


Fig. 6.11 Differences in MAE between "Smart Fusion" and BCLA-Gibbs and BCLA-MAP for 42 subjects are shown in the upper and middle plots, respectively. A positive difference indicates an improvement of MAE (i.e., smaller MAEs) using the BCLA approaches when fusing outputs from six (FFT+AR) algorithms. The lower plot shows the percentage of windows considered when using the "Smart Fusion" for individual subjects (see numbers in red). The numeric labels on the x-axis correspond to the subject IDs. Note that BCLA approaches use all available windows.

A positive difference in MAE indicates an improvement of the BCLA approaches in bpm whereas a negative difference in MAE implies the opposite. The MAE improved with BCLA-MAP (i.e., positive difference) for 69.1% of subjects; when windows were excluded if the standard deviation of RR estimates was greater than four, BCLA-MAP has a MAE improved for 81.0% of subjects. Furthermore, the MAE improved with BCLA-Gibbs for 73.8% of subjects when using all available windows, and 83.3% of subjects for excluding those with a large standard deviation as aforementioned. In addition, BCLA-Gibbs had 61.9% of subjects with smaller MAEs than those of BCLA-MAP using all windows when fusing six algorithms.

Fig. 6.12 shows MAEs across 42 subjects for different methods of fusion. The results were divided into three groups depending on the agreement of the estimates from six algorithms;

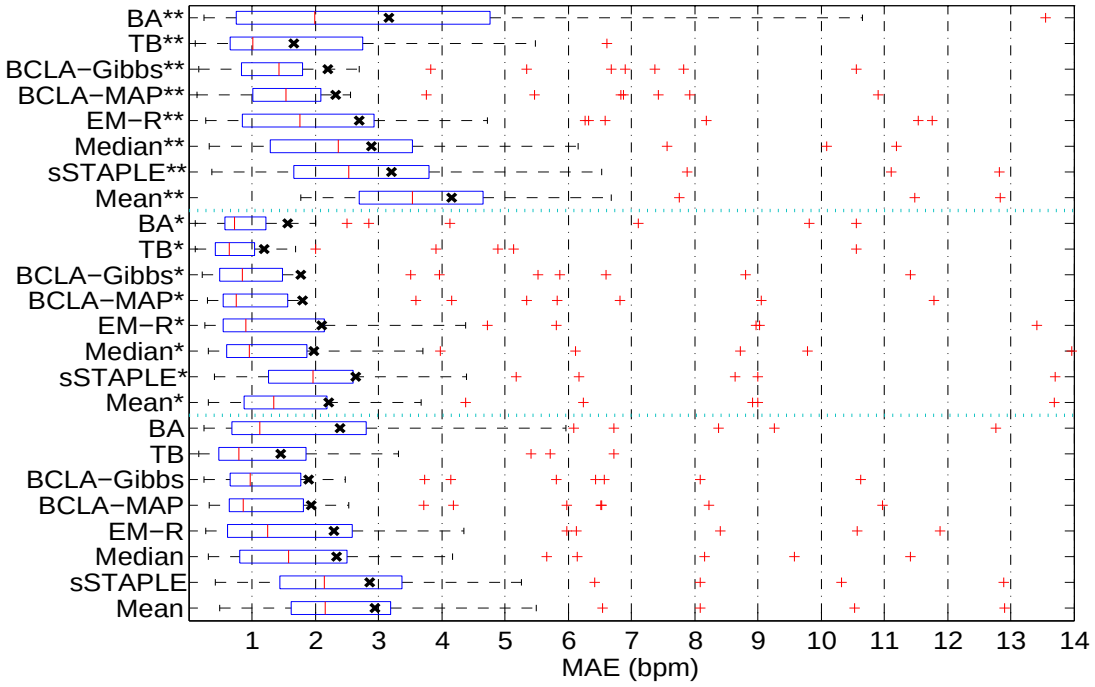


Fig. 6.12 Box plot of MAEs across 42 subjects for different fusion approaches. The median MAE is indicated in red '-' within each box, while the black '×' indicates the mean MAE. The span of each box indicate the interquartile MAE range over 42 subjects for each fusion method when combining six algorithms. Notation: BA – the single best performing algorithm with least MAE across all subjects; TB – theoretical best algorithm with least MAE selected per subject; Mean* – "Smart Fusion". Note that results associated with * indicate that the MAEs were derived from windows excluding those have a standard deviation greater than four bpm, while those associated with ** indicate the results derived only from windows that have a standard deviation greater than four bpm.

i.e., whether the standard deviation of their RR estimates was greater than four bpm: (1) Fusion results with all available windows; (2) Fusion results with windows discarded that have standard deviation greater than four bpm (denoted *); (3) Fusion results with windows that have a standard deviation greater than four bpm (denoted as **). The results demonstrate that both the BCLA-MAP and BCLA-Gibbs methods have optimal performance with their mean MAEs being closest to the theoretical best algorithm (i.e., selecting the best algorithm with the least MAE per subject) for the three different groups. In the case where all windows or only those with standard deviation greater than four bpm were considered, the proposed

BCLA methods outperformed the BA (i.e., "Best" algorithm), EM-R, sSTAPLE, mean and median voting approaches. For the case when discarding windows with standard deviation greater than four bpm, the proposed BCLA methods outperformed others except BA*. The mean MAE results of BCLA-Gibbs always outperform those of BCLA-MAP for all three groups. Although the results of BA* outperformed those of BCLA*, the comparison was only based on 55.4% of the windows. When considering all windows, and thereby reporting RR estimates for all data, the BCLA approaches were superior over other fusion strategies, demonstrating its ability to perform reliable fusion in the face of substantial noise.

The common "fusion" approaches (such as mean, median, and "Smart Fusion") of the RR estimates derived from the different modulations are typically very straightforward (resulting in poor overall fusion results), and tend to substantially reduce the number of windows for which an estimate of RR is computed. In the case for designing non-invasive sensors for real-time RR estimates, it is important to consider all windows and all modulation signals that are available to provide a robust estimation. As the MAE results demonstrated in Fig. 6.12 across 42 subjects, the proposed BCLA approaches produced the least MAEs over other fusion methods like EM-R and sSTAPLE. Although BCLA-MAP had a smaller median MAE than BCLA-Gibbs, it is difficult to conclude its significance based on the results using only 42 subjects. Furthermore, BCLA-Gibbs had 61.9% of subjects with smaller MAEs than those of BCLA-MAP (see Fig. 6.11). Nevertheless, both BCLA-MAP and BCLA-Gibbs had a median MAE below 1 bpm. In terms of computational time for running practical systems, BCLA-MAP has an advantage over BCLA-Gibbs in producing RR estimates. However, BCLA-MAP does not provide an accuracy confidence interval in its estimation, which might be less informative in comparison to a fully-Bayesian approach such as BCLA-Gibbs.

6.6 Summary and Future Work

This chapter has detailed the Gibbs sampling approach for extending inference with the BCLA framework to a fully-Bayesian setting. It was applied to the simulated and PCinC QT datasets, and to the Capnobase RR dataset in an unsupervised manner to infer the unknown ground truth as well as the bias and precision values of each algorithm:

- In the PCinC QT dataset, BCLA-Gibbs outperformed all non-BCLA approaches when considering all annotations available. In comparison to BCLA-MAP, the BCLA-Gibbs method has provided a small improvement in RMSE for Divisions 2 and 4, where the number of annotators was relatively large (i.e., 48 and 69 respectively), but inferior RMSE in Division 3 when 21 algorithms were used. When a comparison was made as a function of annotators and only three automated algorithms were randomly selected, BCLA-Gibbs had on average 9.30%, 12.92%, 21.27%, 23.91%, and 27.81% improvement over the BCLA-MAP, EM-R, sSTAPLE, median and mean voting approaches respectively. The BCLA-Gibbs approach has proved to be an improvement over the BCLA-MAP approach when fewer algorithms (up to 11) are available. This further shows a potential of using BCLA-Gibbs in crowd-sourcing applications, where a task is labelled by a maximum of 10 or less annotators [181, 182].

- In the case where annotations are scarce and fewer annotators are available in the Capnobase RR dataset, the BCLA-MAP and BCLA-Gibbs methods were shown to be superior to other approaches, with BCLA-Gibbs producing the smallest MAE of 1.89 ± 0.36 bpm when fusing six algorithms derived from FFT- and AR-based modelling across 42 subjects. Furthermore, BCLA-Gibbs had 61.9% of subjects with smaller MAEs than those of BCLA-MAP.

For the Capnobase dataset, features such as age could be incorporated into the BCLA model in future, to adjust the mean of the estimated latent ground truth to fit a subject

population closely, noting that different age groups tend to exhibit different ranges of respiratory rate. Furthermore, BCLA assumes that each record is independent as RR was estimated per individual window extracted from modulation signals. Using the same dataset, Pimentel *et al*. [183] has demonstrated the potential of using Gaussian process (GPs) to model the time dependency in the RR estimates. Future work could fuse additional results extracted from GPs to further improve the accuracy in RR estimates.

The current BCLA model assumed consistent performance of each annotator throughout the records; i.e., that the performance is time-invariant. In the context of manual labelling, this might not be true over an extended period of time where a manual annotator's performance might improve through learning, or his performance might drop due to inattention or fatigue. The nature of the datasets being considered in this work are such that one can assume that performance across records is approximately consistent for each algorithm. Future work will include modelling the performance of each manual annotator varying across records and through time to provide a more reliable estimation of the aggregated ground truth for datasets in which intra-annotator performance is highly variant.

Chapter 7

BCLA with Signal Quality Extension

"A wise man, therefore, proportions his belief to the evidence."

– David Hume

7.1 Introduction

This chapter investigates the feasibility of incorporating signal quality into the BCLA model to improve the error in the inferred aggregation of the ground truth. The signal quality is considered as a predictor feature of noise and artefact extracted from the medical data, and therefore can be provided as a confidence measure into the modified BCLA model, denoted as BCLAs. As a proof-of-concept, BCLAs was applied to the PCinC QT dataset, and it was compared to results obtained using BCLA with no features (see Chapters 5 and 6) and with the beat-specific heart rate feature. It was also applied to the Capnobase RR dataset to attempt to improve the inferred RR estimation across individual subjects.

7.2 Background

The BCLA model described previously in Chapters 5 and 6 assumed that the precision of each annotator is not affected by the signal quality of the recordings (i.e., their annotations would remain the same for recordings with different level of signal qualities that are affected by noise). This may not be necessarily true in the context of medical applications. Noise is a prominent source of inaccurate detection of medical data. In the application of electrocardiogram (ECG), large amounts of noise can increase the false-positive alarm rate in arrhythmia analysis [184]. Examples of ECG noise can be categorised as follows [158]: (1) power-line interference, which is also commonly called 50/60Hz noise as it depends on the frequency of the power supply; (2) electrode contact noise that is caused by loss of contact between the electrode and the skin; (3) patient-electrode motion artefact which are changes caused by electrode-skin impedance variation due to electrode motion; (4) electromyographic noise that is caused by muscle contraction near the electrode; and (5) baseline drift due to respiration and changes of electrode impedance. Noise also results in poor diagnostic ECG quality which leads to incremental increases to workloads of physicians when interpreting signals or skewing statistics using an automatic algorithm. Having non-experts (i.e., human novices or automated algorithms) to label ECGs creates high levels of disagreement that could effectively reduce the diagnostic accuracy.

The signal quality of a recording is a predictor of its noise level prior to annotations, and is a measurable variable, which can be obtained in the pre-processing stage preceding the voting algorithm that aggregates annotations to infer a single ground truth. Instead of using the signal quality as an indicator for removing noisy segments which might cause a large sections of data to be discarded [81], we can incorporate signal quality as an independent variable into the BCLA model. Previous studies had used the signal quality in a similar context to estimate confidence for a given segment of medical data: Li *et al.* [185] combined non-linearly weighted signal quality indices to scale the noise covariance in the Kalman filter

to provide a reliable prediction. Tsanas *et al.* [186] applied the same methodology and also included an update of the state noise covariance in the Kalman filter as a function of signal quality. Pimentel *et al.* [187, 188] considered an estimate of signal quality as a confidence measure of the R-peak detections in the ECG and the arterial blood pressure, and combine their confidence scores as a probabilistic input to a hidden semi-Markov model. Vollmer [189] used the difference between maximum and minimum of a smoothed window of ECG to indicate the signal quality of a segment that provided a confidence in the heart beat extraction. Johannesen *et al.* [190] used physiological constraints as a means of estimating signal quality to provide confidence in labels of different heart-beat locations for multiple physiological waveforms. Johnson *et al.* [191] implemented a signal quality estimate to switch between segments from multimodal data to obtain more reliable estimation of R-peak detection.

Signal quality can also be seen as a measure of task difficulty: noisy records are harder to label due to noise contamination of the signals, while clean records can be seen as easier tasks for labelling. Experts are able to "filter" noise to some degree and provide consistently reliable annotations across data of differing noise levels, whereas non-experts might mistake noise for intrinsic features of the signals. An example of a noisy ECG signal is shown in Fig. 4.2 with a large variation in annotation (i.e., 100 ms) between two human annotators. The modified BCLA model (i.e., BCLAs) incorporates the signal quality that acts as a probabilistic score with a range of zero to one, which is described in the following sections.

7.3 BCLA with Signal Quality (BCLAs)

Assuming not all annotators provide labels on the same segment or recording, the signal quality for the ith record from a jth annotator is denoted as t_i^j. The signal quality is assumed to be within the range of (0,1], where 1 indicates a good signal quality and the segment is clean; while any value close to 0 indicates a noisy segment which is corrupted and any

annotations generated from such segment would not produce any meaningful results. Note that the signal quality cannot be 0 as will be explained below.

The graphical representation of BCLAs is illustrated in Fig. 7.1. It assumes that y_i^j is a noisy version of z_i, drawn from a normal distribution as $\mathcal{N}\left(y_i^j \mid z_i + \phi^j, \left(t_i^j \lambda^j\right)^{-1}\right)$. λ^j is the precision of the jth annotator, defined as the estimated inverse-variance of annotator j. t_i^j acts as a scaling factor for λ^j, with $t_i^j \to 0$ indicates annotator j is less precise in labelling the ith recording as it is of noisy quality. If we set $t_i^j = 0$, this would imply zero precision (or infinite variance) for each annotator in the model, and thereby a lower bound of t_i^j needs to be obtained to define the range of "good quality" data and this is specific for a given dataset. The bias for jth annotator, ϕ^j, is also assumed to be drawn from a normal distribution as $\mathcal{N}\left(\phi^j \mid \mu_\phi, 1/\alpha_\phi\right)$. Furthermore, the ground truth, z_i, has its own distribution as $\mathcal{N}(z_i \mid a_i, 1/b)$, where a_i can be expressed as a linear regression function $f(\mathbf{w}, \mathbf{x_i})$ as previously described in the BCLA model. The likelihood with signal quality extension of the parameter $\boldsymbol{\theta} = \left\{\mathbf{w}, \boldsymbol{\lambda}, \boldsymbol{\phi}, \alpha_\phi, b, z_i\right\}$ for a given dataset $\mathbf{D}$ can be written using Bayes' theorem as:

$$
\begin{aligned}
p(\boldsymbol{\theta} \mid \mathbf{D}) &\propto p(\mathbf{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}) \\
&= \text{Gamma}(\alpha_\phi \mid k_\alpha, \vartheta_\alpha)\text{Gamma}(b \mid k_b, \vartheta_b) \times \\
&\quad \left[\prod_{j=1}^{R} \mathcal{N}\left(\phi^j \mid \mu_\phi, 1/\alpha_\phi\right) \text{Gamma}\left(\lambda^j \mid k_\lambda, \vartheta_\lambda\right)\right] \times \\
&\quad \left[\prod_{i=1}^{N} \mathcal{N}(z_i \mid a_i, 1/b) \prod_{j=1}^{R} \mathcal{N}\left(y_i^j \mid z_i + \phi^j, \frac{1}{t_i^j \lambda^j}\right)\right]. \qquad (7.1)
\end{aligned}
$$

7.3.1 The Maximum-a-Posteriori (BCLAs-MAP) Approach

Estimation of $\boldsymbol{\theta}$ can be performed using the maximum-a-posterior (MAP) approach, which maximises the log-likelihood of the parameters, i.e., $\underset{\boldsymbol{\theta}}{\text{argmax}}\{\log p(\boldsymbol{\theta} \mid \mathbf{D})\}$. The log-

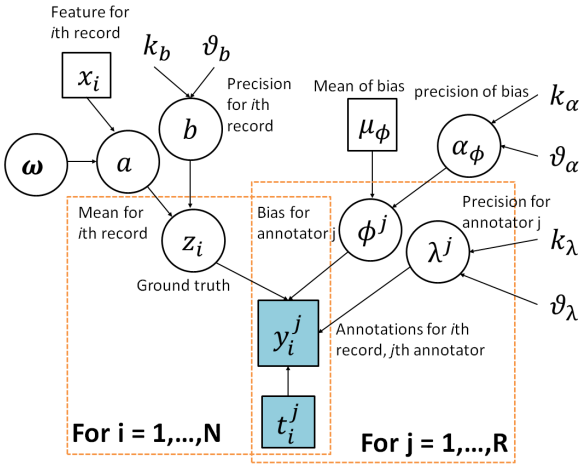


Fig. 7.1 Graphical representation of the BCLAs model: y_i^j corresponds to the annotation provided by the jth annotator for the ith record, and it is modelled by the z_i (the unknown underlying ground truth), the ϕ^j (bias), the λ^j (precision), and signal quality t_i^j. BCLAs is the BCLA model (see Fig. 5.3) with an additional t_i^j component.

likelihood including signal quality t_i^j can be rewritten as:

$$\begin{aligned}
\log p(\boldsymbol{\theta} \mid \mathbf{D}) = &-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{R}\left[\log\left(\frac{2\pi}{t_i^j\lambda^j}\right) + \left(y_i^j - \phi^j - z_i\right)^2 t_i^j \lambda^j\right] \\
&-\frac{1}{2}\sum_{j=1}^{R}\left[\log\left(\frac{2\pi}{\alpha_\phi}\right) + \left(\phi^j - \mu_\phi\right)^2 \alpha_\phi\right] \\
&-\frac{1}{2}\sum_{i=1}^{N}\left[\log\left(\frac{2\pi}{b}\right) + \left(z_i - \mathbf{x}_i^\intercal \mathbf{w}\right)^2 b\right] \\
&+\left[(k_\lambda - 1)\log\lambda^j - \log\left(\Gamma(k_\lambda)\vartheta_\lambda^{(k_\lambda)}\right) - \frac{\lambda^j}{\vartheta_\lambda}\right] \\
&+\left[(k_\alpha - 1)\log\alpha_\phi - \log\left(\Gamma(k_\alpha)\vartheta_\alpha^{(k_\alpha)}\right) - \frac{\alpha_\phi}{\vartheta_\alpha}\right] \\
&+\left[(k_b - 1)\log b - \log\left(\Gamma(k_b)\vartheta_b^{(k_b)}\right) - \frac{b}{\vartheta_b}\right].
\end{aligned} \tag{7.2}$$

The parameters in $\boldsymbol{\theta}$ can be derived by equating the gradient of the log-likelihood to zero respectively as follows:

$$\frac{1}{\lambda^j} = \frac{1}{N_j + 2(k_\lambda - 1)} \left[\sum_{i \in U_j} t_i^j \left(y_i^j - \phi^j - z_i \right)^2 + \frac{2}{\vartheta_\lambda} \right]. \tag{7.3}$$

$$\mathbf{w} = \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^\intercal \right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i z_i. \tag{7.4}$$

$$\phi^j = \frac{\sum_{i \in U_j} t_i^j \left(y_i^j - z_i \right) + \mu_\phi \left(\frac{\alpha_\phi}{\lambda^j} \right)}{\sum_{i \in U_j} t_i^j + \frac{\alpha_\phi}{\lambda^j}}. \tag{7.5}$$

$$\frac{1}{\alpha_\phi} = \frac{1}{R + 2(k_\alpha - 1)} \left[\sum_{j=1}^{R} \left(\phi^j - \mu_\phi \right)^2 + \frac{2}{\vartheta_\alpha} \right]. \tag{7.6}$$

$$z_i = \frac{\sum_{j \in V_i} t_i^j \lambda^j \left(y_i^j - \phi^j \right) + \left(\mathbf{x}_i^\intercal \mathbf{w} \right) b}{\sum_{j \in V_i} t_i^j \lambda^j + b}. \tag{7.7}$$

$$\frac{1}{b} = \frac{1}{N + 2(k_b - 1)} \left[\sum_{i=1}^{N} \left(z_i - \mathbf{x}_i^\intercal \mathbf{w} \right)^2 + \frac{2}{\vartheta_b} \right]. \tag{7.8}$$

U_j is the set of records/segments with annotations provided by the jth annotator, and V_i is the set of annotators that provided annotations for the ith record. N_j is the number of records annotated by the jth annotator.

This MAP problem can be solved using the EM algorithm in a two-step iterative process as before: (i) the E-step estimates the expected *true* annotations for all records, $\hat{\mathbf{z}}$, as a weighted sum of the provided annotations, and can be estimated using equation (7.7); (ii) the M-step is based on the current estimation of $\hat{\mathbf{z}}$ and given the dataset $\mathbf{D}$. The model parameters, $\mathbf{w}$, $\boldsymbol{\phi}$, α_ϕ, b, and $\boldsymbol{\lambda}$ can be updated using equations (7.4), (7.5), (7.6), (7.8), and (7.3) accordingly in a sequential order until convergence, which is described below.

Convergence Criteria for the BCLAs-MAP Approach

The BCLAs-MAP model behaves similarly to BCLA-MAP, and uses the generalised extreme value distribution to model the maxima of the precision distribution as before. We recall that this is performed to restrict the upper bound of the precision values $\boldsymbol{\lambda}$ and to guarantee

convergence of the MAP algorithm. Details of estimating the maxima of precision values can be found in Chapter 5, section 5.3.1.

The BCLAs-MAP Estimate of Precision

Signal quality is introduced in the precision term that models the annotation in the likelihood function, as seen earlier in equation (7.1). We note that the inferred λ^j described in equation (7.3) is in fact the scaled precision for annotator j (denoted $\hat{\lambda}^j$): it is the resulting value of the true precision for annotator j (denoted λ_T^j) de-weighted by the averaged signal quality across N_j records (i.e., $\frac{1}{N_j}\sum_{i\in U_j} t_i^j$). The true precision for annotator j should therefore be corrected as follows:

$$\lambda_T^j = \frac{\hat{\lambda}^j}{N_j}\sum_{i\in U_j} t_i^j. \tag{7.9}$$

7.3.2 The Gibbs Sampler (BCLAs-Gibbs) Approach

A detailed description of the Gibbs sampler was previously described in Chapter 6. In the hth iteration of sampling, we recall that the lth parameter is drawn from its conditional distribution and can be as expressed as:

$$\theta_l^h \sim p\left(\theta_l \mid \boldsymbol{\theta}_{-l}^{h-1}, \mathbf{D}\right), \tag{7.10}$$

where $\boldsymbol{\theta}_{-l}$ denotes all parameters in $\boldsymbol{\theta}$ but l. Each parameter in the likelihood described in equation (7.1) can therefore be updated by sampling from its posterior distribution with its hyperparameters (denoted by *). Derivations of the hyperparameters of conjugate priors are detailed in Appendix B. Different from the BCLA-Gibbs approach previously described in Chapter 6, we note that the Gibbs equations of BCLAs-Gibbs for estimating z_i and ϕ^j are

now dependent on the signal quality t_i^j:

$$z_i \sim \mathcal{N}\left(z_i \mid a_i^*, \frac{1}{b_i^*}\right), \tag{7.11}$$

where

$$a_i^* = \frac{\left(\mathbf{x}_i^\intercal \mathbf{w}\right) b + \sum_{j \in V_i} \left[\left(y_i^j - \phi^j\right) \lambda^j t_i^j\right]}{b + \sum_{j \in V_i} \lambda^j t_i^j}, \quad b_i^* = b + \sum_{j \in V_i} \lambda^j t_i^j.$$

$$\phi^j \sim \mathcal{N}\left(\phi^j \mid \mu_\phi^{j*}, \frac{1}{\alpha_\phi^{j*}}\right), \tag{7.12}$$

where

$$\mu_\phi^{j*} = \frac{\mu_\phi \alpha_\phi + \sum_{i \in U_j} \left(y_i^j - z_i\right) \lambda^j t_i^j}{\alpha_\phi + \sum_{i \in U_j} \lambda^j t_i^j}, \quad \alpha_\phi^{j*} = \alpha_\phi + \sum_{i \in U_j} \lambda^j t_i^j.$$

7.4 Data Description and Methodology of Comparison

7.4.1 Data Description

Signal Quality in the QT Dataset

In the context of the QT dataset considered by this thesis, it is known that an underlying *true* QT interval can be modulated by heart rate variation [1] and signal quality is recording-dependent. As a proof-of-concept, a beat signal quality (denoted as bSQI) proposed by Johnson *et al.* [191] was used in the work described by this chapter. The bSQI variable measures the percentage of the agreement of two R-peak detectors, naming *gqrs* and *jqrs*, within a 5-second segment, evaluated every 1 second. bSQI = 1 indicates 100% agreement between the two detectors, and this serves as a measure that the segment has good quality, and free from noise and artefact. Any segment with bSQI < 1 indicates the disagreement

between detectors which is assumed to be due to the presence of noise and artefact. A description of each detector is outlined below [191]:

(i) *gqrs* is available in Physionet as a Matlab function of the "wfdb" toolbox[1]. The detector consists of a QRS matched filter with a custom-built set of heuristics.

(ii) *jqrs* consists of a window-based peak energy detector proposed by Behar *et al.* [192–194]. Modifications of the detector were made by Johnson *et al.* [191], where (1) the original band-pass filter was replaced with a "Mexican hat" wavelet filtering of the QRS complexes; (2) a heuristic implementation was applied to ensure no detections were made during periods of flat-line; (3) a search-back procedure was included in the case of suspected missed beats.

Signal Quality in the Capnobase RR Dataset

The aforementioned notation of signal quality was also applied to the Capnobase respiratory rate dataset, where a "pulse signal quality" (denoted as pSQI) was obtained from the authors Pimentel *et al.* [188] (and they had adapted their signal quality methodology to be applied to PPG). The pSQI measures the percentage of the agreement between two PPG peak detectors, naming *wabp* and *PUD* within a 32-second window, evaluated every 5 seconds with 50% overlap. The mean of the pSQI values within the 32-second window was estimated to represent the signal quality of that window. A total of 150 pSQI values were obtained for the 8-minute recording for each subject. The PPG peak detectors are outlined below:

(i) *wabp* is available in Physionet as a Matlab function of the "wfdb" toolbox designed by Zong *et al.* [195][2]. The detector locates the onset of PPG pulses that have been re-sampled at 125 Hz, by first converting each waveform into accumulative first-derivative signal, and then performing adaptive thresholding and local-search strategies to identify

[1] http://physionet.org/physiotools/matlab/wfdb-app-matlab/html/gqrs.html.
[2] https://physionet.org/physiotools/matlab/wfdb-app-matlab/html/wabp.html.

pulse onsets. The PPG peak was found subsequently by determining the maximum value in a 150 ms window after the onset locations.

(ii) *PUD* is a pulse waveform delineator, and was adopted from Li *et al.* [145], where the PPG signal was again re-sampled to 125 Hz. The onset, peak, and dicrotic notch of each PPG pulse were detected using *PUD*. The final peak location of each pulse was determined by searching the maximum value within a 150 ms window centred at a detected peak (if a dicrotic notch was not found) that appeared in the pulsatile PPG waveform.

7.4.2 Methodology of Comparison

QT Dataset

As mentioned previously, an optimisation of the lower bound threshold (denoted as T) of bSQI needs to be defined to avoid assigning zero precision values to annotators. The evaluation of T is dataset-specific, and can be estimated analytically: T is set to be ranged within [0.01,1], and with T=1, BCLAs is equivalent to the BCLA model as any values of bSQI ≤ 1 are mapped to 1. Similarly, a value of T = 0.5 indicates that any bSQI values ≤ 0.5 are now mapped to 0.5. For each T value, RMSE of the inferred ground truth is computed, and the optimal threshold for the lower bound of bSQI (denoted as T_{min}) that corresponds to the minimum RMSE can be subsequently found. T_{min} is determined by finding the lower bound on the bSQI that produces the smallest RMSE with respect to the reference provided in the dataset.

The BCLAs model was evaluated in a similar way as the BCLA model, which was described in previous chapters (see Chapters 5 and 6). To test the feasibility of introducing bSQI into the BCLAs model, the same parameter setting (see Table 5.1) was provided as in the BCLA model for latter comparison. The feature matrix of the beat-specific heart rate (bHR)

$\mathbf{X}_{bHR} = \left[\mathbf{x}_1^\intercal, ..., \mathbf{x}_N^\intercal\right]$ for N records was considered, where $\mathbf{x}_i^\intercal = \left[\sqrt{\tilde{RR}_i}^{j=1}, ..., \sqrt{\tilde{RR}_i}^{j=R}\right]$ from R annotators. The $\tilde{RR}_i$ is the median of the five preceding R-R intervals of the annotated beat in the ith record. The results of the 1,000 sub-sampled RMSEs and MAEs using BCLAs with the bHR feature were computed and compared to: (1) EM-R with the beat-specific heart rate and signal quality features described in Chapter 4 (denoted as EM-R with bHRSQIs); (2) BCLA with no features (denoted as NF); (3) BCLA with the bHR feature; (4) sSTAPLE as well as the mean and median voting approaches. Furthermore, BCLAs with the bHR feature inferred precision and bias values of individual algorithms were compared with those estimated using EM-R with the bHRSQIs features, sSTAPLE, and BCLA with the bHR feature.

Capnobase RR Dataset

The BCLAs model was evaluated in the same manner as the BCLA model, which was described in Chapter 6.

7.5 Results

7.5.1 QT Dataset

Fig. 7.2 (a) shows a plot of MAEs across annotators against the reference QT intervals for all recordings provided in Division 4. As there is no direct correlated relationship between MAEs in QT interval estimation and the length of the QT intervals, the figure demonstrates that the errors in annotations were not associated with the physical length of the reference QT intervals. A similar observation is shown in Fig. 7.2(b), where there is no linear relationship between the MAEs across annotators and the averaged bSQI extracted from all recordings. The cumulative distribution function (cdf) of bSQI values for Division 4 is shown in Fig. 7.2(c). The cdf shows that 1.08% of the annotations have bSQI $\leq$ 0.3125, which amounted to 260, 103, and 363 annotations for Divisions 2, 3, and 4, respectively. About

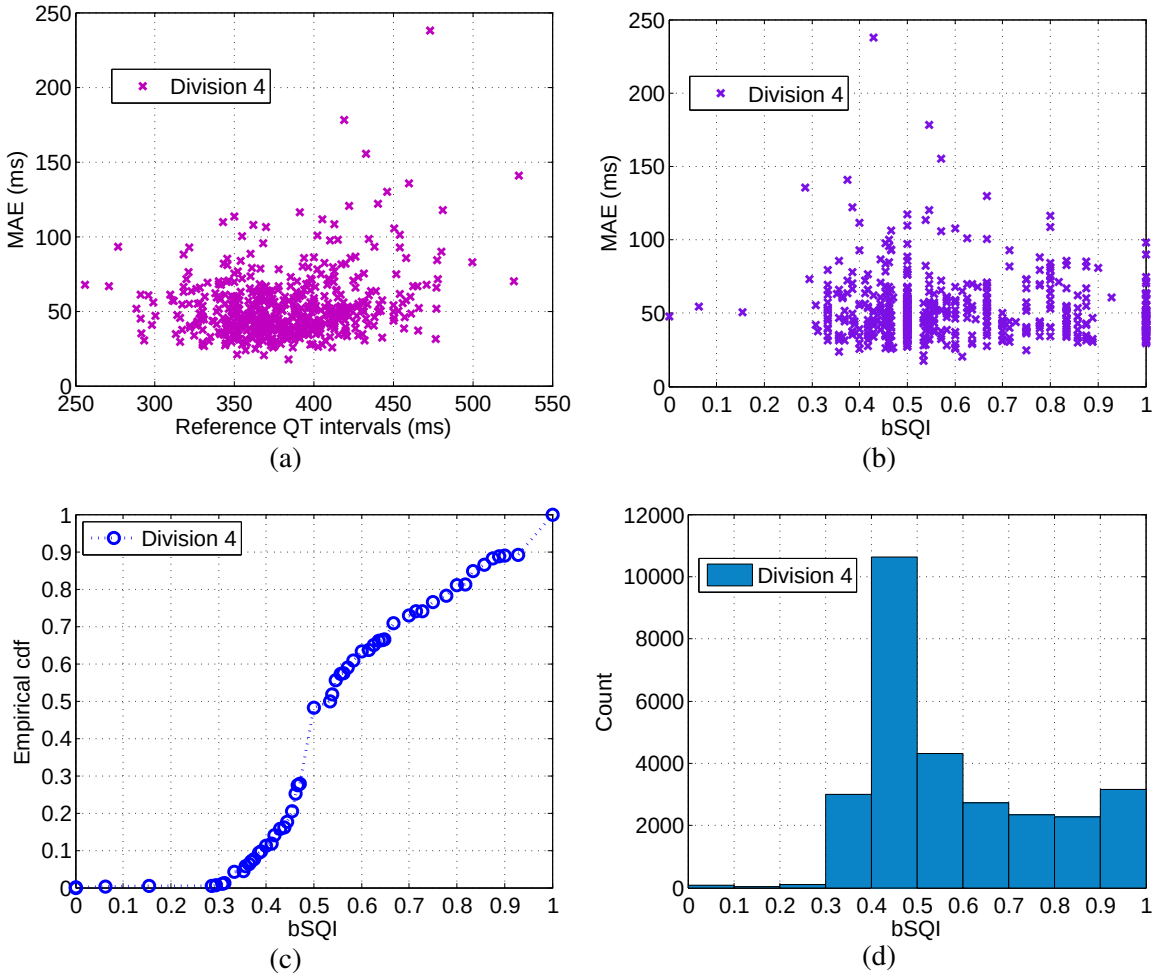
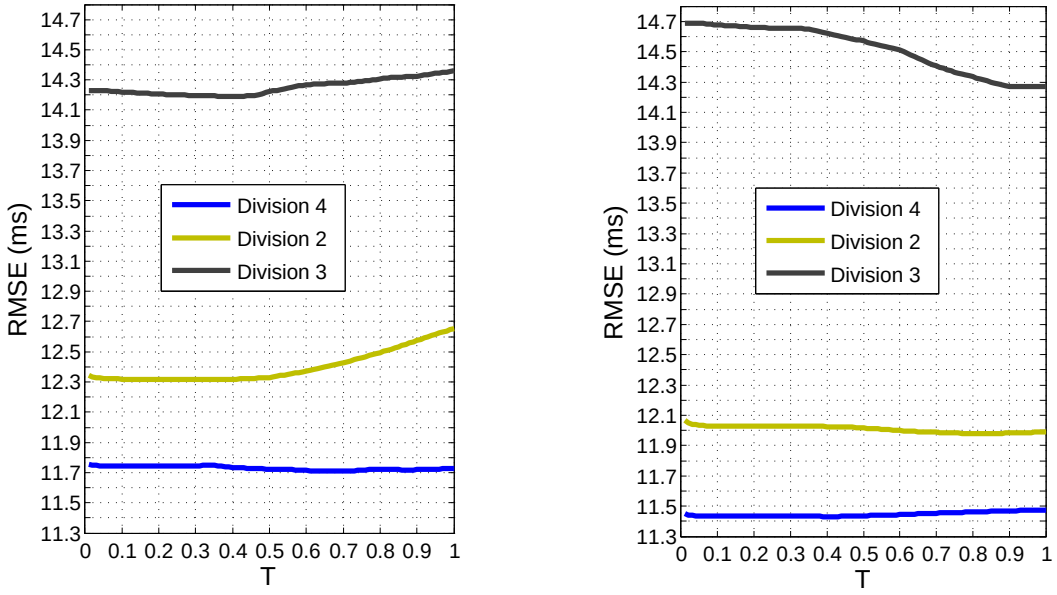


Fig. 7.2 Example of errors and distribution plots for the PCinC QT dataset Division 4: (a) MAEs across annotators for all recordings were plotted against the reference QT intervals; (b) MAEs across annotators versus bSQI values for all recordings; (c) empirical cdf of bSQI values; (d) histogram of the distribution of bSQI values.

15.44% of the annotations have a bSQI ≈ 0.5. This is also demonstrated in the histogram of Fig. 7.2(d), where only 71 annotations had bSQI ≤ 0.1 in Division 4.

Lower Bound Optimisation of bSQI

As descried earlier, the bSQI variable introduced in the BCLAs model takes values within the range of (0,1], where a higher bSQI value approaching 1 indicates that the annotated segment has no obvious noise and artefact, and hence the task of estimating the QT interval may be assumed to be easier than for data with lower bSQI values.



(a) BCLAs-MAP

(b) BCLAs-Gibbs

Fig. 7.3 RMSEs of (a) BCLAs-MAP and (b) BCLAs-Gibbs while varying the threshold of the lower bound T on the bSQI values for different divisions in the PCinC QT database.

Fig. 7.3 shows results of introducing bSQI into the BCLAs model, and the optimal lower bounds on the bSQI values which provide improved RMSEs for the PCinC QT dataset. It was observed across different divisions that a subtle but consistent increase in RMSEs when the lower bound of the bSQI was set to be $T \rightarrow 0$. For Divisions 2 and 3, the effect on RMSE of introducing bSQI into the BCLAs-MAP model may be seem to decrease as the lower bound threshold (i.e., T) increases from 0 to 1 (see Fig. 7.3 (a)). The optimal lower bound of bSQI (i.e., T_{min}) occurs around $T \approx 0.4$ when using the BCLAs-MAP approach, while it appears much later approximately at $T \approx 0.9$ when using the BCLAs-Gibbs approach (see Fig. 7.3 (b)).

Furthermore, it is interesting to note that although different T values affected the RMSEs for Divisions 2 and 3, few changes are observed in Division 4, even though it is the union of Divisions 2 and 3. This could be because as more data are being included into the model to infer the ground truth, BCLAs becomes less reliant on the information encoded by bSQI, and therefore little improvement was observed in RMSE. There also seems to be a trade

Table 7.1 Optimisation of the lower bound T of the bSQI values for each division of the PCinC QT dataset.

Division (number)	BCLAs-MAP		BCLAs-Gibbs	
	RMSE (T_{min})	**T Range**	**RMSE (T_{min})**	**T Range**
2 (48)	12.32 ms (0.33)	0.05 – 0.46	11.98 ms (0.83)	0.73 – 0.94
3 (21)	14.19 ms (0.41)	0.35 – 0.45	14.27 ms (0.95)	0.90 – 1.00
4 (69)	11.71 ms (0.67)	0.60 – 0.74	11.43 ms (0.42)	0.05 – 0.48

Note that a range of T values indicated as T Range for each division produced the same minimum RMSE as the values were rounded to the nearest hundredth accuracy.

off between the number of annotators available versus the RMSE of the inferred ground truth: RMSE for Division 2 (with 48 algorithms) is smaller than that of Division 3 (with 21 algorithms) as shown in Table 7.1.

Table 7.1 shows the optimised lower bound T of the bSQI values for different divisions of the QT dataset. Note that a range of T values (see T Range in Table 7.1) produced the same minimum RMSE (rounded to two decimal places). As the current QT dataset has missing annotations for various algorithms, it was not trivial to perform direct comparison between Divisions 2 and 3 concerning the lower bound on bSQI, as the analysis involves different algorithms. Nevertheless, there was a range of T values for which they overlap between the two divisions, and which resulted in smallest RMSE: [0.35, 0.45] when using the BCLAs-MAP approach and [0.9, 0.94] for the BCLAs-Gibbs approach.

In terms of the performance using MATLAB R2011a on a 3.3GHz Intel Xeon processor, BCLAs-MAP with features took about 20 seconds for running 5,000 iterations with 20,712 annotations. In comparison, BCLAs-Gibbs with features took approximately 114 seconds to run 2,000 draws, with half of the draws being discarded as burn-in to minimise correlation among samples, as before.

Fig. 7.4 shows the inferred bias and precision values of each annotator using BCLAs and BCLA when compared to the reference provided in the PCinC QT dataset. BCLAs with the bHR feature results in inferred σ values that lie close to the line of identity in Fig. 7.4 (a),

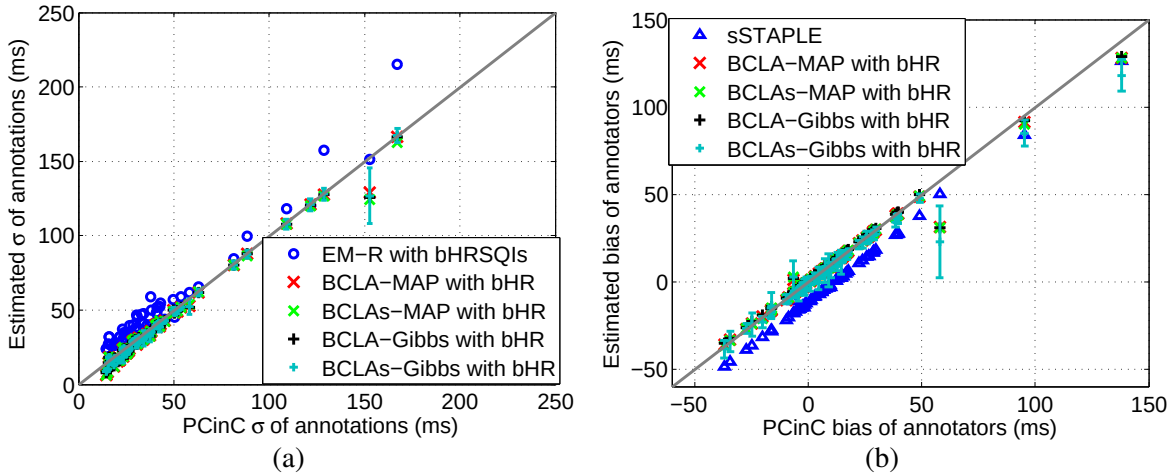


Fig. 7.4 A comparison of the PCinC QT dataset reference, and (a) inferred σ and (b) bias of each annotator for Division 4.

indicating that the BCLAs model can provide a reliable estimation of the precision of each annotator just as well as the BCLA model. In term of the bias estimation as shown in Fig. 7.4 (b), BCLAs with the bHR feature modelled the bias of each annotator accurately.

Comparison of Results

Table 7.2 shows a comparison of the results from BCLAs with the optimised lower bound of the bSQI (T_{min}) and the bHR feature, and those when no features (i.e., NF) or only the bHR feature was used in the BCLA model. They were also compared to the results of EM-R with bHRSQIs, sSTAPLE, mean, and median voting approaches. It may be seen that including the bHR feature into the BCLA model does not always significantly improve results compared with BCLA with no features (NF) (see column 6 versus 7, and column 9 versus 10 in Table 7.2). The RMSE for Division 3 was significantly smaller ($p < 0.05$) when using the BCLA-Gibbs model with the bHR feature versus NF (13.87$\pm$0.78 ms for the bHR feature versus 14.80$\pm$0.84 ms for NF), while this was the opposite for the BCLA-MAP model with the bHR feature (14.39$\pm$0.89 ms for the bHR feature versus 14.19$\pm$0.87 ms for NF).

Table 7.2 RMSEs of inferred labels using different voting approaches with and without features in the PCinC QT dataset.

RMSE (ms)										
Division (number)	sSTAPLE	Mean	Median	EM-R with bHRSQIs	MAP Approach			Gibbs Approach		
					BCLA with		BCLAs with bHR	BCLA with		BCLAs with bHR
					NF	bHR		NF	bHR	
2 (48)	15.20±0.99†‡	16.17±0.57†‡	15.29±0.56†‡	14.48±0.54†‡	12.65±0.70†‡	12.59±0.65*†‡	12.32±0.59†	12.17±0.68†‡	12.15±0.68†‡	12.02±0.59‡
3 (21)	22.33±1.08†‡	30.68±1.46†‡	19.13±0.83†‡	16.38±0.80†‡	14.19±0.87†‡	14.39±0.89*†‡	14.23±0.91†	14.80±0.84††	13.87±0.78*†‡	14.08±0.81‡
4 (69)	16.32±0.61†‡	17.66±0.57†‡	14.44±0.52†‡	14.41±0.53†‡	11.89±0.66†‡	11.81±0.65*†‡	11.72±0.59†	11.59±0.56†‡	11.58±0.56†‡	11.66±0.57‡

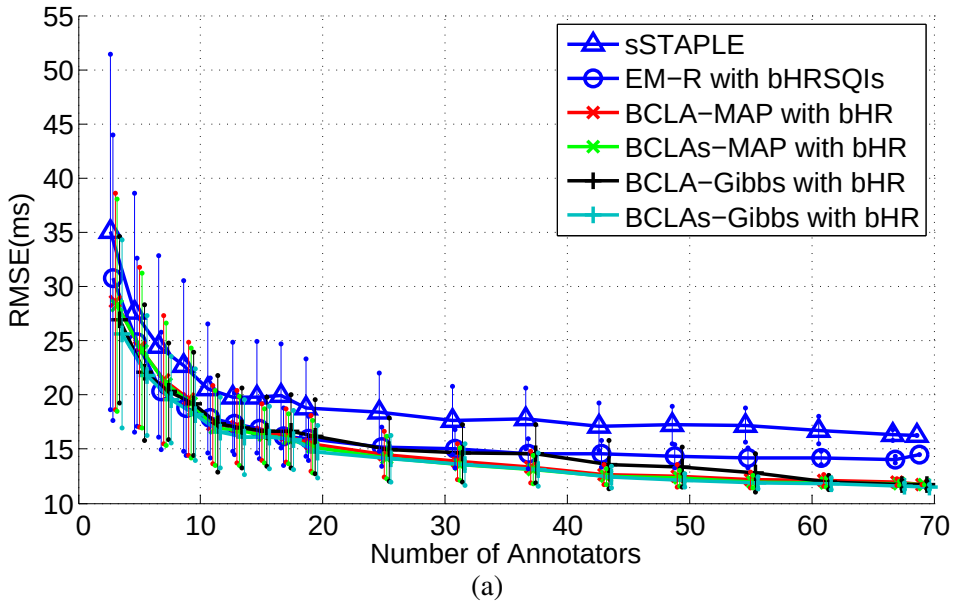
Note: the above are the mean ± standard deviation of the RMSE results obtained from 1,000 subsamples. Results of BCLAs were obtained using the T_{min} as stated in Table 7.1. The † indicates BCLAs-Gibbs results with the bHR feature were significantly different (for column 2 to 10 only) from the rest of the voting strategies ($p < 0.05$) using the two-tailed Wilcoxon rank-sum test. The * (columns 6 versus 7, and columns 9 versus 10 only) indicates the significance ($p < 0.05$) of introducing bHR feature to the BCLA-MAP and BCLA-Gibbs model respectively. The ‡ (columns 2 to 7, and columns 9 to 11) indicates the BCLAs-MAP results with the bHR feature were significantly different from others ($p < 0.05$). MAEs results are in Appendix C.1 Table C.1.

BCLAs-MAP and BCLAs-Gibbs with the bHR feature significantly outperformed the BCLA model with or without the bHR feature for Division 2. RMSE for BCLAs-Gibbs was significantly higher than that for BCLA-Gibbs when both had the bHR feature for Divisions 3 and 4. However, significant improvement was observed in BCLAs-MAP with bHR when compared to BCLA-MAP with bHR for all divisions. Nevertheless, the RMSEs of BCLAs-Gibbs with bHR significantly outperformed those of BCLAs-MAP with bHR for all divisions. Even without the signal quality extension, BCLA-Gibbs with bHR had superior performance over BCLAs-MAP with bHR.

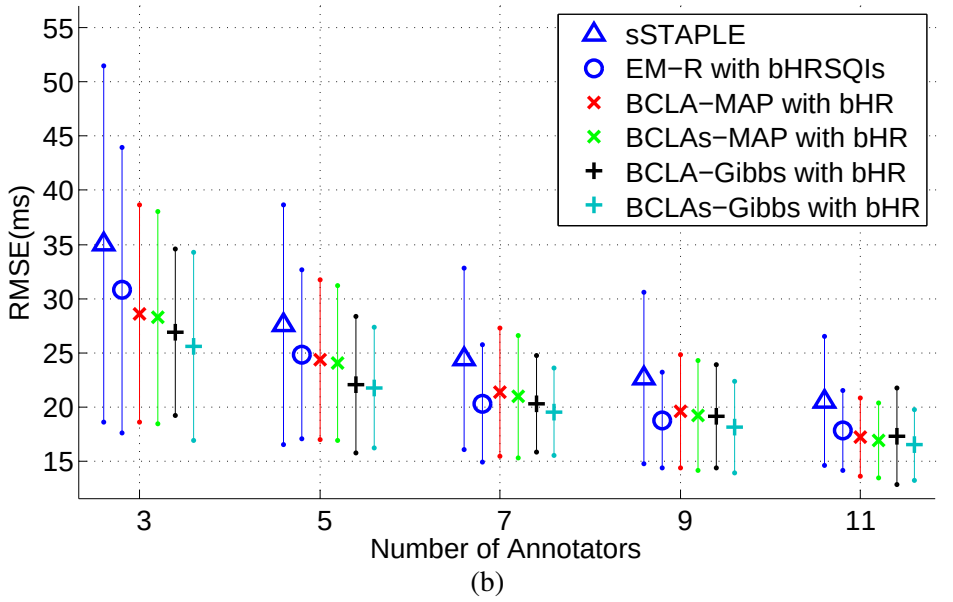
Both the BCLA and BCLAs models surpassed the performance of benchmarking algorithms for all divisions. In particular, BCLAs-Gibbs with the bHR feature has a RMSE improvement of 16.99%, 14.04%, and 19.08% over EM-R with bHRSQIs for Divisions 2, 3, and 4, respectively.

In further assessment of the performance of BCLAs as a function of the number of annotators, a random number of annotators was selected 100 times, as before, as shown in Fig. 7.5 (a). The figure shows that using physiological features such as bHR provides improvement on the estimate of the inferred ground truth (i.e., produced smaller RMSEs), whereas an algorithm like sSTAPLE, which operates with no features, performed worst across different numbers of annotators. Furthermore, the results of BCLAs with the bHR feature consistently outperformed those of BCLA with the bHR feature, particular for small number of annotators, as shown in Fig. 7.5 (b).

For numbers of annotators varied from three to 11, RMSE values are shown in Table 7.3. These numbers of annotators are chosen to provide a realistic comparison that reflects the number of available automated algorithms as would occur in practice. There was no significant difference between BCLA-MAP with the bHR feature versus EM-R with the bHRSQIs features. In comparison, BCLAs-MAP outperformed upon BCLA-MAP slightly with smaller RMSEs, and had a significant reduction in the RMSE when compared to EM-R



(a)



(b)

Fig. 7.5 (a) the mean and standard deviation of the RMSE results as a function of the number of annotators for Division 4 in the PCinC QT dataset; (b) RMSE results when using three to 11 annotators.

Table 7.3 RMSEs (ms) of using three to eleven algorithms for different voting strategies.

Number of Annotators	EM-R with bHRSQIs	BCLA with bHR		BCLAs with bHR	
		MAP	Gibbs	MAP	Gibbs
3	30.79±13.16	28.64±9.99	26.92±7.66	28.28±9.78	25.61±8.68*†
5	24.86±7.79	24.37±7.37	22.06±6.28*	24.09±7.13	21.79±5.56*†
7	20.34±5.42	21.41±5.92	20.31±4.45	20.99±5.63	19.56±4.02
9	18.82±4.42	19.61±5.22	19.18±4.74	19.28±5.06	18.17±4.20
11	17.87±3.71	17.25±3.61	17.33±4.47	16.94±3.47*	16.55±3.28*

Note that the annotations of the selected number of annotators were sampled 100 times from Division 4 provided in the PCinC QT dataset. The * indicates results of using BCLA and BCLAs with the bHR feature were significantly different from EM-R with the bHRSQIs features ($p < 0.05$) using a two-sided Wilcoxon rank-sum test. The † shows results where BCLAs-Gibbs with bHR is significantly different from BCLAs-MAP with bHR using the same test ($p < 0.05$).

when considering 11 annotators. The BCLA-Gibbs model showed significant improvement over EM-R when five annotators were considered. Through incorporating the signal quality extension into BCLA (i.e., BCLAs), we conclude that it provides an additional and significant improvement for BCLAs-Gibbs when three, five, and 11 annotators were considered. Overall, BCLAs-Gibbs with the bHR feature provides the best estimation of the inferred ground truth with the lowest RMSE when compared to other methods. More analysis of the BCLA and BCLAs models with the bHR feature versus no features, as a function of the number of automated annotators, is shown in Appendix C.1.

7.5.2 Capnobase RR Dataset

The results of the RR estimation using the BCLAs models were similar to those using the BCLA model when fusing FFT-based, AR-based, and all (FFT+AR) algorithms. Noting in Appendix C.1 Fig. C.1 (a) that pSQI ≈ 1 for most windows, with 2.47% of windows across 42 subjects having pSQI ≤ 0.9, meaning that there was high agreement between the two peak detectors throughout most windows. As discussed above, when pSQI ≈ 1, the BCLAs model is equivalent to the BCLA model, which explains why no improvement in MAE was

observed. Furthermore, the reference labels were provided from the capnometry waveform, the signal quality measured in the PPG might not be directly comparable to those of the reference signal. The extraction of RRs was performed on the modulation signals, which implied that the pSQI values were not associated with the MAEs of the RR annotations (see Appendix C.1 Fig. C.1 (b)).

7.6 Summary and Future Work

In this study, a signal quality extension was introduced in BCLAs to reduce the error of the inferred ground truth when aggregating annotations. In the PPG application for respiratory rate estimation, the record-specific pSQI was considered as a signal quality metric to be applied to the Capnobase RR dataset. However, as the pSQI values of the dataset were close to one, no improvement was observed. Furthermore, the extraction of RRs was performed on the modulation signals, which might have signal qualities different from those from the PPG. Birrenkott *et al.* [196] have recently proposed respiratory signal quality indices (RQIs) which explicitly modelled the characteristics of the modulation signals. Four RQIs were derived using Fourier transform, autocorrelation, autoregression, and Hjorth complexity. These RQIs could be incorporated in BCLAs in future work to provide more accurate RR estimates.

For the ECG application, results of BCLAs have demonstrated that performance can be improved by incorporating record- and beat-specific signal quality (i.e., bSQI) and heart rate (i.e., bHR) extracted from the ECG independently from the annotations. It was shown that using bSQI in the BCLAs model can provide consistent improvement in the estimation of the inferred ground truth for annotators varied between three to 69 automated algorithms, while it still guaranteed accurate estimation of the precision and bias values of each annotator. Moreover, the nature of the dataset is that the recordings are of high quality with very little noise. The BCLAs method is therefore expected to have an even larger benefit for noisier datasets.

As there are other signal quality indices (SQIs) which were considered in Chapter 4 for ECG signals, future work could combine these SQIs. Note that the SQIs are highly correlated, which could limit traditional approaches to coefficient significance testing for feature selection [127]. Furthermore, as certain SQIs are expected to be useful in only a subset of scenarios (e.g., manual annotators are sensitive to ECG T-wave elevation while automated algorithms are not), the significance of each coefficient could be too variable for reasonable generalisation across datasets. In the case of RQIs proposed by Birrenkott *et al*. [196] for RR estimates, each RQI was evaluated independently and the optimal RQI was dataset-specific. A robust fusion of all RQIs is deemed to provide a better respiratory quality of an underlying modulation signal. A possible solution would be using regularised logistic regression to combine these SQIs/RQIs; i.e., transforming the signal quality metric described here into a linear function of SQIs/RQIs to be incorporated in the BCLAs model. The regression coefficients would then be learned through either sampling the posterior of the likelihood or maximising the log-likelihood in the model. Furthermore, the regularisation can be achieved by adding a zero mean Laplace (i.e., double exponential) prior distribution over the coefficients in the BCLAs model.

When combining aforementioned signal quality metrics for medical data, a simulation study of recordings should be generated with various types of noise, such as motion artefacts, electrode contact noise, and baseline drift. The annotations provided on the simulated recordings would help to identify how each individual annotator varies across the same record given the different noise levels. By selecting a data-specific set of signal quality metrics, it allows BCLAs to better learn the variance of the annotations provided, thereby further improving the estimation of the inferred ground truth.

Chapter 8

BCLA with Correlated Labellers

"When evaluating a model, at least two broad standards are relevant. One is whether the model is consistent with the data. The other is whether the model is consistent with the 'real world.' "

– Kenneth A. Bollen

8.1 Introduction

In this chapter, the framework of BCLA with correlated annotators (BCLAc) is proposed, and our two models of BCLAc are described. The Gibbs sampler and the Metropolis Hastings (MH) algorithm are used by the work described in this chapter to estimate parameters for the BCLAc models (denoted as BCLAc-Gibbs and BCLAc-MH, respectively). We have previously described the Gibbs sampler, and here go on to consider the MH algorithm that can be used when appropriate conjugate priors of the parameters of interest are not readily available.

8.2 A Generative Model of Correlated Annotators

The BCLA model described in Chapter 5, section 5.2.3 assumed that annotators are conditionally independent given the features which defined the ground truth. It does not factor in the possible dependency (or correlation) of errors between individual annotators which might not be the case for automated algorithms. For example, a set of algorithms based on the same analytical method might be expected to yield correlated errors in their respective labels. Incorporating a correlation measure into the annotator model could therefore allow for a better aggregation of the inferred ground truth. Annotators who are considered to be anomalous (i.e., those that are highly correlated to other annotators but which have large variances and biases) should be penalised with lower weighting for their labels; expert annotators (i.e., those that are highly correlated to other annotators but which have small variances and biases) should have their labels weighted more heavily in the model. Two generative frameworks for modelling BCLA with correlated annotators (denoted *BCLAc*) are now described.

8.2.1 The $\mathscr{M}_{\Sigma}$ Model

A multivariate normal distribution[1] can be applied to the annotator model, using the covariance matrix (denoted $\boldsymbol{\Sigma}$) to describe the correlation among annotators, as well as providing constraint to the biases $\boldsymbol{\phi}$. The Inverse-Wishart distribution[2] is used as a prior for the covariance matrix $\boldsymbol{\Sigma}$, and the bias values $\boldsymbol{\phi}$ for all annotators are modelled using a multivariate normal distribution with mean $\boldsymbol{\mu}_{\phi\Sigma}$ and covariance $\boldsymbol{\Sigma}/k_0$. The graphical representation of

[1]The probability density function of the d-dimensional multivariate normal distribution can be defined as $\mathscr{N}(z \mid \boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^d|\boldsymbol{\Sigma}|)^{(-1/2)} exp(-\frac{1}{2}(z-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(z-\boldsymbol{\mu})$, where $\boldsymbol{\mu}$ is 1-by-d and $\boldsymbol{\Sigma}$ is a d-by-d symmetric positive-definite matrix.

[2]The probability density function of a Inverse-Wishart distribution for d-by-d symmetric positive definite matrices X and T, and v as a scalar greater than or equal to d, is given as $\frac{|\mathbf{S}|^{v/2}|X|^{-(v+d+1)/2}\exp(-\frac{1}{2}\text{trace}(\mathbf{S}X^{-1}))}{2^{vd/2}\Gamma_d(v/2)}$, where $\Gamma_d(v/2) = \pi^{d(d-1)/4}\prod_{i=1}^{d}\Gamma\left(\frac{v+1-i}{2}\right)$ is a multivariate gamma function.

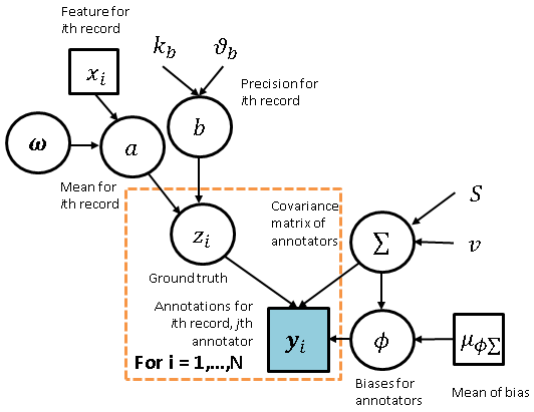


Fig. 8.1 Graphical representation of the BCLAc with $\mathscr{M}_\Sigma$ model: $\mathbf{y}_i$ corresponds to the annotations provided by all annotators for the ith record, and it is modelled by the z_i (the unknown underlying ground truth), the $\boldsymbol{\phi}$ (biases), and the $\boldsymbol{\Sigma}$ (covariance matrix). Furthermore, z_i is drawn from a Gaussian distribution with parameters mean a_i and variance $1/b$, where a_i can be a function of feature vector $\mathbf{x}_i$ and coefficients $\mathbf{w}$. $\boldsymbol{\phi}$ are modelled from a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}_{\phi\Sigma}$ and covariance $\boldsymbol{\Sigma}$ over a scalar factor k_0. The $\boldsymbol{\Sigma}$ is drawn from an Inverse-Wishart (denoted as IW) distribution with parameters v and $\mathbf{S}$. The b is drawn from a Gamma distribution (denoted as Gamma) with parameters k_b and ϑ_b.

BCLA with correlated annotators for modelling covariance (denoted as BCLAc with $\mathscr{M}_\Sigma$) is shown in Fig. 8.1.

As it is illustrated in Fig. 8.1, only the annotator model has been modified in BCLAc when compared to the BCLA model, by introducing the covariance measure among annotators. Assuming that each record is independent, the conditional probability density function of the modified annotator model with covariance becomes the following:

$$p\left(\mathbf{y} \mid z_i, \boldsymbol{\phi}, \boldsymbol{\Sigma}\right) = \prod_{i=1}^{N} \mathscr{N}\left(z_i + \boldsymbol{\phi}, \boldsymbol{\Sigma}\right), \tag{8.1}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the R annotators and there is a total of N recordings.

Matrix $\boldsymbol{\Sigma}$ can be further decomposed into a correlation matrix and the precision values of the annotators. Using the separation strategy proposed by Barnard *et al.* [197], $\boldsymbol{\Sigma}$ is formulated as:

$$\boldsymbol{\Sigma} = \mathbf{Q}\boldsymbol{\rho}\mathbf{Q}, \tag{8.2}$$

where $\mathbf{Q}$ is an R by R diagonal matrix with entries being $\frac{1}{\sqrt{\lambda^{j=1}}}, ..., \frac{1}{\sqrt{\lambda^{j=R}}}$. λ^j is the precision value for the jth annotator, and $\boldsymbol{\rho}$ is the correlation matrix of the annotation errors among R annotators. The $\boldsymbol{\rho}$ matrix measures the Pearson product-moment correlation coefficients [198] within a range of [-1,1]. The correlation coefficient ρ_{pq} of two sets of N annotations (i.e., $\mathbf{y}^p$ and $\mathbf{y}^q$), each provided by annotator p and q, can be written as:

$$\rho_{pq} = \frac{N\sum_{i=1}^{N} y_i^p y_i^q - \sum_{i=1}^{N} y_i^p \sum_{i=1}^{N} y_i^q}{\sqrt{N\sum_{i=1}^{N} y_i^{p2} - \left(\sum_{i=1}^{N} y_i^p\right)^2}\sqrt{N\sum_{i=1}^{N} y_i^{q2} - \left(\sum_{i=1}^{N} y_i^q\right)^2}}. \tag{8.3}$$

$\rho_{pg} = 0$ when both annotators' errors are independent from each other, whereas having ρ_{pq} being in the range of [-1,0) or (0,1] indicates that annotators' errors are negatively or positively correlated in a linear manner, respectively (i.e., their errors decrease or increase together throughout recordings).

The biases of individual annotators are now assumed to be drawn from a multivariate normal distribution constrained by $\boldsymbol{\Sigma}$, with conditional probability density:

$$p\left(\boldsymbol{\phi} \mid \boldsymbol{\mu}_{\phi\Sigma}, \boldsymbol{\Sigma}\right) = \mathcal{N}\left(\boldsymbol{\phi} \mid \boldsymbol{\mu}_{\phi\Sigma}, \boldsymbol{\Sigma}/k_0\right), \tag{8.4}$$

where $\boldsymbol{\mu}_{\phi\Sigma}$ is the prior mean for $\boldsymbol{\phi}$, and k_0 is a positive scalar that expresses our belief on $\boldsymbol{\mu}_{\phi\Sigma}$.

8.2.2 The $\mathscr{M}_\rho$ Model

Although the Inverse-Wishart prior is commonly used for modelling covariance matrices [199], it might be inadequate for the BCLAc model: one may have a different prior belief in the precision matrix $\mathbf{Q}$ over the correlation matrix $\boldsymbol{\rho}$. In particular from equation (8.2), the covariance matrix $\boldsymbol{\Sigma}$ is assumed to be jointly modelled by two independent parameters: the precision values and the correlation of the errors of the annotators. The graphical

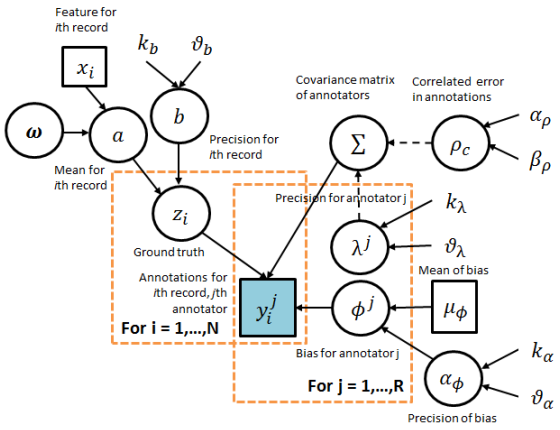


Fig. 8.2 Graphical representation of the BCLAc with $\mathscr{M}_\rho$ model: y_i^j corresponds to the annotation provided by the jth annotator for the ith record, and it is modelled by the z_i (the unknown underlying ground truth), the ϕ^j (bias), and the $\mathbf{\Sigma}$ (covariance matrix). Furthermore, z_i is drawn from a Gaussian distribution with parameters mean a_i and variance $1/b$, where a_i can be a function of feature vector $\mathbf{x}_i$. ϕ^j is modelled from a Gaussian distribution with mean μ_ϕ and variance $1/\alpha_\phi$. $\mathbf{\Sigma}$ is composed of $\boldsymbol{\lambda}$ and $\boldsymbol{\rho}_c$, where $\boldsymbol{\rho}_c$ is a correlation matrix drawn from a Beta distribution with parameters α_ρ and β_ρ. The λ^j is the precision of the jth annotator. The b, λ, and α_ϕ are drawn from a Gamma distribution (denoted as Gamma) with parameters k_b, ϑ_b, k_λ, ϑ_λ, and k_α, ϑ_α respectively.

representation of BCLA with correlated annotators for explicitly modelling their correlation (denoted as BCLAc with $\mathscr{M}_\rho$) is shown in Fig. 8.2.

Assuming that each record is independent, the conditional probability density function of the modified annotator model with correlation becomes the following:

$$p(\mathbf{y} \mid z_i, \boldsymbol{\phi}, \mathbf{Q}, \boldsymbol{\rho}) = \prod_{i=1}^{N} \mathscr{N}(z_i + \boldsymbol{\phi}, \mathbf{Q}\boldsymbol{\rho}\mathbf{Q}). \tag{8.5}$$

Modelling the Prior of the Correlation in the BCLAc with $\mathscr{M}_\rho$ Model

As discussed in the previous section, the correlation of errors and precision values are modelled using two different priors: the Gamma prior is considered for modelling the precision parameter as described in the BCLA model, but an additional Beta distribution[3] with

[3]The probability density function of a Beta distribution for any $a > 0$, $b > 0$, and $\delta \in [0,1]$ is given as $p(\delta \mid a,b) = \frac{\delta^{a-1}(1-\delta)^{b-1}}{B(a,b)}$, where $B(a,b) = \int_0^1 \delta^{a-1}(1-\delta)^{b-1} d\delta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ is the beta function, and a and b are shape parameters.

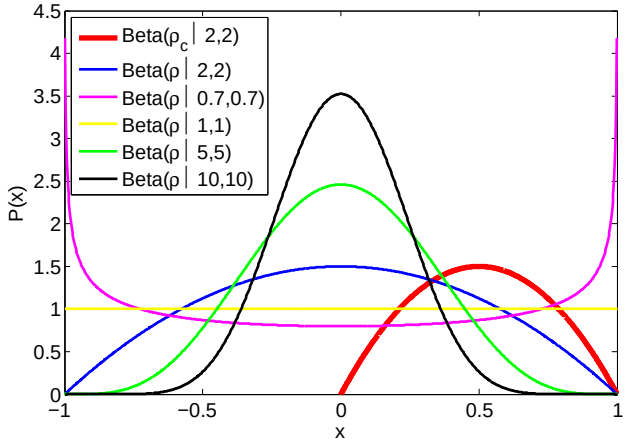


Fig. 8.3 Probability density function of a Beta distribution for modelling the correlation matrix, $\boldsymbol{\rho}$: The original Beta distribution is plotted in red; i.e., Beta($\boldsymbol{\rho}_c \mid \alpha_\rho, \beta_\rho$), and its transformed function is plotted as Beta($\boldsymbol{\rho} \mid \alpha_\rho, \beta_\rho$) for $\alpha_\rho = \beta_\rho = 2$. Other shapes of Beta($\boldsymbol{\rho} \mid \alpha_\rho, \beta_\rho$) are plotted for $\alpha_\rho = \beta_\rho = 0.7$, 1, 5, and 10, respectively.

parameters α_ρ and β_ρ is considered in the BCLAc with $\mathscr{M}_\rho$ model. This Beta distribution acts as the prior for the correlation matrix $\boldsymbol{\rho}$. Furthermore, $\boldsymbol{\rho}_c$ functions as an internal parameter that maps the correlation coefficients from the range [-1, 1] onto the range of [0,1], such that:

$$\boldsymbol{\rho}_c = \frac{\boldsymbol{\rho} + 1}{2}, \tag{8.6}$$

where $\boldsymbol{\rho}_c \sim \text{Beta}(\boldsymbol{\rho}_c \mid \alpha_\rho, \beta_\rho)$. An example of the Beta distribution is shown in Fig. 8.3. The Beta distribution shown in blue is a linear transformation of the original Beta distribution (as shown in red) from the range of [0,1] to [-1,1]. Note that when the parameters of the Beta are set to unity, the Beta distribution becomes uniformly distributed with probability value of one within the range of [-1,1]. When the parameter values are below one, the implication is that annotators' errors in annotation are more likely to be positively and negatively correlated, rather than being independent. If there is no *a priori* knowledge of the performance of the annotators, α_ρ and β_ρ are set to be equal values, and the majority of the annotators' errors in annotation are assumed to be independent of each other (where their correlation

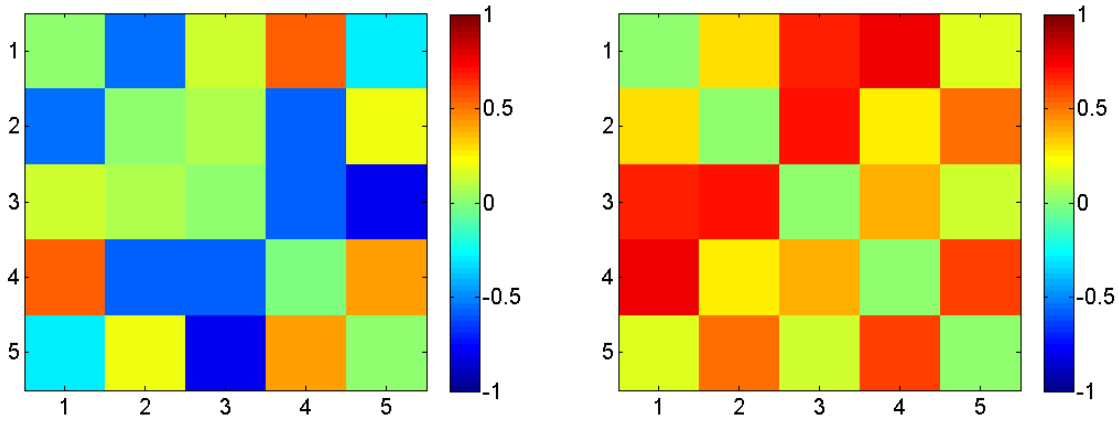
coefficients are mostly zeros, and there is an equal probability of having errors in annotation being positively or negative correlated).

8.2.3 Simulated QT Datasets with Correlated Annotators

Simulation of the BCLAc with $\mathscr{M}_\Sigma$ Model

A total of 1,000 simulated records were generated, each having five annotators with correlated QT annotations. The number of annotators and annotations are selected here for the purpose of demonstration (Application to other datasets using different numbers of annotators and annotations will be described later, in Chapter 8). The simulated dataset considers the following parameter values: the *true* annotation for each record, $z_i \sim \mathscr{N}(400,40)$, a Gaussian distribution with a mean, $a_i = 400$ ms and a standard deviation $b^{-1/2} = 40$ ms. No particular features were considered in this simulation (i.e., $x_i = 1$) as it was solely interested in the performance of the model in estimating correlation, but an intercept was assumed for $f(\mathbf{w},\mathbf{x})$. Furthermore, it was assumed that $\alpha_\phi \sim \text{Gamma}(3,0.0005)$, and the $b \sim \text{Gamma}(3,0.0002)$. The simulated annotations of five annotators for a particular record, $\mathbf{y}_i \sim \mathscr{N}(z_i + \boldsymbol{\phi}, \boldsymbol{\Sigma})$. The $\boldsymbol{\Sigma} \sim \text{IW}(\tau,5)$ for five annotators, where τ is assumed to be a diagonal matrix with diagonal elements being the expected mean of the $\text{Gamma}(4,0.0003)$; their biases, $\boldsymbol{\phi} \sim \mathscr{N}(\boldsymbol{\mu}_{\phi\Sigma}, \boldsymbol{\Sigma}/k_0)$, a Gaussian distribution with 0 ms mean vector and a covariance of $\boldsymbol{\Sigma}/k_0$. The k_0 describes the confidence of the mean $\boldsymbol{\mu}_{\phi\Sigma}$ of the distribution of biases $\boldsymbol{\phi}$ and was set to be 1. The *true* precision values of individual annotators, $\boldsymbol{\lambda}$, can be estimated by finding the inverse of the values in the diagonal elements of $\boldsymbol{\Sigma}$ where each of their correlations is $\rho = 1$. The correlation matrix of the annotation errors are then obtained through decomposing the $\boldsymbol{\Sigma}$, and the correlation of a pair of annotators (e.g., a_1 and a_2) can be estimated as follows:

$$\rho_{a_1,a_2} = \frac{\Sigma_{a_1,a_2}}{1/\sqrt{(\lambda_{a_1}\lambda_{a_2})}}. \tag{8.7}$$



(a) (b)

Fig. 8.4 Simulated correlation of errors of the BCLAc with $\mathscr{M}_\Sigma$ Model: (a) The estimated correlation matrix of errors from five annotators using equation (8.7). (b) The correlation matrix derived directly from the annotations provided by these annotators. Note that each correlated matrix is subtracted from an identity matrix to remove the correlation of an annotator with itself.

The estimated correlation of errors of the BCLAc with $\mathscr{M}_\Sigma$ model for five annotators are shown in Fig. 8.4 (a). It differs from the correlation matrix derived directly from the annotations provided by these annotators (see Fig. 8.4 (b)).

Simulation of the BCLAc with $\mathscr{M}_\rho$ Model

A total of 1,000 simulated records were generated, each having five annotators with correlated QT annotations (more variant of datasets using different number of annotators and annotations will be described later, in Chapter 8). The simulated dataset considers the following parameter values: annotators have precision values, $\boldsymbol{\lambda} \sim \text{Gamma}(4, 0.0003)$; their biases, $\boldsymbol{\phi} \sim \mathscr{N}(0, 25)$, a Gaussian distribution with 0 ms mean and a standard deviation $\alpha_\phi^{-1/2} = 25$ ms. The *true* annotation for each record, $z_i \sim \mathscr{N}(400, 40)$, a Gaussian distribution with a mean, $a_i = 400$ ms and a standard deviation $b^{-1/2} = 40$ ms. No particular features were considered in this simulation (i.e., $x_i = 1$) as it was solely interested in the performance of the model in estimating correlation, but an intercept was assumed for $f(\mathbf{w}, \mathbf{x})$. Furthermore, it was assumed that $\alpha_\phi \sim \text{Gamma}(3, 0.0005)$, and the $b \sim \text{Gamma}(3, 0.0002)$. The simulated

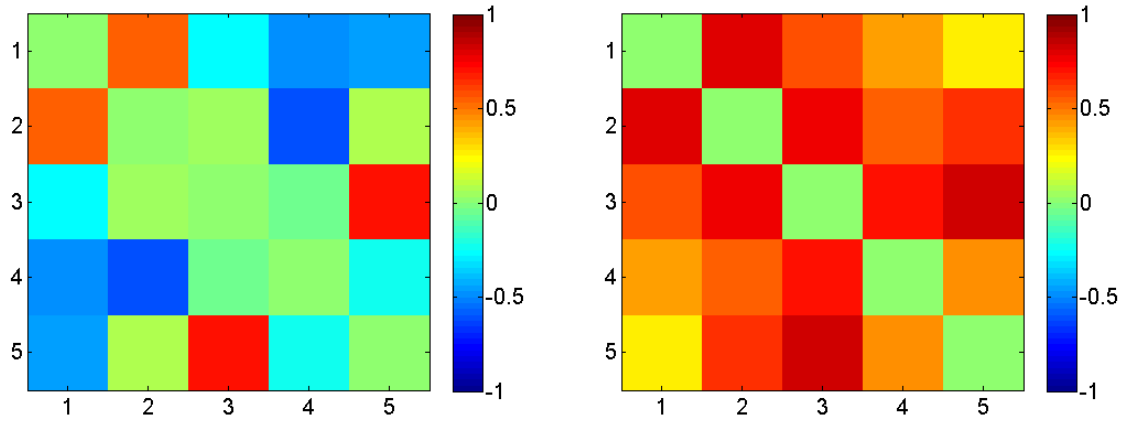


Fig. 8.5 Simulated correlation of errors of the BCLAc with $\mathscr{M}_\rho$ Model: (a) The generated correlation matrix of the errors in annotation from five annotators. (b) The correlation matrix derived directly from the annotations provided by these annotators. Note that each correlated matrix is subtracted from an identity matrix to remove the correlation of an annotator with itself.

annotations of five annotators for a particular record, $\mathbf{y}_i \sim \mathscr{N}(z_i + \boldsymbol{\phi}, \Sigma)$. The $\boldsymbol{\Sigma}$ is calculated from $\mathbf{Q}$ and $\boldsymbol{\rho}$ using equation (8.2), where $\mathbf{Q}$ is composed of the precision values $\boldsymbol{\lambda}$, and $\boldsymbol{\rho}$ was formulated as: assuming the off-diagonal elements in the correlation matrix were drawn from $\text{Beta}(\boldsymbol{\rho}_c \mid 2,2)$, the values in $\boldsymbol{\rho}_c$ were linearly transformed into $\boldsymbol{\rho}$ using equation (8.6); the correlation matrix (see Fig. 8.5(a)) was then created using the improved vine method proposed by Lewandowski *et al.* [200] to ensure its positive semi-definite property.

Correlation of Errors

The correlation coefficients of the generated annotations from two models (see Fig. 8.4 (b) and Fig. 8.5 (b)) were estimated for further comparison: it is interesting to observe that the annotations from all five annotators are positively correlated with $\boldsymbol{\rho} > 0$ as shown by Fig. 8.4 (b) and Fig. 8.5 (b), respectively. This phenomenon can be explained by the fact that the correlation measures the change in trend (i.e., an increase or decrease together, or an increase and decrease inversely) for two set of annotations from a pair of annotators. As long as the annotations both increase similarly, the correlation coefficients would therefore be positive.

However, the *true* correlation of errors among annotators is an estimation of correlation obtained from the difference between the annotations provided and underlying ground truth, as shown in Fig. 8.4 (a) for the $\mathscr{M}_{\Sigma}$ model and Fig. 8.5 (a) for the $\mathscr{M}_{\rho}$ model, respectively.

As annotations with no ground truth are provided in practice, BCLAc will therefore have to infer the *true* annotations in an unsupervised manner as well as predicting the bias and precision of each annotator, and the correlation of errors among them accurately.

8.3 BCLAc with Correlated Annotators (BCLAc)

8.3.1 Gibbs Sampling for BCLAc with $\mathscr{M}_{\Sigma}$ (BCLAc-Gibbs)

The BCLAc with $\mathscr{M}_{\Sigma}$ model was previously described in section 8.2.1, where it considered the modelling of the covariance matrix among annotators directly. The graphical form of the BCLA with $\mathscr{M}_{\Sigma}$ model is shown in Fig. 8.1. The posterior of the parameter $\boldsymbol{\theta}$ for a given dataset $\mathbf{D}$ can be written using Bayes' theorem as:

$$\begin{aligned} & p(\boldsymbol{\theta} \mid \mathbf{D}) \propto p(\mathbf{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}) \\ & = \mathscr{N}\left(\boldsymbol{\phi} \mid \boldsymbol{\mu}_{\phi\Sigma}, \boldsymbol{\Sigma}/k_0\right) \mathrm{IW}(\boldsymbol{\Sigma} \mid v, \mathbf{S}) \times \\ & \mathit{Gamma}(b \mid k_b, \vartheta_b) \left[\prod_{i=1}^{N} \mathscr{N}(z_i \mid a_i, 1/b)\, \mathscr{N}(\mathbf{y_i} \mid z_i + \boldsymbol{\phi}, \boldsymbol{\Sigma})\right]. \end{aligned} \tag{8.8}$$

The Gibbs sampler described in Chapter 6 can be used to model the covariance (i.e., $\boldsymbol{\Sigma}$) directly without modelling the precision and correlation individually. The ground truth can be estimated using the precision values derived from the estimated $\boldsymbol{\Sigma}$. The posterior of biases and covariance can be modelled jointly using a conjugate prior defined by the multivariate normal inverse-wishart distribution (see Appendix B.2 for details). Each parameter in the likelihood described in equation (8.8) can therefore be updated by sampling from its conjugate prior

distribution with posterior hyperparameters (denoted by *) as follows:

$$
\begin{aligned}
& z_i \sim \mathscr{N}\left(z_i \mid a_i^*, \frac{1}{b_i^*}\right). \\
& \boldsymbol{\phi} \sim \mathscr{N}\left(\boldsymbol{\phi} \mid \boldsymbol{\mu}_{\phi\Sigma}^*, \boldsymbol{\Sigma}_{\phi}^*\right). \\
& b \sim Gamma\left(b \mid k_b^*, \vartheta_b^*\right). \\
& \boldsymbol{\Sigma} \sim \mathrm{IW}\left(\boldsymbol{\Sigma} \mid v^*, \mathbf{S}^*\right).
\end{aligned} \tag{8.9}
$$

where

$$
\begin{aligned}
& a_i^* = \frac{\left(\mathbf{x}_i^{\intercal}\mathbf{w}\right) b + \sum_{j\in V_i}\left[\left(y_i^j - \phi^j\right)\lambda^j\right]}{b + \sum_{j\in V_i}\lambda^j}, \quad b_i^* = b + \sum_{j\in V_i}\lambda^j. \\
& \boldsymbol{\mu}_{\phi\Sigma}^* = \frac{k_0\boldsymbol{\mu}_{\phi\Sigma}}{k_0+N} + \frac{\mathbf{U}\bar{\mathbf{y}}_b}{k_0+\mathbf{U}}, \quad \boldsymbol{\Sigma}_{\phi}^* = \frac{\boldsymbol{\Sigma}}{k_0+N}. \\
& k_b^* = k_b + \frac{N}{2}, \quad \frac{1}{\vartheta_b^*} = \frac{\sum_{i=1}^N (z_i - \bar{z})^2}{2} + \frac{1}{\vartheta_b}. \\
& v^* = v + N, \\
& \mathbf{S}^* = \mathbf{S} + \sum_{i=1}^N (\mathbf{y}_i - z_i - \bar{\mathbf{y}}_b)^T (\mathbf{y}_i - z_i - \bar{\mathbf{y}}_b) + \frac{k_0 N}{k_0+N}\left(\bar{\mathbf{y}}_b - \boldsymbol{\mu}_{\phi\Sigma}\right)^T \left(\bar{\mathbf{y}}_b - \boldsymbol{\mu}_{\phi\Sigma}\right).
\end{aligned} \tag{8.10}
$$

Note that V_i is the set of annotators that provided annotations for the ith record. $\mathbf{U}$ is a 1 by R vector, and each of its element indicates the total number of annotations provided by a respective annotator excluding missing annotations. $\bar{\mathbf{y}}_b = [\bar{y}_b^{j=1}, \cdots, \bar{y}_b^{j=R}]$, and $\bar{y}_b^j = \frac{1}{N_j}\sum_{i=1}^N \left(y_i^j - z_i\right)$ is the sample mean difference between the inferred ground truth and annotations across N_j recordings provided by the jth annotator (excluding records with missing annotations). $\boldsymbol{\mu}_{\phi\Sigma}$ is the prior mean for $\boldsymbol{\phi}$, and k_0 defines the belief on this prior mean. $\mathbf{S}$ is proportional to the prior mean for $\boldsymbol{\Sigma}$, and v defines our belief concerning $\mathbf{S}$. v also must satisfy the condition that $v > R-1$. The precision values, $\boldsymbol{\lambda}$ for R annotators, can be estimated as $[\mathrm{diag}\,(\boldsymbol{\Sigma})]^{-1}$. After M iterations of the Gibbs sampler, the expectation of the

value of each parameter can be approximated by calculating the mean after the burn-in period (i.e., $\geq \frac{M}{2}$).

Note that in the $\mathscr{M}_{\Sigma}$ model, having an Inverse-Wishart prior might be an undesirable feature as it acts as a single parameter that controls the covariance matrix $\boldsymbol{\Sigma}$, which subsequently models jointly the precision matrix $\mathbf{Q}$ and the correlation matrix of errors $\boldsymbol{\rho}$. In comparison, BCLAc with $\mathscr{M}_{\rho}$ is a more flexible model which provides different prior belief in $\mathbf{Q}$ over $\boldsymbol{\rho}$ as discussed in section 8.2.2. However, the conjugate prior of the Beta distribution for $\boldsymbol{\rho}$ is not readily available for Gibbs sampling, thereby an alternative MCMC approach such as the MH algorithm is proposed to compute the parameters described in the BCLAc with $\mathscr{M}_{\rho}$ model.

8.3.2 Background on Metropolis Hastings Algorithm

In previous chapters, Gibbs sampling was considered for inference with the BCLA model with independent annotators. Although it is possible to describe mathematically the posterior of the BCLAc model, it is difficult to derive the conditional distribution for the correlation coefficients in the model. Furthermore, we have shown that there was a strong correlation between the inferred ground truth and annotators' biases, which might result in slow mixing for the Gibbs sampler (see discussion in Chapter 6, section 6.5.1). An alternative method such as the Metropolis Hastings (MH) algorithm [201], a commonly-used MCMC method, can be considered to draw samples from the full joint posterior distribution of all the parameters, as well as from distribution for each parameter of interest.

Suppose we have a target distribution over parameters, $\boldsymbol{\theta}$, and given the observation dataset $\mathbf{D}$, the posterior of $\boldsymbol{\theta}$ can be inferred using Bayes' theorem as follows:

$$p(\boldsymbol{\theta} \mid \mathbf{D}) = \frac{p(\mathbf{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int p(\mathbf{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}} = \frac{1}{Z} p(\mathbf{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}), \tag{8.11}$$

where Z represents a constant term (sometimes called the "partition function") which is independent of $\boldsymbol{\theta}$. The MH algorithm allows us to draw samples from the full posterior of $\boldsymbol{\theta}$ up to some constant Z. The MH algorithm involves generating a draw $\boldsymbol{\theta}^*$ from a proposal distribution $Q(\boldsymbol{\theta})$. A rejection/acceptance criterion is then applied, such that the draw is accepted according to the posterior target distribution.

Pseudo code for the MH algorithm is shown in Algorithm 2, where the proposal distribution $Q(\boldsymbol{\theta})$ can take different forms and is user-specified. When the proposal distribution is symmetric, the term $\mathrm{Q}\left(\boldsymbol{\theta}^h \mid \boldsymbol{\theta}^*\right)$ is equivalent to the $\mathrm{Q}\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^h\right)$ (i.e., they have equal probability densities), and the MH algorithm is then equivalent to the Metropolis algorithm [202]. In the case where the proposal distribution is asymmetric, such as with the log-normal or Gamma distributions, the ratio of $\mathrm{Q}\left(\boldsymbol{\theta}^h \mid \boldsymbol{\theta}^*\right) / \mathrm{Q}\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^h\right)$ acts as a "correction factor" to alter the acceptance probability. In fact, the Gibbs sampler is a special case of the MH algorithm where the proposal distributions are the posterior conditionals, yielding the acceptance probability (denoted as α) of being one.

Algorithm 2 The Metropolis-Hastings Algorithm

$\boldsymbol{\theta}^0 \leftarrow \mathbf{x}$ ▷ $\mathbf{x}$ is an initialisation vector
for each sequence $h = 0$ to M **do**
  Draw $\boldsymbol{\theta}^* \sim \mathrm{Q}\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^h\right)$
  Draw $u \sim U(0,1)$
  if $u < min\left\{1, \alpha = \frac{p(\boldsymbol{\theta}^*|\mathbf{D})\mathrm{Q}(\boldsymbol{\theta}^h|\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^h|\mathbf{D})\mathrm{Q}(\boldsymbol{\theta}^*|\boldsymbol{\theta}^h)}\right\}$ **then**
    Set $\boldsymbol{\theta}^{h+1} \leftarrow \boldsymbol{\theta}^*$
  else
    Set $\boldsymbol{\theta}^{h+1} \leftarrow \boldsymbol{\theta}^h$
  end if
end for

Note that the acceptance probability α is a ratio, and hence the constant Z values cancel. This is advantageous, because we need not consider the value of Z, which is often not straightforward to compute. $\boldsymbol{\theta}$ can also be further partitioned into L parameters $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_L\}$, and each parameter can be updated using an individual proposal distribution.

The acceptance probability of each parameter can either be evaluated together as one entire $\boldsymbol{\theta}$ or in a sequence, as is similar to the Gibbs sampler as detailed in Chapter 6.

The Proposal Distribution

The proposal distribution of each parameter of interest serves an an important factor for influencing the efficiency of the MH algorithm, irrespective of the asymmetry of that proposal distribution [176]. If a Gaussian proposal is assumed for $\mathrm{Q}\left(\boldsymbol{\theta}^{*} \mid \boldsymbol{\theta}^{h}\right)$ in Algorithm 2, its variance parameter σ_v^2 would need to be tuned during the burn-in period, guided by the acceptance rate. The acceptance rate is defined as the fraction of the proposed samples that is accepted of the most recent M_s samples (where $M_s < M$ in Algorithm 2), and its desired value depends on the dimension of the target distribution. Roberts *et al.* [203] have shown that the ideal acceptance rate for a one-dimensional Gaussian distribution is approximately 45%, but can reduce to 23% as the dimensionality approaches infinity.

When a proposal distribution takes small updating steps with small variance (i.e., the proposed draw is similar to the value from the previous iteration), then the acceptance probability α is high, yielding a higher acceptance rate for the proposed samples. However, this leads to slow mixing - the Markov chain might be stuck in the local parameter space, producing highly correlated draws, and will take longer to reach its stationary distribution. On the contrary, if the proposal distribution generates large steps (i.e., has large variance) which corresponds to taking big jumps in the parameter space, then the acceptance probability α is low, with a subsequently low acceptance rate, which again results in slow convergence.

The same convergence diagnostic tests described in Chapter 6, section 6.2.2 for the Gibbs sampler can be applied to the MH algorithm to measure its rate of convergence; i.e., (i) autocorrelation; (ii) the Geweke diagnostic; (iii) the Raftery-Lewis diagnostic; and (iv) the Gelman-Rubin diagnostic.

8.3.3 The Metropolis Hastings Algorithm for BCLAc with $\mathscr{M}_\rho$ (BCLAc-MH)

The BCLAc with $\mathscr{M}_\rho$ model previously proposed in section 8.2.2 (see Fig. 8.2) provides a direct modelling of the correlation of errors, and it is separated from the modelling of the precision values of individual annotators. The likelihood of the parameter $\boldsymbol{\theta}$ with correlation extension for a given dataset $\mathbf{D}$ can be written using Bayes' theorem as:

$$\begin{aligned} p(\boldsymbol{\theta} \mid \mathbf{D}) &\propto p(\mathbf{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}) \\ &= \text{Gamma}\left(\alpha_\phi \mid k_\alpha, \vartheta_\alpha\right) \text{Gamma}\left(b \mid k_b, \vartheta_b\right) \text{Beta}\left(\boldsymbol{\rho}_c \mid \alpha_\rho, \beta_\rho\right) \times \\ &\quad \left[\prod_{j=1}^{R} \mathscr{N}\left(\phi^j \mid \mu_\phi, 1/\alpha_\phi\right) \text{Gamma}\left(\lambda^j \mid k_\lambda, \vartheta_\lambda\right)\right] \times \\ &\quad \left[\prod_{i=1}^{N} \mathscr{N}\left(z_i \mid a_i, 1/b\right) \mathscr{N}\left(\mathbf{y_i} \mid z_i + \boldsymbol{\phi}, \mathbf{Q}\boldsymbol{\rho}\mathbf{Q}\right)\right]. \end{aligned} \tag{8.12}$$

As the Beta distribution does not yield a closed form expression for the posterior, the MH algorithm serves as an alternative to the Gibbs Sampler to compute the parameters in BCLAc with $\mathscr{M}_\rho$. The acceptance rate of BCLAc-MH was set to be within [20%, 40%] following [203] as the model is in a hierarchical structure, and has individual parameters, each with different dimensions, thereby a range of acceptance rates was considered to cater for variable dimensions of the proposal distribution.

The Proposal Distribution for BCLAc-MH

As shown in line 3 of Algorithm 2, the new sample of a parameter θ^* is drawn from a proposal distribution given the current value of θ in the hth sequence of the M iterations, $\text{Q}\left(\theta^* \mid \theta^h\right)$. Gaussian proposal distributions were considered for updating the inferred ground truth for N recordings $\mathbf{z}$, the bias values for the R annotators $\boldsymbol{\phi}$, the mean value of the ground truth a, and the correlation matrix of errors $\boldsymbol{\rho}$. Since $\boldsymbol{\rho}$ is symmetrical along its main diagonal

(where it has values of one) with identical elements in [-1,1] off the diagonal, only half of the off-diagonal elements were updated at each proposal. Note that each correlation coefficient (i.e., ρ_o) in the correlation matrix was assumed to be in a range of [-∞,+∞] during the update step, then it was transformed to the correct range of [-1,1] (denoted as ρ_n) when estimating the acceptance probability (i.e., α in line 5 of the Algorithm 2) using the following equation:

$$\rho_n = \frac{1-\exp(-\rho_o)}{1+\exp(-\rho_o)}. \tag{8.13}$$

Log-normal[4] proposal distributions were considered for the precision values of the R annotators $\boldsymbol{\lambda}$, the variance of the ground truth $1/b$, and the variance of the bias values for the R annotators $1/\alpha_\phi$. Note that the log-normal distribution is asymmetric and ranges from zero to +∞, thereby a correction factor for each parameter needed to be calculated when estimating the acceptance probability. Alternatively, one can transform the parameter from log-normal to normal scale, and perform the update of θ in the normal space using a Gaussian proposal.

Tuning the Variance Value in a Proposal Distribution for BCLAc-MH

In order to make sure that the proposal update of each parameter falls within the defined ranges of the acceptance rate, the aforementioned variance of the proposal distribution needs to be tuned during the burn-in period. A simple way of updating the variance in a Gaussian or log-normal proposal is to make it as a multiple of the step length, k. The tuning process is as follows:

1. Fix all parameter (i.e., $\mathbf{z}$, $\boldsymbol{\phi}$, a, $\boldsymbol{\rho}$, $\boldsymbol{\lambda}$, b, and α_ϕ) values but one.
2. Initialise the variance σ_v^2 of its proposal distribution to be a function of ck, where c is a scalar acting as a multiplication factor and which is initialised arbitrarily. The step

[4] A univariate log-normal distribution can be defined as $\log \mathcal{N}\left(x \mid \mu, \sigma^2\right) = x\left(2\pi\sigma^2\right)^{(-1/2)} \exp\left(-\left(\log x-\mu\right)^2/2\sigma^2\right)$, where μ and σ correspond to the location and scale parameters. The mean m and variance v of a log-normal distribution are functions of μ and σ, and can be estimated as $m = \exp\left(\mu+\sigma^2/2\right)$ and $v = \exp\left(2\mu+\sigma^2\right)\left(\exp\left(\sigma^2\right)-1\right)$. Conversely, $\mu = \log\left(m^2/\sqrt{v+m^2}\right)$ and $\sigma = \sqrt{\log\left(v/m^2+1\right)}$.

length k is initialised to be $k = 0.25$. The updated parameter θ^* is defined[5] as:

$$\theta^* = \mathcal{N}(\theta, ck) \quad or \quad \theta^* = \theta \log \mathcal{N}(0, ck).$$

3. Run the MH algorithm for 20,000 iterations or until the posterior of the parameter value has converged: for every 1,000 iterations, the acceptance rate r is estimated as the percentage of the proposal that is accepted. If $r < 20\%$, reduce the value of k by setting $k' = k/1.1$, or if $r > 40\%$, increase k by setting $k' = 1.1k$.
4. Compare the value of k after convergence k_{conv} with its initialised value: a large k_{conv} value ($k_{conv} >> 10k$) indicates that c could probably be increased, and vice versa. Essentially c was set to maximise the convergence rate for a given value of r.
5. Repeat steps 1 to 4 for the remaining parameters. The variance of a proposal distribution for an individual parameter is therefore set in a way that all parameters would have similar acceptance rates.

Once the variance values are defined, the MH algorithm will run for M iterations, and the expectation of the value of each parameter can be approximated by calculating the mean after the burn-in period (i.e., $h \geq \frac{M}{2}$).

Dealing with Incomplete Data

In the case where the annotations are missing, this does not affect the update step in BCLAc-MH. However, when calculating the probability densities $p(\boldsymbol{\theta}^* \mid \mathbf{D})$ or $p\left(\boldsymbol{\theta}^h \mid \mathbf{D}\right)$ to compute the acceptance probability α, special care is required for dealing with missing annotations: the recordings are associated with multivariate normal densities with different means and covariance matrices that describe the annotations provided by annotators. Instead of estimating the density per recording which is computationally intensive as N becomes large, the annotations for all records may be partitioned into multiple clusters according to the patterns of the missing labels provided by different annotators. The probability densities for multiple

[5] N.B.: $\theta^* = \theta \log \mathcal{N}(0, ck) = \exp \mathcal{N}(\log(\theta), ck)$.

recordings within the same cluster (i.e., those with the same pattern of missing labels) can then be estimated simultaneously in a more efficient manner.

8.4 Data Description and Methodology of Comparison

8.4.1 Datasets

Simulated Datasets

Simulated datasets of 100, 500, and 1000 records were generated with parameter values described in section 8.2.3, each with (i) five and (ii) 10 correlated QT annotations from corresponding annotators. These were generated from both BCLAc with $\mathscr{M}_{\Sigma}$ and with $\mathscr{M}_{\rho}$, forming a total of $3 \times 2 \times 2 = 12$ simulated datasets.

Real Datasets

There are two datasets used to validate the proposed BCLAc models, namely the Capnobase RR dataset, and a subset of the PCinC QT dataset (denoted as PCinCnew). For PCinCnew, only those annotators that had provided more than 50% of annotations were considered, and thereby resulting 40 annotators in Division 2, 15 annotators in Division 3, and 55 annotators in Division 4.

8.4.2 Methodology of Comparison

Simulated QT Datasets

To test the robustness of the BCLAc-Gibbs and BCLAc-MH approaches, each was applied to the 12 simulated QT datasets generated from the BCLAc with $\mathscr{M}_{\Sigma}$ as well as with $\mathscr{M}_{\rho}$ models. BCLAc-Gibbs and BCLAc-MH were evaluated in the same manner as the BCLA model, which was described in Chapter 6.

Real Datasets

The BCLAc models were evaluated in the same manner as the BCLA model for the Capnobase RR dataset, which was described in Chapter 6. Noting that the errors of three annotators (i.e., three sets of 150 RR estimates) might be correlated as their annotations were derived from the same methodology (see Chapter 3), the BCLAc models were explored to uncover such relationship to further improve the inferred RR estimates. The BCLAc-Gibbs approach was considered for the PCinCnew dataset, and was evaluated in the same way as BCLA as described in Chapter 6.

8.5 Results

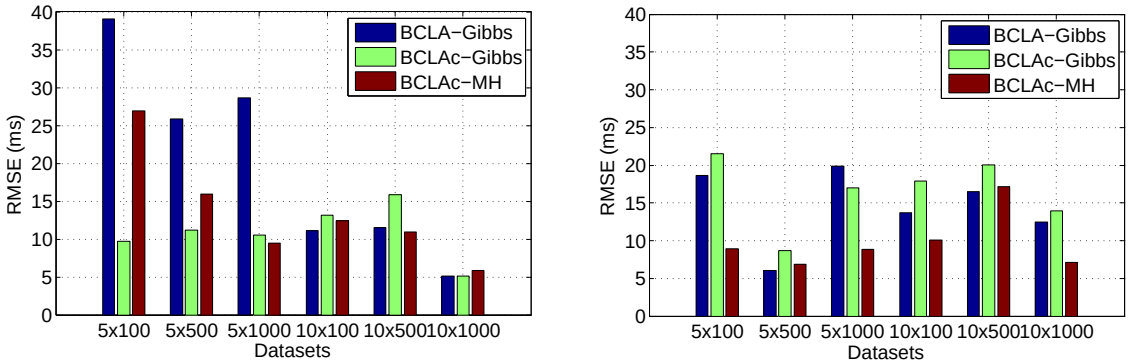
8.5.1 Simulated Datasets

BCLA-Gibbs took about 350 seconds for running 10,000 draws of the 2,500 annotations created from five annotators for 500 recordings using the same system as before. In comparison, BCLAc-Gibbs took approximately 404 seconds to run the same number of draws. BCLAc-MH was allowed to run for 35 hours, resulting in 20,000,000 iterations, where the parameter values were stored every 1,000 iterations; this process resulted in a total of 20,000 samples of parameter values. Half of the samples were discarded as burn-in, as before.

Simulated Datasets with Five Annotators

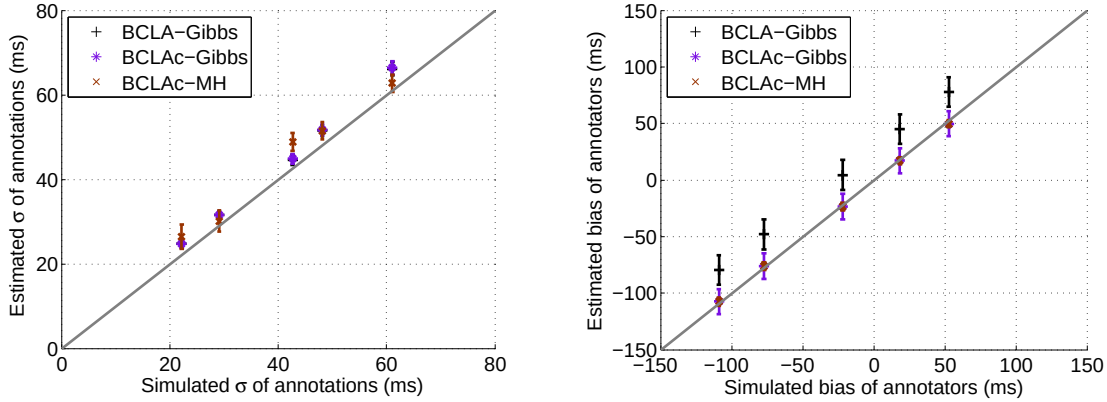
When five annotators were considered for generating different datasets with 100, 500, and 1,000 annotations respectively, RMSE results (see Fig. 8.6) varied depending on how the correlation of errors among annotators was generated.

Datasets from BCLAc with $\mathscr{M}_\Sigma$ For datasets generated from BCLAc with $\mathscr{M}_\Sigma$, we assumed that the uncertainty of the bias values $\boldsymbol{\phi}$ was controlled by the covariance matrix $\boldsymbol{\Sigma}$. In

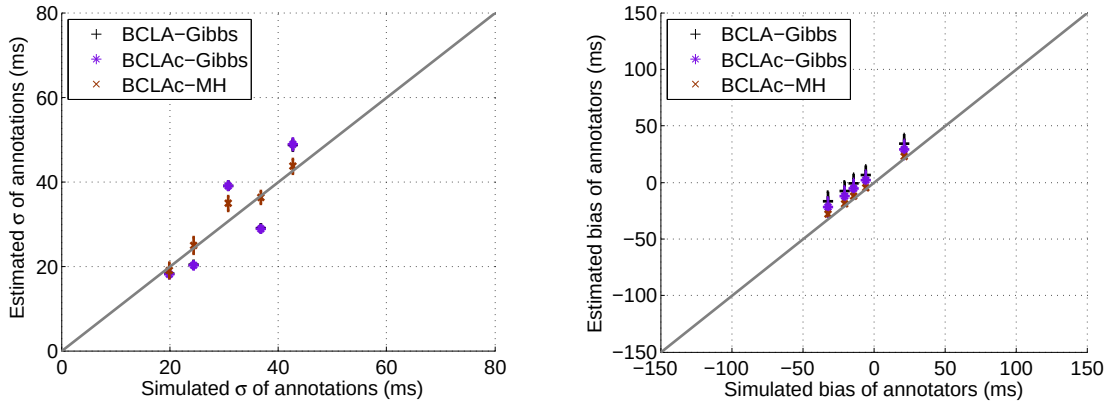


(a) The $\mathscr{M}_\Sigma$ model (b) The $\mathscr{M}_\rho$ model

Fig. 8.6 Bar plots of RMSE using different strategies for simulated datasets; for (a) BCLAc with $\mathscr{M}_\Sigma$ and (b) BCLAc with $\mathscr{M}_\rho$. In each dataset, the number of annotators and number of annotations are indicated on the left and right of "$\times$" respectively.

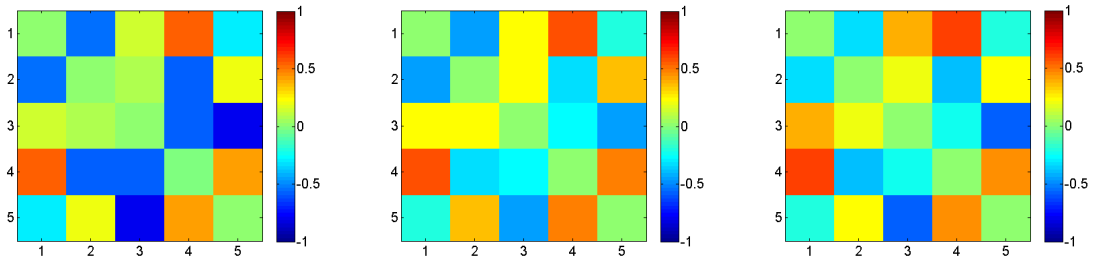


(a) the $\mathscr{M}_\Sigma$ model: 1000 annotations $\times$ 5 annotators

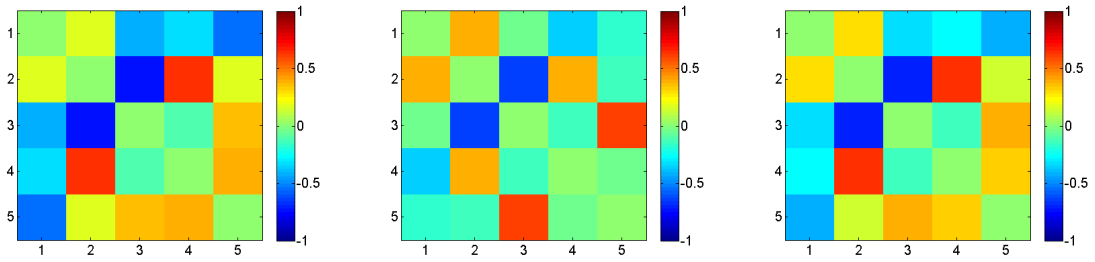


(b) the $\mathscr{M}_\rho$ model: 1000 annotations $\times$ 5 annotators

Fig. 8.7 A comparison of the simulated and inferred σ and bias values of five annotators in the simulated datasets generated from (a) the $\mathscr{M}_\Sigma$ model and (b) the $\mathscr{M}_\rho$ model. The diagonal line indicates a perfect match between the simulated and estimated results. The results are plotted as with one standard deviation from the mean.



(a) The $\mathscr{M}_{\Sigma}$ model: 1000 annotations $\times$ 5 annotators



(b) The $\mathscr{M}_{\rho}$ model: 1000 annotations $\times$ 5 annotators

Fig. 8.8 A comparison of the simulated and inferred correlation of errors among five annotators in the simulated datasets generated from (a) the $\mathscr{M}_{\Sigma}$ model, and (b) the $\mathscr{M}_{\rho}$ model. In each row, the *true* $\boldsymbol{\rho}$ (left), is compared with its inferred values using BCLAc-Gibbs (middle), and BCLAc-MH (right). Note that each correlated matrix is subtracted from an identity matrix to remove the correlation of an annotator with itself.

comparison, the BCLA-Gibbs and BCLAc-MH approaches assume the bias values $\boldsymbol{\phi}$ were modelled independently from the covariance matrix $\boldsymbol{\Sigma}$. It was therefore predicted that both approaches would provide less reliable estimation of the true bias values $\boldsymbol{\phi}$.

The BCLA-Gibbs approach performed consistently worse than the other two approaches as it over-estimated the bias values $\boldsymbol{\phi}$ of the annotators constantly (see Fig. 8.7 (a)). The BCLAc-MH approach, on the other hand, showed improvement over BCLA-Gibbs in its bias prediction (see Fig. 8.7 (a)), because it incorporates an additional modelling of the correlation of errors $\boldsymbol{\rho}$ among annotators (see Fig. 8.8 (a)). Such correlation of errors contributed to the covariance matrix $\boldsymbol{\Sigma}$ that improved on the estimation of the underlying ground truth.

Overall, BCLAc-Gibbs was able to provide accurate estimation of the correlated errors (see Fig. 8.8 (a)), the inferred ground truth (see Fig. 8.6) (a)), and the precision and bias values of annotators (see Fig. 8.7 (a)). Both the BCLAc-Gibbs and the BCLAc-MH approaches produced smaller RMSEs for the three datasets when compared to those produced by BCLA-

Gibbs (see Fig. 8.6 (a) for datasets with five annotators). In particular, the performance of BCLAc-Gibbs was superior over BCLAc-MH when fewer number of annotations are available.

Datasets from BCLAc with $\mathscr{M}_\rho$ The BCLAc-Gibbs approach was less robust in performance as it produced less reliable estimation of the correlation of errors $\boldsymbol{\rho}$ (see Fig. 8.8 (b)). This has subsequently affected the estimation of the inferred precision values $\boldsymbol{\lambda}$ (see Fig. 8.7 (b)), as BCLAc-Gibbs assumed both the precision values $\boldsymbol{\lambda}$ and the correlation of errors $\boldsymbol{\rho}$ were directly derived from the same covariance matrix $\boldsymbol{\Sigma}$. Furthermore, BCLAc-Gibbs assumed the uncertainty of the bias values $\boldsymbol{\phi}$ were proportional to the covariance $\boldsymbol{\Sigma}$, therefore their inferred bias values were also affected (see Fig. 8.8 (b)). In comparison of RMSE (see Fig. 8.6 (b)), the results of BCLA-Gibbs varied depending on the simulated correlation of errors. As BCLAc with $\mathscr{M}_\rho$ assumed the precision and bias values of annotators were modelled independently and separately from the correlated errors of annotators $\boldsymbol{\rho}$, the BCLA-Gibbs approach (which assumes independent $\boldsymbol{\lambda}$ and $\boldsymbol{\phi}$) could sometimes produce smaller RMSE of the ground truth (see results of dataset 5×500 in Fig. 8.6 (b)) over the others.

It was evidenced that BCLAc-MH was more robust over the other two approaches as it provided accurate estimation of the correlated errors (see Fig. 8.8 (b)), reliable estimation of the ground truth (see Fig. 8.6) (b)), as well as precision and bias values of annotators (see Fig 8.7 (b)).

Simulated Datasets with 10 Annotators

The results of inferred bias and precision values for 10 annotators were similar to those obtained for five annotators (see Appendix C.2 for details). Further examining the overall RMSEs of the inferred ground truth as shown in Fig. 8.6, the BCLA-Gibbs approach was sufficient to provide reliable estimation without introducing the correlation of errors among annotators when annotator number was increased to 10.

Table 8.1 The parameters of BCLAc for modelling the PCinCnew QT dataset.

Symbol	Definition	Value
k_b	shape of Gamma distribution for b	3 ‡
ϑ_b	scale of Gamma distribution for b	0.0002 ‡
$\mathbf{S}$	the scale matrix of Inverse-Wishart distribution for $\mathbf{\Sigma}$	$\frac{1}{\lambda_M} I$
v	the degrees of freedom of Inverse-Wishart distribution for $\mathbf{\Sigma}$	R
k_0	a scale factor of $\mathbf{\Sigma}$	10
$\boldsymbol{\mu}_{\phi\Sigma}$	mean vector of the bias distribution	10 †

Note: b is the precision parameter for the model of the ground truth. $\mathbf{\Sigma}$ is the covariance matrix. $\frac{1}{\lambda_M} I$ is an R $\times$ R diagonal matrix with entries obtained from a scalar, λ_M, as the inverse of the mean of the Gamma$(4, 0.003)$, with the assumption that the annotations provided by the best performing algorithm is ± 5 ms away from the reference. $\boldsymbol{\lambda}$ refers to annotators' precision values. The values with † are derived from [159–161]. The values with ‡ are derived from [12, 162, 163].

BCLAc-MH seemed to be more robust in providing reliable estimation for datasets generated from both the $\mathscr{M}_\Sigma$ and $\mathscr{M}_\rho$ models, except when datasets were generated from the $\mathscr{M}_\Sigma$ model with fewer number of annotations provided by a smaller number of annotators (see RMSE results in Fig. 8.6 (a) for datasets 5×100 and 5×500).

8.5.2 QT Dataset

BCLA-Gibbs took about 152 seconds for running 5,000 draws of the 548 annotations, each provided from five algorithms using the same system as before. With the same dataset, BCLAc-Gibbs took approximately 70 seconds to run 2,000 draws for convergence. When fusing large number of algorithms (e.g., 55 with 26,922 annotations), BCLA-Gibbs required 5,000 draws and took about 500 seconds, whereas BCLAc-Gibbs ran for 290 seconds to compute 3,000 draws. Both methods discarded the fist half of the samples as burn-in, as before.

The resulting values of the parameters and hyperparameters of the BCLA model for the QT dataset were described previously in Chapter 5, Table 5.1. They were also considered in the BCLAc models with additional parameters as shown in Table 8.1.

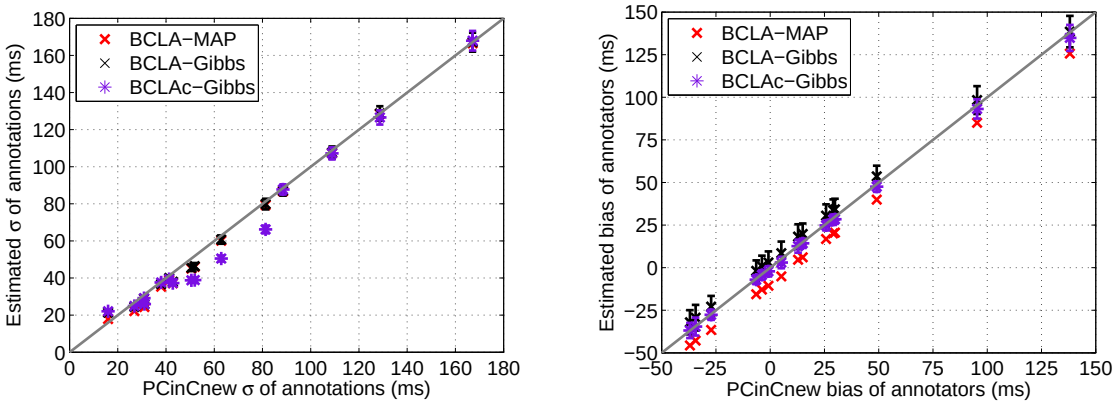


Fig. 8.9 A comparison of the PCinCnew QT reference and inferred σ and bias of each annotator for Division 3. The results are plotted as with one standard deviation from the mean (as '+').

The inferred precision and bias results estimated using BCLA-MAP, BCLA-Gibbs, and BCLAc-Gibbs are compared for the PCinCnew QT dataset. Both the BCLA and BCLAc methods produced accurate estimation of the bias values for Division 2 and 4. In the case of Division 3 (see Fig. 8.9), BCLAc-Gibbs produced more accurate estimation of the bias values in comparison to those computed using BCLA-Gibbs and BCLA-MAP. However for σ prediction, the BCLA methods were more reliable than BCLAc-Gibbs. This might be due to the fact that both the correlation of errors and the precision values were jointly modelled as a whole using one distribution (i.e., the IW distribution) in the BCLAc-Gibbs model, limiting its accuracy as one small imperfect estimation of the correlation of errors would affect the estimation of the precision values directly.

An example of the inferred correlation of errors using BCLAc-Gibbs is shown in Fig. 8.10 for 55 algorithms. Although the exact coefficient values were not fully recovered, BCLAc-Gibbs was able to identify the key relationship of errors in annotation among annotators, while a direct estimation of the correlation of annotations fail to do so. In comparison to the estimation of the ground truth for all divisions, the RMSE results are given in Table 8.2, where BCLAc-Gibbs produced the smallest RMSEs as compared to all other BCLA-based methods, in addition to outperforming all other voting strategies.

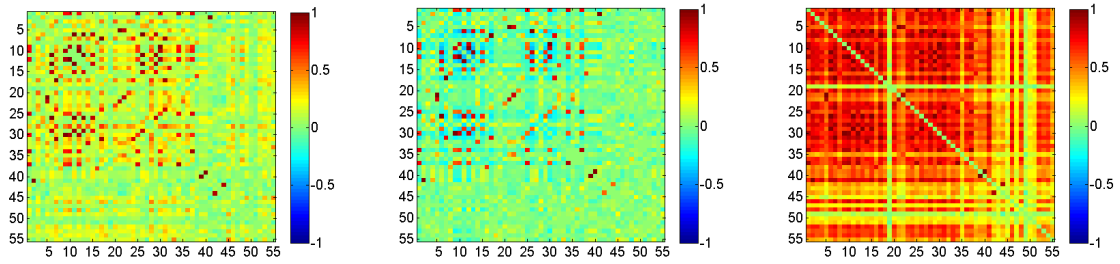


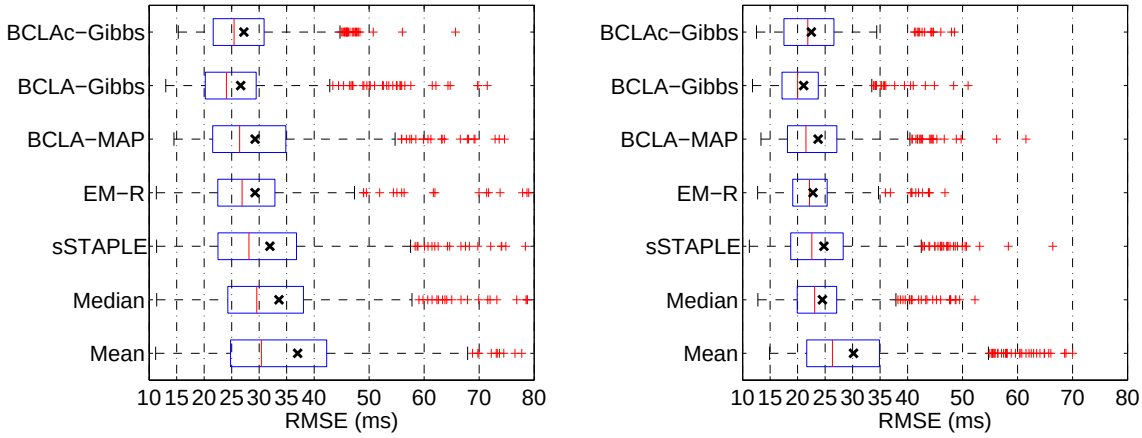
Fig. 8.10 A comparison of the reference and inferred correlation of errors among among 55 algorithms for the Division 4 in the PCinCnew QT dataset: In each row, the reference $\boldsymbol{\rho}$ (left), is compared with its inferred values using BCLAc-Gibbs (middle), and the correlation estimated directly from the annotations provided (right). Note that each correlated matrix is subtracted from an identity matrix to remove the correlation of an algorithm with itself.

Table 8.2 RMSEs of the inferred labels using different strategies in the PCinCnew QT dataset.

Division (number)	Mean	Median	EM-R	sSTAPLE	BCLA-MAP	BCLA-Gibbs	BCLAc-Gibbs
2(40)	16.38	15.52	15.40	15.16	12.71	12.06	11.95
3(15)	32.14	20.55	24.64	20.54	17.41	16.20	16.16
4(55)	18.28	15.10	16.85	15.13	12.65	11.66	11.66

Fig. 8.11 shows a further evaluation of the accuracies (in terms of RMSE) of different voting strategies as a function of the number of annotators. The results were generated by sub-sampling annotators 500 times with three to nine annotators selected from Division 4. Similar results were obtained as those described in Chapter 6: BCLA-Gibbs outperformed all other voting strategies with least mean or median RMSE of 500 samples for selecting annotator varied from three to nine (RMSE of 26.64 ± 10.32 ms, 21.10 ± 5.64 ms, 19.26 ± 4.83 ms, 18.34 ± 4.03 ms for selecting three, five, seven, and nine annotators, respectively). The RMSE results of BCLA-Gibbs were significant smaller than others when performing a two-sided Wilcoxon rank-sum test ($p < 0.01$), with an exception of comparing to the BCLAc-Gibbs approach while selecting nine annotators.

Inspecting the performance of BCLAc-Gibbs, it produced smaller RMSEs as compared to the BCLA-MAP approach when number of annotators was less than nine: 27.22 ± 8.34 ms versus 29.26 ± 11.41 ms, 22.49 ± 6.28 ms versus 23.73 ± 7.67 ms, 20.77 ± 6.15 ms versus 21.24 ± 6.26 ms, and 19.58 ± 5.87 ms versus 19.15 ± 4.40 ms for three, five, seven, and nine



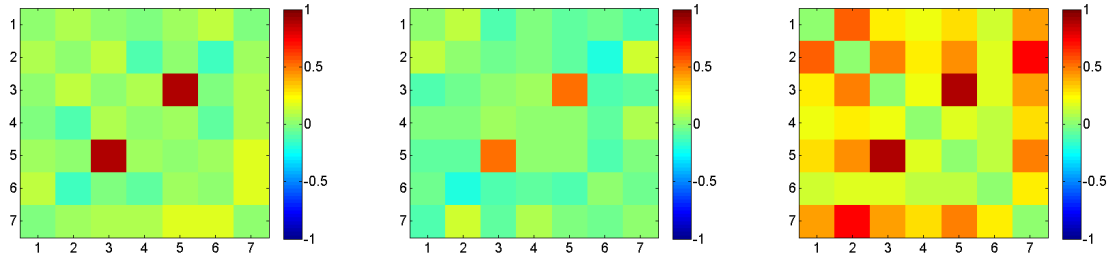
(a) Selection of three annotators

(b) Selection of five annotators

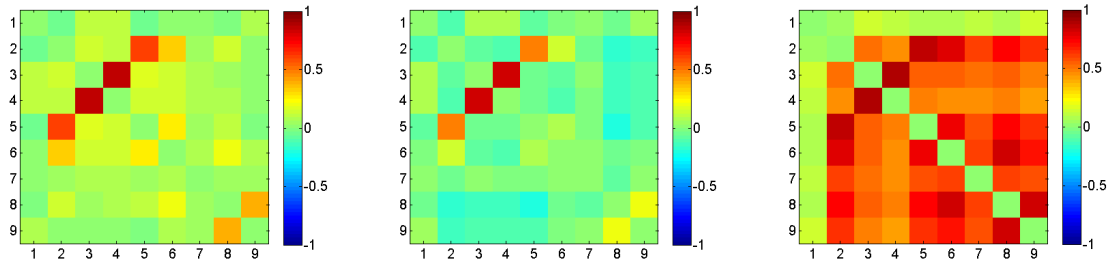
(c) Selection of seven annotators

(d) Selection of nine annotators

Fig. 8.11 Box plots of the RMSE results of using different voting approaches are shown as a function of the number of automated annotators. Each plot was generated by randomly sampling the annotators 500 times. The median RMSE is indicated in red '-' within each box, while the black '×' indicates the mean RMSE. The span of each box represents the interquartile range.



(a) Selection of seven annotators.



(b) Selection of nine annotators

Fig. 8.12 A comparison of the reference and inferred correlation of errors among seven (a) and nine (b) automated annotators in the PCinCnew QT datasets: In each row, the reference $\boldsymbol{\rho}$ (left), is compared with its inferred values using BCLAc-Gibbs (middle), and the correlation estimated directly from the annotations provided (right). Note that each correlated matrix is subtracted from an identity matrix to remove the correlation of an annotator with itself.

annotators accordingly. Although the performance of BCLAc-Gibbs was sub-optimal when compared to BCLA-Gibbs, it only required approximately half of the sampling time, and it provides an additional modelling of the correlation of errors among annotators in their annotations. Examples of the inferred correlation of errors using the BCLAc-Gibbs approach for selecting seven and nine annotators are demonstrated in Fig. 8.12. It was expected that the direct estimation of correlation using the annotations provided was unable to indicate any meaningful relationship of the correlated errors among annotators (see right correlation matrix as compared to the reference $\boldsymbol{\rho}$ on the left matrix for (a) and (b) respectively in Fig. 8.12). In comparison, BCLAc-Gibbs was able to identify the key correlated relationship among annotators (see the middle matrix for (a) and (b) respectively in Fig. 8.12).

Table 8.3 The parameters of BCLAc for modelling the Capnobase RR dataset

Symbol	Definition	Value
k_b	shape of Gamma distribution for b	3
ϑ_b	scale of Gamma distribution for b	0.006
$\mathbf{S}$	the scale matrix of Inverse-Wishart distribution for $\mathbf{\Sigma}$	$\frac{1}{\lambda_M} I$
v	the degrees of freedom of Inverse-Wishart distribution for $\mathbf{\Sigma}$	R
k_0	a scale factor of $\mathbf{\Sigma}$	10
$\boldsymbol{\mu}_{\phi\Sigma}$	mean vector of the bias distribution	0
μ_ϕ	mean of the bias distribution	variable †
k_λ	shape of Gamma distribution for $\boldsymbol{\lambda}$	3 ‡
ϑ_λ	scale of Gamma distribution for $\boldsymbol{\lambda}$	0.02 ‡
k_α	shape of Gamma distribution for α_ϕ	5
ϑ_α	scale of Gamma distribution for α_ϕ	0.1
α_ρ, β_ρ	shape parameters of Beta distribution for $\boldsymbol{\rho}$	10,10

N.B.: b is the precision for the estimate of the ground truth. $\mathbf{\Sigma}$ is the covariance matrix. $\frac{1}{\lambda_M} I$ is an R $\times$ R diagonal matrix with entries obtained from a scalar, λ_M, as the inverse of the mean of the Gamma$(k_\lambda, \vartheta_\lambda)$. $\boldsymbol{\lambda}$ refers to signal-specific precisions. For the BCLAc-MH approach only: α_ϕ is the precision for the estimate of the bias from ground truth. The values with ‡ are determined with the assumption that the RR estimates provided by the best modulation signal is ± 2 bpm away from the reference [179, 180]. The values with † are estimated from the median RR estimates provided by the algorithms.

8.5.3 Capnobase RR Dataset

BCLA-Gibbs took about 38.8 seconds for running 5,000 draws of the 900 annotations provided from six algorithms using the same system as before. Similarly, BCLAc-Gibbs took approximately 68 seconds, and the first 3,000 samples were discarded as burn-in for both methods. The BCLAc-MH approach took approximately 21 hours to complete 10,000,000 iterations, where the parameter values were stored every 1,000 iterations; this process resulted in a total of 10,000 samples, and half of the them were discarded as burn-in.

The resulting values of the parameters and hyperparameters of the BCLA model for the Capnobase dataset were described previously in Chapter 6, Table 6.3. They were also considered in the BCLAc models with additional parameters as shown in Table 8.3.

Fig. 8.13 shows MAE results as before, across 42 subjects. We recall that the results were divided into three groups: (1) all available windows; (2) windows are discarded that have

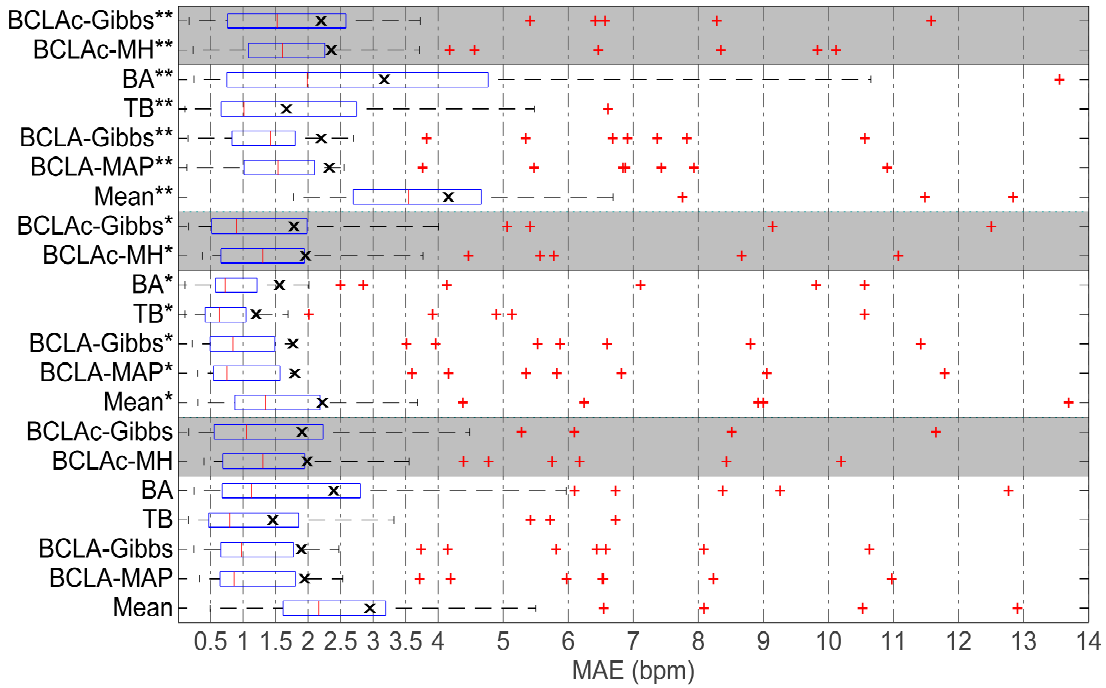


Fig. 8.13 Box plot of MAEs across 42 subjects for different fusion approaches. The median MAEs are indicated in red '-' in each box, while the black '×' indicates the mean MAE. The span of each box indicate the interquartile MAE range over 42 subjects for each fusion method when combining six algorithms. Notation: BA – the single best performing algorithm with least MAE across all subjects; TB – theoretical best algorithm with least MAE selected per subject; Mean* – the "Smart Fusion". Note that results associated with * indicate the MAEs were derived from windows excluding those have a standard deviation greater than four bpm, while those associated with ** indicate the results derived only from windows that have a standard deviation greater than four bpm. Note that the results of other approaches except BCLAc (shaded in grey colour) were reproduced from Fig. 6.12 for the purpose of comparison.

standard deviation greater than four bpm (*); (3) those windows that have standard deviation greater than four bpm (**). It may be seen that BCLAc-Gibbs consistently outperformed BCLAc-MH for all three groups. Furthermore, the results show that both BCLAc-Gibbs and BCLAc-MH have similar performance when compared to the BCLA approaches, with their mean MAEs being closest to the "theoretical best" (TB) algorithm. Out of 42 subjects when considering all windows, BCLAc-Gibbs produced a MAE of 1.91 ± 0.36 bpm, and had 17 subjects with smaller MAEs than BCLA-Gibbs, but had 22 and 30 subjects over BCLA-MAP and the "Smart Fusion". In comparison, BCLAc-MH produced a MAE of

1.98±0.33 bpm, and had 14, 16, 19, and 26 subjects with smaller MAEs than BCLA-Gibbs, BCLA-MAP, BCLAc-Gibbs, and the "Smart Fusion". It was unsurprising that the mean MAEs of BCLAc-MH were worse than the BCLAc-Gibbs approach as the RR dataset was considered *a priori* be described well by BCLAc with $\mathscr{M}_\Sigma$: as six RR estimates per window were derived from a set of three modulations, each using the AR- and FFT-based approaches, their bias values $\boldsymbol{\phi}$ were highly correlated. This relationship is assumed in BCLAc with $\mathscr{M}_\Sigma$, and it models the uncertainty in $\boldsymbol{\phi}$ using the scaled covariance matrix $\boldsymbol{\Sigma}/k_0$.

The results obtained with BCLAc-MH were also similar to those findings with the simulated datasets: the BCLAc-MH approach can be less reliable in its prediction when the number of recordings is relatively small with a small number of annotators (see the dataset of 100 annotations with five annotators as shown in Fig. 8.6 (a)). It was therefore expected that more recordings or annotators are necessary to see improved performance of the BCLAc-MH model in this setting.

8.6 Summary and Future Work

This chapter evaluated the proposed Bayesian Continuous-Valued Label Aggregrator with correlation annotators (i.e., BCLAc) to model the correlation of errors among automated algorithms. Two variants of the BCLAc framework were further discussed as to model the correlation of errors directly (i.e., the $\mathscr{M}_\rho$ model) and indirectly through the covariance matrix (i.e., the $\mathscr{M}_\Sigma$ model). It was shown that BCLAc was capable of inferring a reliable and robust estimation of the underlying ground truth, precision and bias values of annotators in an unsupervised manner, as well as providing a measure of similarity in the decision making process among annotators.

Currently, BCLAc with $\mathscr{M}_\Sigma$ used a single distribution to define the joint relationship of the precision of each annotator and their correlated errors in annotation, which might be less desire in practice. Furthermore, it cannot deal with the missing values in an optimal

manner since the covariance matrix in the model is only annotator number dependent, and more care need to be considered to provide decomposition of the covariance matrix to cater for missing annotations. Future work would be following the suggestions of the matrix re-parameterisation provided by Dominici *et al.* [204] to further optimise the BCLAc-Gibbs approach.

In comparison, BCLAc with $\mathscr{M}_\rho$ provides more flexibility in modelling the parameters of interest. However, the simulated and real datasets have shown that it requires a larger number of recordings/samples or number of annotators to provide a reliable estimation. Furthermore, as the posterior cannot be solved directly using a Gibbs sampler, instead a Metropolis Hastings (MH) (i.e., BCLAc-MH) algorithm was considered. This increased the complexity of the model as not only much larger number of draws required for convergence, but also extra number of parameters are necessary for fine tuning the proposal distribution in the MH algorithm. Future work will focus on speeding up the process of the BCLAc-MH approach.

In the Capnobase RR dataset, it was observed that although the precision values of the algorithms were accurately predicted, the mean of the distribution of their bias values varied for different subjects. In future work, a multilevel structure focusing on the population-based bias modelling should be incorporated in the proposed models, thereby providing an improvement of the RR estimation per individual subject.

Chapter 9

Conclusions and Future Work

9.1 Summary of Results

Expert labelling remains the gold standard in medical applications. However, scarcity and costs limit expert availability, and the labelling process is time consuming. While automated algorithms offer advantages of time efficiency, repeatability, and cost-saving benefits compared to manual labelling, there remains a substantial discrepancy in their estimation as well as reliability that limit their use. This thesis considered two commonly-encountered and clinically-important examples for the use of concept labels in clinical practice: the QT interval in the electrocardiogram and respiratory rate estimation from the photoplethysmogram. The "experts" could be either humans (such as trained clinicians), or automated algorithms, or some combination of both. This thesis focused on the case in which the experts are existing algorithms, where we sought to combine their outputs in a principled manner.

In this thesis, we have presented a series of methods for inferring a consensus in a scenario when only noisy annotations of some presumed underlying ground truth are provided. Two aforementioned medical applications are considered as exemplars to validate the proposed methods for aggregating the outputs of a group of mixed imperfect automated algorithms

in an unsupervised manner. A series of methods and their results are summarised in the following sub-sections.

9.1.1 The EM-R Algorithm

The EM-R algorithm (see Chapter 4) is the maximum likelihood approach proposed by Raykar *et al.* [89] using the expectation-maximisation for estimating consensus labels when only the noisy versions of the unknown true labels are available. We have implemented the EM-R algorithm and further improved its performance by proposing contextual features (such as beat-specific heart rate bHR, and beat-specific signal quality indices bSQIs) for the QT dataset as a proof-of-concept. The inferred results of both QT and respiratory rate datasets have shown that the EM-R algorithm outperformed the best-performing single algorithm as well as mean and median voting strategies. One of the short-comings of the EM-R algorithm is that it does not model the possibility of a bias inherent to an individual annotator. Incorporating the bias of each annotator could possibly allow for a better aggregation of the inferred ground truth, as annotators who are consistent but offset from the aggregated labels are heavily penalised in the current algorithm.

9.1.2 The sSTAPLE Algorithm

To investigate the effect of bias of an individual annotator, the scalar version of the Simultaneous Truth and Performance Level Estimation (sSTAPLE) proposed by Warfield *et al.* [3] was implemented and detailed in Chapter 5 of this thesis, and served as a second benchmarking approach for inferring the underlying true labels. The sSTAPLE algorithm was also adapted to deal with any dataset with missing labels. In the simulated dataset, the performance of sSTAPLE was superior over the EM-R algorithm, producing smaller RMSEs of the inferred labels. However, the former had larger RMSEs and MAEs than the latter when they were applied to the QT and the respiratory rate datasets. Our experiment of bias simulation has further shown that sSTAPLE yielded equal likelihood for any bias values,

which was in agreement with the study of Xing *et al*. [130]: sSTAPLE always assumes the mean of the bias values of annotators to be zero. This result explains the sub-optimal performance of the sSTAPLE algorithm as the automated algorithms tend to overestimate the QT intervals with a positive mean bias value. Automated algorithms also consistently underestimate or overestimate the respiratory rates. Furthermore, unlike EM-R, the sSTAPLE algorithm does not model the ground truth as a function of feature inputs, thereby making it less ideal for aggregating medical labels in practice.

9.1.3 The BCLA Model

Building upon the prior investigations of the two state-of-the-art benchmarking algorithms, the BCLA model was proposed in Chapter 5 as a generative model that aggregates medical labels in an unsupervised manner, to estimate jointly the assumed bias and precision of each annotator. The model was then used to infer the underlying ground truth with consideration of two approaches that are now described.

The BCLA-MAP Approach

A maximum-a-posterior approach for formulating the BCLA framework (i.e., BCLA-MAP) with independent annotators was applied to the QT interval estimation using labels from a simulated dataset, as well as from the QT dataset. When applied to the QT and respiratory rate datasets, BCLA-MAP significantly outperformed previously-mentioned voting strategies.

The BCLA-Gibbs Approach

Chapter 6 introduced the Gibbs sampler for taking a fully-Bayesian approach for BCLA (i.e., BCLA-Gibbs) with independent annotators. It was applied to both the simulated and QT datasets, as well as the respiratory rate dataset. The performance of the BCLA-Gibbs method was compared to all methods considered in previous chapters. In the QT dataset, although

the RMSEs of BCLA-Gibbs was similar to those of BCLA-MAP, the BCLA-Gibbs method not only provided similarly accurate estimates but also produced a confidence interval in its estimation. Furthermore, BCLA-Gibbs was more robust for inferring the true labels particularly when fewer number of annotators are available (e.g., up to 11 annotators in the QT dataset). Similar results were also observed with the respiratory rate dataset.

9.1.4 The BCLAs Model

The BCLA model assumed that the precision of each annotator is not affected by the signal quality of the recordings (i.e., their annotations would remain the same for recordings with different level of signal qualities that are affected by noise). This may not be necessarily true in the context of medical labels. Noise is a prominent cause of inaccurate detection of medical data. Thereby an extension to BCLA with independent annotators incorporating the notion of signal quality (i.e., BCLAs) was proposed in Chapter 7; this involves a confidence measure relating to the quality of the input waveform. We validated the BCLAs model using our two clinical exemplars as described above. Two approaches, based on MAP and Gibbs sampling, were proposed for BCLAs (denoted BCLAs-MAP and BCLAs-Gibbs, respectively). The results obtained with the QT and respiratory rate datasets demonstrated that the BCLAs model was able to provide reliable estimation of the ground truth using the signal quality confidence extension. No improvement was observed in the respiratory rate dataset as the PPGs are of high quality with very little noise. For the QT dataset, more substantial improvements were shown when selecting fewer number of annotators (e.g., up to 11 annotators): the results of the BCLAs approaches outperformed those of the BCLA approaches consistently. Once again, a fully-Bayesian approach (such as with the Gibbs sampler) was shown to provide more reliable results as compared to the MAP approach in both the BCLA and BCLAs models.

9.1.5 The BCLAc Model

The assumption made in the proposed BCLA model is that all annotators are conditionally independent given the ground truth. This simplified assumption may be violated as the annotation errors of the automated annotators might be correlated as they were estimated using variants of the same methodology. In order to model the relationship among the annotators, the BCLA with correlation annotators (BCLAc) model was explored in Chapter 8 to uncover its potential in further improving the inferred medical labels. Two variations of the BCLAc framework were presented, where one involves modelling the correlation matrix between expert labels (i.e., BCLAc with $\mathscr{M}_{\rho}$), and the other one involves modelling the covariance matrix between expert labels (i.e., BCLAc with $\mathscr{M}_{\Sigma}$). In both simulated and real datasets, the results demonstrated that the proposed BCLAc methodologies can be used to combine medical data from multiple sources, to infer a reliable and robust estimation of the underlying ground truth in an unsupervised manner. Similar to the BCLA approaches, no training data needed to be held out for optimising the parameters of the BCLAc framework, and no prior knowledge of the performance of each algorithm was given. In addition to similar performance as the BCLA-Gibbs approach, the BCLAc models provide a measure of the correlation of errors among annotators, which imply that annotators can be grouped based on their correlated decision making process.

9.1.6 Key Contributions

This thesis has made the following novel contributions:

- A literature review on the relevant models that aim to fuse noisy labels to infer jointly the consensus, and precision and bias of an annotator, via unsupervised learning.
- Implementation of the EM-R algorithm and further proposed an incorporation of contextual features to improve its accuracy (see corresponding journal and conference papers [13, 127, 205–209]).

- Implementation of sSTAPLE and further investigated the effect of bias modelling.
- A proposal of the BCLA framework with independent annotators as a generative model to aggregate continuous-valued labels in an unsupervised manner, to estimate jointly the assumed bias and precision of each annotator. The MAP (see corresponding journal and conference papers [210, 211]) and Gibbs sampling approaches were further proposed to derive parameters in the BCLA model.
- A proposal of BCLAs that serves as an extension to BCLA with independent annotators incorporating the notion of signal quality, and subsequently the MAP and Gibbs sampling approaches were provided to derive parameters of interest.
- A proposal of the BCLA framework with correlated annotators (i.e., BCLAc) and its two variations, BCLAc with $\mathscr{M}_{\rho}$ and BCLAc with $\mathscr{M}_{\Sigma}$. The Metropolis Hastings and Gibbs sampling approaches were further proposed to derive the parameters in BCLAc.

9.1.7 Discussions and Limitations

To demonstrate the importance of formulating a consensus where the ground truth is not available, two commonly-encountered and clinically-important examples are described in the thesis: (1) the QT interval in the electrocardiogram and (2) respiratory rate estimation from the photoplethysmogram. In the current clinical setting, there is no standardised guidelines to define the targeted accuracies for automated estimations of the QT interval, as well as for respiratory rate, therefore making it difficult to assess the acceptability of automated annotation algorithms. Moreover, We have challenged the definition of an expert, as there exist large inter- and intra-variabilities in both humans and algorithms in their accuracies. It is often impossible to assess who is the best-performing annotator without prior knowledge, especially on unseen data. Here we have presented a series of models that deal with this particular issue and guaranteed that the consensus at least as well, if not better, than the best person or algorithm available. Furthermore, both humans and algorithms vary in their skill

as a function of context. Our proposed models can optimise the amount of information that can be drawn from each individual in an unsupervised manner.

We have validated our models in each aforementioned clinical dataset where the inferred consensus was generated from the estimates provided by automated algorithms in a unsupervised manner. These inferred estimates of consensus (from either the automated QT intervals in the electrocardiogram or the automated respiratory rate estimates from the photoplethysmogram) were compared with the unseen reference datasets (i.e., where QT intervals generated by up to 15 manual experts or respiratory labels provided by a manual expert on an invasive capnogram):

- In the QT dataset, a minimum RMSE of 11.68±0.58 ms and a MAE of 8.64±0.48 ms was achieved when using proposed models, which significantly outperformed the best-performing automated algorithm (RMSE of 15.12±1.22 ms and MAE of 10.73±0.86 ms). In the reference dataset, although the best-performing manual expert had a RMSE of 6.65 ms, there were only a total of four (out of 20) manual experts achieved a RMSE below 12 ms. In the case of selecting annotations from three automated algorithms with at least 50% records annotated, our proposed models had a minimum RMSE of 26.64±10.32 ms and MAE of 20.48±10.25 ms, which were comparable to errors made by manual experts ranging from 10 to 30 ms [22, 23, 28, 212].
- In the respiratory dataset, a minimum MAE of 1.89±0.36 bpm was achieved when fusing six AR- and FFT-based algorithms (each derived from three respiratory-induced modulations) across 42 subjects. In the case when only considering three FFT-based algorithms (from three modulations), the MAE was 1.95±0.39 bpm. Nevertheless, our proposed models outperformed the best-performing algorithm (MAE of 2.39±0.43 bpm) in both cases. Although the MAE results were not significantly different, the comparison was made for only 42 subjects and can not be generalised to a population. Currently, there is no clinical guideline for defining accuracy in automated respiratory

rate estimation, an error within 2 bpm is generally considered acceptable for non-invasive respiratory sensors [179, 180]. Furthermore, the modulation which derived the best-performing algorithm may not always be readily available depending on the age and health condition of a subject, making it impossible to choose a specific modulation for respiratory rate estimation in practice. It is therefore necessary to combine all possible modulations available in a robust manner using our proposed approaches. In comparison to the state-of-art "Smart fusion" [81] for combing all AR- and FFT-based estimates using all modulations (MAE of 2.22$\pm$0.41 bpm) which discarded 55.4% of respiratory estimates as artefacts, our proposed models considered all estimates and had 73.8% subjects with smaller MAEs.

In terms of limitations, our models currently assume that the performance of an annotator is consistent throughout records, which might not be the case as the performance can be features dependent (such as difficulty of the task, quality of the recordings). Furthermore, our models assume that each record is independent, which might limit its use in application for continuous monitoring as data are correlated through time. In the case of manual annotators, his performance is time-variant as he might improve through learning or might drop due to inattention or fatigue. Finally, our models operate in an unsupervised manner, where they aggregate annotations by processing the entire dataset, which might be computationally expensive for a large dataset. Future work will be described in the next section to address these limitations.

9.2 Future Work

9.2.1 BCLAc with Contextual Features and Signal Quality Extension

This thesis has demonstrated the potential of using the BCLAc models to describe the correlation of errors among annotators. Future work could include the use of physiological

features in the proposed models and further validate their performance, as well as extending the BCLAc models to incorporate the signal quality extension.

9.2.2 Modelling of the Feature Components

Training a Classifier without Ground Truth

The ground truth labels considered in this thesis were assumed to be modelled from a Gaussian distribution with mean described as a linear regression function (i.e., $f(\mathbf{x}, \mathbf{w})$) with contextual features (i.e., $\mathbf{x}$) capturing the underlying physiology. The regression coefficients, $\mathbf{w}$, were learned jointly with the ground truth and annotators' competence in an unsupervised manner. Such coefficients can also be used to predict the future unseen labels for a given set of features. Future work could focus on exploring other functions to describe the mean of the ground truth: kernel functions [95] can be considered to transform the relationship of the input features to better describe the underlying ground truth. Furthermore, kernel functions used with Gaussian processes can be considered to model the time-variant relationship across recordings of the underlying true labels.

Feature-Specific Competence

Recently, Yan *et al.* [114] described the competence of an annotator as a logistic function of the input features for combining binary labels. Ouyang *et al.* [129] considered an annotator to have K pairs of bias and precision values, each depending on the underlying difficulty level of a recording: the proposed model assumed the observed annotations follow essentially a Gaussian mixture distribution, and that the mixture coefficients were the fractions of recordings assigned to the corresponding levels of task difficulty. Instead of modelling task difficulty, Zhang *et al.* [115] assumed that the sensitivity and specificity of an annotator for binary labels were mixture component dependent, where a total of K mixture components was fitted to the distribution of the input features.

Future work could likewise use a Gaussian mixture model to approximate the input features, then infer the precision and bias of each annotator, the ground truth labels, as well as the mixture coefficients jointly. Physiological features such as age and pathological condition could be used to help us to learn the precision and bias of an algorithm for different clustered groups. Furthermore, in a resource-constrained setting where gathering outputs from all annotators for a new recording is not possible, one could select the label provided by an annotator from a feature-specific cluster with the highest accuracy to represent the underlying truth label.

Mixture of Experts

The Mixture of Experts approach [213] is a technique used for assigning weights for models and combines their outputs through a generalised non-linear rule. The weights or the mixture coefficients are feature-specific and determined by a gating network, and it is commonly trained using the EM algorithm [214]. Mixture of Experts can also be seen as model-combination method, because different models contribute to the distribution in different regions of the input features, which are determined by the gating function [95]. This can be seen as being an extension to the Gaussian mixture model described previously. The current thesis focuses on combining the outputs of experts in which the experts are existing automated algorithms, future work can be done by relaxing this assumption to combine a mixture of manual and automated annotators, and each can be considered as an "expert".

9.2.3 On-line Adaptive Learning

Currently, the proposed algorithms in this thesis aggregate annotations by processing the entire dataset as a batch, which might be computationally expensive for a large dataset. In the case where recordings are supplied in a sequential order (i.e., individual recording is available at different time intervals to be labelled), sequential learning might be more appropriate for real-time applications. A sequential Bayesian update can be used to infer the

ground truth labels: by observing annotations provided previously for $N-1$ recordings, the posterior of the parameters of interest (such as the underlying true labels, precision and bias of an individual annotator) can be estimated using Bayes' theorem, where it is proportional to the multiplication of the priors of the parameters and the likelihood of the dataset **D**. Upon observing the Nth recording, the revised posterior of the parameters can be estimated as the multiplication of their posterior after observing annotations provided for $N-1$ recordings and the likelihood of the annotations for the Nth recording given the parameters [95]. This can be seen as a single state in a dynamic system that changes over time following the first-order Markovian property, where the current state N only depends on the previous state $N-1$. Practical approaches such as Kalman filtering and particle filtering could be used to model the posterior at the Nth recording or the Nth state, based on a linearity assumption for the dynamic system, and whether the posterior density of an individual parameter is Gaussian or non-Gaussian, respectively [215]. Similar to the concept proposed in Simpson *et al.* [119] for their confusion matrix, the proposed future work could model the precision and bias of each annotator as a function of time.

9.2.4 Future Applications

An Intelligent Electrophysiological Annotation Platform

CrowdLabel, a web-based open-source annotation system was developed to provide a platform for medical labelling [13]. Our recent study [209] has demonstrated the process of integrating the mobile-end of the system with a back-end annotation system (i.e., *CrowdLabel*) for reviewing and scoring the quality of an ECG signal. This proposed intelligent cardiac health monitoring and review system contains an improved version of the existing state-of-the-art EM-R algorithm running in the back-end to provide an adjudication of ECG annotations from the annotators (such as trainees and/or automated algorithms) in real-time. The proposed algorithms in this thesis such as BCLA, BCLAc, and BCLAs have shown their

superior performance over the EM-R algorithm, and hence can be implemented in the system to further improve the estimations of the unknown true labels in an unsupervised manner.

An Accreditation System for Electrocardiographic Readers

Currently, there is no standardisation for testing competency of ECG reading, both in terms of accuracy and consistency, as well as no empirical method for measuring level of expertise, even though the annotation heavily relies on the experts' experiences. The *CrowdLabel* annotation system could act as a standardised annotation tool as part of an accreditation system for electrocadiographic readers, allowing clinicians to be regularly tested in an ECG annotation practice. This system could also be potentially used for educational purposes where training medical students in ECG labelling. To monitor the progress of a user in *CrowdLabel*, adaptive BCLA could be used to keep track of the user's learning curve that changes with time and provide real-time feedback on his/her performance.

References

[1] International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use, "Guidance for Industry E14: Clinical Evaluation of QT/ QTc Interval Prolongation and Proarrhythmic Potential for Non-Antiarrhythmic Drugs," 2014.

[2] S. M. Salerno, P. C. Alguire, and H. S. Waxman, "Competency in interpretation of 12 - lead electrocardiograms: a summary and appraisal of published evidence," *Annals of Internal Medicine*, vol. 138, no. 9, pp. 751–760, 2003.

[3] S. K. Warfield, K. H. Zou, and W. M. Wells, "Validation of image segmentation by estimating rater bias and variance," *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences*, vol. 366, pp. 2361–2375, 2008.

[4] R. R. Bond, T. Zhu, D. D. Finlay, B. Drew, P. D. Kligfield, D. Guldenring, C. Breen, A. G. Gallagher, M. J. Daly, and G. D. Clifford, "Assessing Computerised Eye Tracking Technology for Gaining Insight into Expert Interpretation of the 12-lead Electrocardiogram: An Objective Quantitative Approach," *Journal of Electrocardiology*, vol. 47, no. 6, pp. 895 – 906, 2014.

[5] J. P. Metlay, W. N. Kapoor, and M. J. Fine, "Does this patient have community-acquired pneumonia?: Diagnosing pneumonia by history and physical examination," *Journal of the American College of Cardiology*, vol. 278, no. 17, pp. 1440–1445, 1997.

[6] F. Molinari, L. Gentile, P. Manicone, R. Ursini, L. Raffaelli, M. Stefanetti, A. D'Addona, T. Pirronti, and L. Bonomo, "Interobserver variability of dynamic MR imaging of the temporomandibular joint," *La Radiologia Medica*, vol. 116, no. 8, pp. 1303–1312, 2011.

[7] I. Neamatullah, M. Douglass, L. Lehman, A. Reisner, M. Villarroel, W. Long, P. Szolovits, G. Moody, R. Mark, and G. D. Clifford, "Automated de-identification of free-text medical records," *BMC Medical Informatics and Decision Making*, vol. 8, no. 1, p. 32, 2008.

[8] O. S. Ofer Dekel, "Good Learners for Evil Teachers," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009.

[9] H. Valizadegan, Q. Nguyen, and M. Hauskrecht, "Learning Medical Diagnosis Models from Multiple Experts," in *American Medical Informatics Association Annual Symposium Proceedings*, pp. 921–930, AMIA, 2012.

[10] World Health Organisation, "Cardiovascular diseases." http://www.who.int/mediacentre/factsheets/fs317/en/index.html, August 2015.

[11] World Health Organization, "Global action plan for the prevention and control of noncommunicable diseases 2013-2020," tech. rep., World Health Organization, Geneva, Switzerland, 2013.

[12] G. D. Clifford, F. Azuaje, and P. E. McSharry, *Advanced Methods and Tools for ECG Analysis*. Engineering in Medicine and Biology, Norwood, MA, USA: Artech House, October 2006.

[13] T. Zhu, J. Behar, T. Papastylianou, and G. D. Clifford, "CrowdLabel: A crowdsourcing platform for electrophysiology," in *Computing in Cardiology Conference*, pp. 789–792, Sept 2014.

[14] M. A. Oudijk, A. Kwee, G. H. Visser, S. Blad, E. J. Meijboom, and K. G. Rosén, "The effects of intrapartum hypoxia on the fetal QT interval," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 111, no. 7, pp. 656–660, 2004.

[15] T. Davis, B. Singh, K. E. Choo, J. Ibrahim, J. L. Spencer, and A. St John, "Dynamic assessment of the electrocardiographic QT interval during citrate infusion in healthy volunteers," *British Heart Journal*, vol. 73, no. 6, pp. 523–526, 1995.

[16] S. Severi, E. Grandi, C. Pes, F. Badiali, F. Grandi, and A. Santoro, "Calcium and potassium changes during haemodialysis alter ventricular repolarization duration: in vivo and in silico analysis," *Nephrology Dialysis Transplantation*, vol. 23, no. 4, pp. 1378–1386, 2008.

[17] P. E. Foglia, A. Bettinelli, C. Tosetto, C. Cortesi, L. Crosazzo, A. Edefonti, and M. G. Bianchetti, "Cardiac work up in primary renal hypokalaemia – hypomagnesaemia (gitelman syndrome)," *Nephrology Dialysis Transplantation*, vol. 19, no. 6, pp. 1398–1402, 2004.

[18] L. X. Cubeddu, "Iatrogenic QT abnormalities and fatal arrhythmias: mechanisms and clinical significance," *Current Cardiology Reviews*, vol. 5, no. 3, pp. 166–176, 2009.

[19] W. Zareba, "Drug induced QT prolongation," *Cardiology Journal*, vol. 14, no. 6, pp. 523–533, 2007.

[20] G. K. Isbister and C. B. Page, "Drug induced QT prolongation: the measurement and assessment of the QT interval in clinical practice," *British Journal of Clinical Pharmacology*, vol. 76, no. 1, pp. 48–57, 2013.

[21] D. M. Roden, "Drug-Induced Prolongation of the QT Interval," *New England Journal of Medicine*, vol. 350, no. 10, pp. 1013–1022, 2004.

[22] I. Christov, I. Dotsinsky, I. Simova, R. Prokopova, E. Trendafilova, and S. Naydenov, "Dataset of manually measured QT intervals in the electrocardiogram," *Biomedical Engineering Online*, vol. 5, p. 31, 2006.

[23] F. A. Ehlert, J. J. Goldberger, J. E. Rosenthal, and A. H. Kadish, "Relation between QT and RR intervals during exercise testing in atrial fibrillation," *American Journal of Cardiology*, vol. 70, no. 3, pp. 332–338, 1992.

[24] E. Lepeschkin and B. Surawicz, "The measurement of the QT interval of the electrocardiogram," *Circulation*, vol. 6, no. 3, pp. 378–388, 1952.

[25] M. Malik and A. J. Camm, "Evaluation of Drug-Induced QT Interval Prolongation," *Drug Safety*, vol. 24, no. 5, pp. 323–351, 2001.

[26] P. G. Postema, J. S. D. Jong, I. A. V. der Bilt, and A. A. Wilde, "Accurate electrocardiographic assessment of the QT interval: Teach the tangent," *Heart Rhythm*, vol. 5, no. 7, pp. 1015 – 1018, 2008.

[27] G. D. Clifford and M. Villarroel, "Model-based determination of QT intervals," in *Computers in Cardiology*, pp. 357–360, Sept 2006.

[28] S. M. Al-Khatib, N. M. Allen LaPointe, J. M. Kramer, A. Y. Chen, B. G. Hammill, L. Delong, and R. M. Califf, "A survey of health care practitioners' knowledge of the QT interval," *Journal of General Internal Medicine*, vol. 20, no. 5, pp. 392–396, 2005.

[29] S. Viskin, U. Rosovski, A. J. Sands, E. Chen, P. M. Kistler, J. M. Kalman, L. R. Chavez, P. I. Torres, F. E. Cruz, O. A. Centurion, A. Fujiki, P. Maury, X. Chen, A. D. Krahn, F. Roithinger, L. Zhang, G. M. Vincent, and D. Zeltser, "Inaccurate electrocardiographic interpretation of long QT: the majority of physicians cannot recognize a long QT when they see one.," *Heart Rhythm*, vol. 2, pp. 569–574, 2005.

[30] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[31] B. Charbit, E. Samain, P. Merckx, and C. Funck-Brentano, "QT Interval MeasurementEvaluation of Automatic QTc Measurement and New Simple Method to Calculate and Interpret Corrected QT Interval," *Anesthesiology*, vol. 104, no. 2, pp. 255–260, 2006.

[32] K. Hnatkova, Y. Gang, V. N. Batchvarov, and M. Malik, "Precision of QT Interval Measurement by Advanced Electrocardiographic Equipment," *Pacing and Clinical Electrophysiology*, vol. 29, no. 11, pp. 1277–1284, 2006.

[33] P. Kligfield, E. W. Hancock, E. D. Helfenbein, E. J. Dawson, M. A. Cook, J. M. Lindauer, S. H. Zhou, and J. Xue, "Relation of QT Interval Measurements to Evolving Automated Algorithms from Different Manufacturers of Electrocardiographs," *American Journal of Cardiology*, vol. 98, no. 1, pp. 88 – 92, 2006.

[34] M. Malik, "Errors and misconceptions in ECG measurement used for the detection of drug induced QT interval prolongation," *Journal of Electrocardiology*, vol. 37, Supplement, pp. 25 – 33, 2004.

[35] M. Sano, Y. Aizawa, Y. Katsumata, N. Nishiyama, S. Takatsuki, S. Kamitsuji, N. Kamatani, and K. Fukuda, "Evaluation of Differences in Automated QT/QTc Measurements between Fukuda Denshi and Nihon Koden Systems," *PLoS ONE*, vol. 9, p. e106947, Sept 2014.

[36] I. Savelieva, G. Yi, X. hua Guo, K. Hnatkova, and M. Malik, "Agreement and Reproducibility of Automatic Versus Manual Measurement of QT Interval and QT Dispersion," *The American Journal of Cardiology*, vol. 81, no. 4, pp. 471 – 477, 1998.

[37] G. B. Moody, H. Koch, and U. Steinhoff, "The PhysioNet/ Computers in Cardiology Challenge 2006: QT interval measurement," in *Computing in Cardiology Conference*, pp. 313 –316, 2006.

[38] J. De Trefford and K. Lafferty, "What does photoplethysmography measure?," *Medical and Biological Engineering and Computing*, vol. 22, no. 5, pp. 479–480, 1984.

[39] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiological Measurement*, vol. 28, no. 3, p. R1, 2007.

[40] K. H. Shelley, "Photoplethysmography: beyond the calculation of arterial oxygen saturation and heart rate," *Anesthesia and Analgesia*, vol. 105, pp. S31–6, Dec 2007.

[41] L. Tarassenko and D. Clifton, "Semiconductor wireless technology for chronic disease management," *Electronics Letters*, vol. 47, pp. S30–S32, December 2011.

[42] M. Cretikos, J. Chen, K. Hillman, R. Bellomo, S. Finfer, and A. Flabouris, "The objective medical emergency team activation criteria: A case - control study," *Resuscitation*, vol. 73, no. 1, pp. 62 – 72, 2007.

[43] N. Farrohknia, M. Castrén, A. Ehrenberg, L. Lind, S. Oredsson, H. Jonsson, K. Asplund, and K. E. Göransson, "Emergency department triage scales and their components: a systematic review of the scientific evidence," *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, vol. 19, no. 42, pp. 1–13, 2011.

[44] J. Fieselmann, M. Hendryx, C. Helms, and D. Wakefield, "Respiratory rate predicts cardiopulmonary arrest for internal medicine inpatients," *Journal of General Internal Medicine*, vol. 8, no. 7, pp. 354–360, 1993.

[45] J. Kause, G. Smith, D. Prytherch, M. Parr, A. Flabouris, and K. Hillman, "A comparison of Antecedents to Cardiac Arrests, Deaths and EMergency Intensive care Admissions in Australia and New Zealand, and the United Kingdom – the ACADEMIA study," *Resuscitation*, vol. 62, no. 3, pp. 275 – 282, 2004.

[46] World Health Organization, "Pocket book of hospital care for children: Guidelines for the management of common illnesses with limited resources," tech. rep., World Health Organization, Geneva, Switzerland, 2005.

[47] T. J. Hodgetts, G. Kenward, I. G. Vlachonikolis, S. Payne, and N. Castle, "The identification of risk factors for cardiac arrest and formulation of activation criteria to alert a medical emergency team," *Resuscitation*, vol. 54, no. 2, pp. 125–131, 2002.

[48] R. C. Bone, R. A. Balk, F. B. Cerra, R. P. Dellinger, A. M. Fein, W. A. Knaus, R. M. Schein, and W. J. Sibbald, "Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. the ACCP/SCCM consensus conference committee. american college of chest physicians/society of critical care medicine," *Chest*, vol. 101, no. 6, pp. 1644–1655, 1992.

[49] M. A. Cretikos, R. Bellomo, K. Hillman, J. Chen, S. Finfer, and A. Flabouris, "Respiratory rate: the neglected vital sign," *Medical Journal of Australia*, vol. 188, no. 11, pp. 657–659, 2008.

[50] A. Khalil, G. Kelen, and R. E. Rothman, "A simple screening tool for identification of community-acquired pneumonia in an inner city emergency department," *Emergency Medicine Journal*, vol. 24, no. 5, pp. 336–338, 2007.

[51] M. Seyal, L. M. Bateman, T. E. Albertson, T. Lin, and C. Li, "Respiratory changes with seizures in localization-related epilepsy: Analysis of periictal hypercapnia and airflow patterns," *Epilepsia*, vol. 51, no. 8, pp. 1359–1364, 2010.

[52] World Health Organization, "Fourth programme report, 1988 –1989: ARI Programme for Control of Acute Respiratory Infections," tech. rep., World Health Organization, Geneva, Switzerland, 1990.

[53] Z. V. Edmonds, W. R. Mower, L. M. Lovato, and R. Lomeli, "The reliability of vital sign measurements," *Annals of Emergency Medicine*, vol. 39, no. 3, pp. 233 – 237, 2002.

[54] E. A. Hooker, D. J. O'Brien, D. F. Danzl, J. A. Barefoot, and J. E. Brown, "Respiratory rates in emergency department patients," *The Journal of Emergency Medicine*, vol. 7, no. 2, pp. 129–132, 1989.

[55] P. B. Lovett, J. M. Buchwald, K. Stürmann, and P. Bijur, "The vexatious vital: Neither clinical measurements by nurses nor an electronic monitor provides accurate measurements of respiratory rate in triage," *Annals of Emergency Medicine*, vol. 45, no. 1, pp. 68 – 76, 2005.

[56] E. Simoes, R. Roark, S. Berman, L. Esler, and J. Murphy, "Respiratory rate: measurement of variability over time and accuracy at different counting periods.," *Archives of Disease in Childhood*, vol. 66, no. 10, pp. 1199–1203, 1991.

[57] E. E. Wang, B. J. Law, D. Stephens, J. M. Langley, N. E. MacDonald, J. L. Robinson, S. Dobson, J. McDonald, F. D. Boucher, V. de Carvalho, *et al.*, "Study of interobserver reliability in clinical assessment of RSV lower respiratory illness: A pediatric investigators collaborative network for infections in Canada (PICNIC) study," *Pediatric Pulmonology*, vol. 22, no. 1, pp. 23–27, 1996.

[58] World Health Organization, "IMCI technical seminar on Acute Respiratory Infection," tech. rep., World Health Organization, 2001.

[59] S. Gupta, C. Gennings, and R. P. Wenzel, "R = 20: bias in the reporting of respiratory rates," *The American Journal of Emergency Medicine*, vol. 26, no. 2, pp. 237 – 239, 2008.

[60] C. J. Cook and G. B. Smith, "Do textbooks of clinical examination contain information regarding the assessment of critically ill patients?," *Resuscitation*, vol. 60, no. 2, pp. 129 – 136, 2004.

[61] H. Gao, A. McDonnell, D. Harrison, T. Moore, S. Adam, K. Daly, L. Esmonde, D. Goldhill, G. Parry, A. Rashidian, C. Subbe, and S. Harvey, "Systematic review and evaluation of physiological track and trigger warning systems for identifying at-risk patients on the ward," *Intensive Care Medicine*, vol. 33, no. 4, pp. 667–679, 2007.

[62] G. B. Smith, D. R. Prytherch, P. E. Schmidt, and P. I. Featherstone, "Review and performance evaluation of aggregate weighted 'track and trigger' systems," *Resuscitation*, vol. 77, no. 2, pp. 170 – 179, 2008.

[63] L. Tarassenko, D. A. Clifton, M. R. Pinsky, M. T. Hravnak, J. R. Woods, and P. J. Watkinson, "Centile-based early warning scores derived from statistical distributions of vital signs," *Resuscitation*, vol. 82, no. 8, pp. 1013 – 1018, 2011.

[64] National Institute for Clinical Excellence, "Acutely ill patients in hospital: Recognition of and response to acute illness in adults in hospital," tech. rep., National Institute for Clinical Excellence, 2007.

[65] Royal College of Physicians, "National early warning scores (NEWS): Standardising the assessment of acute-illness severity in the NHS," tech. rep., Royal College of Physicians, 2012.

[66] D. R. Prytherch, G. B. Smith, P. E. Schmidt, and P. I. Featherstone, "Views – towards a national early warning score for detecting adult inpatient deterioration," *Resuscitation*, vol. 81, no. 8, pp. 932 – 937, 2010.

[67] G. B. Smith, D. R. Prytherch, P. Meredith, P. E. Schmidt, and P. I. Featherstone, "The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death," *Resuscitation*, vol. 84, no. 4, pp. 465 – 470, 2013.

[68] J. Hogan, "Why don't nurses monitor the respiratory rates of patients?," *British Journal of Nursing*, vol. 15, no. 9, pp. 489–492, 2006.

[69] S. Ullah, H. Higgins, B. Braem, B. Latre, C. Blondia, I. Moerman, S. Saleem, Z. Rahman, and K. S. Kwak, "A comprehensive survey of wireless body area networks," *Journal of Medical Systems*, vol. 36, no. 3, pp. 1065–1094, 2012.

[70] G. D. Clifford and D. A. Clifton, "Wireless technology in disease management and medicine," *Annual Review of Medicine*, vol. 63, pp. 479–492, 2012.

[71] C. Free, G. Phillips, L. Galli, L. Watson, L. Felix, P. Edwards, V. Patel, and A. Haines, "The effectiveness of mobile-health technology-based health behaviour change or disease management interventions for health care consumers: A systematic review," *PLoS Medicine*, vol. 10, p. e1001362, 01 2013.

[72] L. Clifton, D. A. Clifton, M. A. F. Pimentel, P. J. Watkinson, and L. Tarassenko, "Predictive Monitoring of Mobile Patients by Combining Clinical Observations With Data From Wearable Sensors," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, pp. 722–730, May 2014.

[73] H. Rutter, C. Velardo, C. Toms, V. Williams, L. Tarassenko, A. Farmer, *et al.*, "Using a Mobile Health Application to Support Self –Management in COPD –Development of Alert Thresholds Derived from Variability in Self-Reported and Measured Clinical Variables," *American Journal of Respiratory and Critical Care Medicine*, vol. 189, p. A1396, 2014.

[74] T. Wu, V. Blazek, and H. J. Schmitt, "Photoplethysmography imaging: a new noninvasive and noncontact method for mapping of the dermal perfusion changes," in *Optical Techniques and Instrumentation for the Measurement of Blood Composition, Structure, and Dynamics*, vol. 4163, pp. 62–70, 2000.

[75] K. Humphreys, T. Ward, and C. Markham, "Noncontact simultaneous dual wavelength photoplethysmography: a further step toward noncontact pulse oximetry," *Review of Scientific Instruments*, vol. 78, no. 4, p. 044304, 2007.

[76] F. Wieringa, F. Mastik, and A. Steen, "Contactless Multiple Wavelength Photoplethysmographic Imaging: A First Step Toward "SpO_2 Camera" Technology," *Annals of Biomedical Engineering*, vol. 33, no. 8, pp. 1034–1041, 2005.

[77] M. Poh, D. McDuff, and R. Picard, "Advancements in Noncontact, Multiparameter Physiological Measurements Using a Webcam," *IEEE Transactions on Biomedical Engineering*, vol. 58, pp. 7–11, Jan 2011.

[78] L. Tarassenko, M. Villarroel, A. Guazzi, J. Jorge, D. A. Clifton, and C. Pugh, "Non-contact video-based vital sign monitoring using ambient light and auto-regressive models," *Physiological Measurement*, vol. 35, no. 5, pp. 807–831, 2014.

[79] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light," *Optics Express*, vol. 16, no. 26, pp. 21434–21445, 2008.

[80] P. Addison, J. Watson, M. Mestek, and R. Mecca, "Developing an algorithm for pulse oximetry derived respiratory rate (RRoxi): a healthy volunteer study," *Journal of Clinical Monitoring and Computing*, vol. 26, no. 1, pp. 45–51, 2012.

[81] W. Karlen, S. Raman, J. Ansermino, and G. Dumont, "Multiparameter Respiratory Rate Estimation From the Photoplethysmogram," *IEEE Transactions on Biomedical Engineering*, vol. 60, pp. 1946–1953, July 2013.

[82] A. J. Buda, M. R. Pinsky, N. B. Ingels, G. T. Daughters, E. B. Stinson, and E. L. Alderman, "Effect of intrathoracic pressure on left ventricular performance," *New England Journal of Medicine*, vol. 301, no. 9, pp. 453–459, 1979.

[83] D. Meredith, D. Clifton, P. Charlton, J. Brooks, C. Pugh, and L. Tarassenko, "Photoplethysmographic derivation of respiratory rate: a review of relevant physiology," *Journal of Medical Engineering & Technology*, vol. 36, no. 1, pp. 1–7, 2012.

[84] L. Nilsson, A. Johansson, and S. Kalman, "Respiratory variations in the reflection mode photoplethysmographic signal. Relationships to peripheral venous pressure," *Medical and Biological Engineering and Computing*, vol. 41, no. 3, pp. 249–254, 2003.

[85] P. Grossman, F. H. Wilhelm, and M. Spoerle, "Respiratory sinus arrhythmia, cardiac vagal control, and daily activity," *American Journal of Physiology - Heart and Circulatory Physiology*, vol. 287, no. 2, pp. H728–H734, 2004.

[86] C. Orphanidou, S. Fleming, S. Shah, and L. Tarassenko, "Data fusion for estimating respiratory rate from a single-lead ECG," *Biomedical Signal Processing and Control*, vol. 8, no. 1, pp. 98 – 105, 2013.

[87] A. Johansson, "Neural network for photoplethysmographic respiratory rate monitoring," *Medical and Biological Engineering and Computing*, vol. 41, no. 3, pp. 242–248, 2003.

[88] P. Leonard, N. Grubb, P. Addison, D. Clifton, and J. Watson, "An algorithm for the detection of individual breaths from the pulse oximeter waveform," *Journal of Clinical Monitoring and Computing*, vol. 18, no. 5-6, pp. 309–312, 2004.

[89] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *Journal of Machine Learning Research*, pp. 1297–1322, 2010.

[90] R. Dragusin, P. Petcu, C. Lioma, B. Larsen, H. L. Jørgensen, I. J. Cox, L. K. Hansen, P. Ingwersen, and O. Winther, "Findzebra: A search engine for rare diseases," *International Journal of Medical Informatics*, vol. 82, no. 6, pp. 528–538, 2013.

[91] CrowdMed Team, "Crowdmed." https://www.crowdmed.com/, August 2015.

[92] Webicina Team, "Webicina." http://www.webicina.com/, August 2015.

[93] R. E. Schapire, "The strength of weak learnability," *Machine Learning*, vol. 5, no. 2, pp. 197–227, 1990.

[94] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[95] C. M. Bishop, *Pattern Recognition and Machine Learning*. Secaucus, NJ, USA: Springer, 2006.

[96] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[97] L. von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum, "reCAPTCHA: Human-Based Character Recognition via Web Security Measures," *Science*, vol. 321, no. 5895, pp. 1465–1468, 2008.

[98] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm," in *the 30th Annual Symposium on Foundations of Computer Science*, pp. 256–261, IEEE, 1989.

[99] S. C. Warby, S. L. Wendt, P. Welinder, E. G. Munk, O. Carrillo, H. B. Sorensen, P. Jennum, P. E. Peppard, P. Perona, and E. Mignot, "Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods," *Nature Methods*, vol. 11, no. 4, pp. 385–392, 2014.

[100] V. S. Sheng, F. Provost, and P. G. Ipeirotis, "Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, (New York, NY, USA), pp. 614–622, ACM, 2008.

[101] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Journal of the Royal Statistical Society Series C Applied Statistics*, vol. 28, no. 1, pp. 20–28, 1979.

[102] D. Spiegelhalter and P. Stovin, "An analysis of repeated biopsies following cardiac transplantation," *Statistics in Medicine*, vol. 2, no. 1, pp. 33–40, 1983.

[103] P. Smyth, U. Fayyad, M. Burl, P. Perona, and P. Baldi, "Inferring ground truth from subjective labelling of venus images," *Advances in Neural Information Processing Systems*, vol. 7, pp. 1085–1092, 1995.

[104] R. Jin and Z. Ghahramani, *Learning with multiple labels. In Advances in Neural Information Processing Systems 15*, pp. 897–904. Cambridge, MA, USA: MIT Press, 2003.

[105] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Transactions on Medical Imaging*, vol. 23, pp. 903–921, July 2004.

[106] S. R. Cholleti, S. A. Goldman, A. Blum, D. G. Politte, S. Don, K. Smith, and F. Prior, "Veritas: Combining expert opinions without labeled data," *International Journal on Artificial Intelligence Tools*, vol. 18, no. 05, pp. 633–651, 2009.

[107] R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Machine Learning*, vol. 37, no. 3, pp. 297–336, 1999.

[108] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *Computational Learning Theory*, pp. 23–37, Springer, 1995.

[109] B. Carpenter, "Multilevel Bayesian models of categorical data annotation," *Unpublished manuscript*, 2008.

[110] H. C. Kim and Z. Ghahramani, "Bayesian classifier combination," in *International Conference on Artificial Intelligence and Statistics*, pp. 619–627, 2012.

[111] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," *Advances in Neural Information Processing Systems*, vol. 22, pp. 2035–2043, 2009.

[112] Amazon, "Amazon Mechanical Turk." https://www.mturk.com/mturk/welcome, August 2016.

[113] V. Raykar, S. Yu, L. Zhao, *et al.*, "Supervised learning from multiple experts: Whom to trust when everyone lies a bit," in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 889–896, ACM, 2009.

[114] Y. Yan, R. Rosales, G. Fung, R. Subramanian, and J. Dy, "Learning from multiple annotators with varying expertise," *Machine Learning*, vol. 95, no. 3, pp. 291–327, 2014.

[115] P. Zhang, W. Cao, and Z. Obradovic, "Learning by aggregating experts and filtering novices: a solution to crowdsourcing problems in bioinformatics," *BMC Bioinformatics*, vol. 14, no. Suppl 12, p. S5, 2013.

[116] K. Audhkhasi and S. Narayanan, "A Globally-Variant Locally-Constant Model for Fusion of Labels from Multiple Diverse Experts without Using Reference Labels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 769–783, April 2013.

[117] V. E. Johnson, "On Bayesian analysis of multirater ordinal data: An application to automated essay grading," *Journal of the American Statistical Association*, vol. 91, no. 433, pp. 42–51, 1996.

[118] E. Simpson, S. J. Roberts, A. Smith, and C. Lintott, "Bayesian combination of multiple, imperfect classifiers," 2011.

[119] E. Simpson, S. J. Roberts, I. Psorakis, and A. Smith, "Dynamic Bayesian Combination of Multiple Imperfect Classifiers," in *Decision Making and Imperfection* (T. V. Guy, M. Karny, and D. Wolpert, eds.), vol. 474 of *Studies in Computational Intelligence*, pp. 1–35, Springer Berlin Heidelberg, 2013.

[120] G. Chittaranjan, O. Aran, and D. Gatica-Perez, "Exploiting observers' judgements for nonverbal group interaction analysis," in *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*, pp. 734–739, IEEE, 2011.

[121] S. Oyama, Y. Baba, Y. Sakurai, and H. Kashima, "Accurate integration of crowdsourced labels using workers' self-reported confidence scores," in *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pp. 2554–2560, AAAI Press, 2013.

[122] H. Kajino and H. Kashima, "Convex formulations of learning from crowds," *Transactions of the Japanese Society for Artificial Intelligence*, vol. 27, pp. 133–142, 2012.

[123] H. Valizadegan, Q. Nguyen, and M. Hauskrecht, "Learning classification models from multiple experts," *Journal of Biomedical Informatics*, vol. 46, no. 6, pp. 1125–1135, 2013.

[124] Q. Nguyen, H. Valizadegan, and M. Hauskrecht, "Learning classification models with soft-label information," *Journal of the American Medical Informatics Association*, vol. 21, no. 3, pp. 501–508, 2014.

[125] A. Burnap, Y. Ren, R. Gerth, G. Papazoglou, R. Gonzalez, and P. Y. Papalambros, "When crowdsourcing fails: A study of expertise on crowdsourced design evaluation," *Journal of Mechanical Design*, vol. 137, no. 3, pp. 031101–031109, 2015.

[126] P. Welinder and P. Perona, "Online crowdsourcing: Rating annotators and obtaining cost-effective labels," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pp. 25–32, 2010.

[127] T. Zhu, A. E. Johnson, J. Behar, and G. D. Clifford, "Crowd-Sourced Annotation of ECG Signals Using Contextual Information," *Annals of Biomedical Engineering*, vol. 42, no. 4, pp. 871–884, 2014.

[128] P. Welinder, S. Branson, P. Perona, and S. J. Belongie, "The Multidimensional Wisdom of Crowds," in *Advances in Neural Information Processing Systems*, pp. 2424–2432, 2010.

[129] R. W. Ouyang, L. Kaplan, P. Martin, A. Toniolo, M. Srivastava, and T. J. Norman, "Debiasing crowdsourced quantitative characteristics in local businesses and services," in *Proceedings of the 14th International Conference on Information Processing in Sensor Networks*, (New York, NY, USA), pp. 190–201, ACM, 2015.

[130] F. Xing, S. Soleimanifard, J. L. Prince, and B. A. Landman, "Statistical fusion of continuous labels: identification of cardiac landmarks," in *International Society for Optics and Photonics Medical Imaging*, pp. 7962 –796206, 2011.

[131] J. M. Wiebe, R. F. Bruce, and T. P. O'Hara, "Development and Use of a Gold-standard Data Set for Subjectivity Classifications," in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, (Stroudsburg, PA, USA), pp. 246–253, Association for Computational Linguistics, 1999.

[132] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and Fast-but is It Good?: Evaluating Non-expert Annotations for Natural Language Tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (Stroudsburg, PA, USA), pp. 254–263, Association for Computational Linguistics, 2008.

[133] O. Commowick and S. Warfield, "A Continuous STAPLE for Scalar, Vector, and Tensor Images: An Application to DTI Analysis," *IEEE Transactions on Medical Imaging*, vol. 28, pp. 838–846, June 2009.

[134] P. G. Ipeirotis, F. Provost, and J. Wang, "Quality Management on Amazon Mechanical Turk," in *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, (New York, NY, USA), pp. 64–67, ACM, 2010.

[135] Y. Baba and H. Kashima, "Statistical quality estimation for general crowdsourcing tasks," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, (New York, NY, USA), pp. 554–562, ACM, 2013.

[136] G. Cabrera, C. Miller, and J. Schneider, "Systematic labeling bias: De-biasing where everyone is wrong," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pp. 4417–4422, Aug 2014.

[137] F. Xing, A. J. Asman, J. L. Prince, and B. A. Landman, "Finding seeds for segmentation using statistical fusion," in *International Society for Optics and Photonics Medical Imaging*, pp. 8314 – 8330, 2012.

[138] A. Akhondi-Asl, A. Hans, B. Scherrer, J. Peters, and S. Warfield, "Whole brain group network analysis using network bias and variance parameters," in *9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1511–1514, May 2012.

[139] M. Nasir, B. Baucom, P. Georgiou, and S. Narayanan, "Redundancy analysis of behavioral coding for couples therapy and improved estimation of behavior from noisy annotations," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1886 –1890, April 2015.

[140] E. Kamar, A. Kapoor, and E. Horvitz, "Identifying and Accounting for Task-Dependent Bias in Crowdsourcing," in *Third AAAI Conference on Human Computation and Crowdsourcing*, 2015.

[141] S. A. Bousseljot R, Kreiseler D, "Nutzung der EKG-Signaldatenbank CARDIODAT der PTB uber das Internet," *Biomedizinische Technik*, vol. 40(1), pp. 317–318, 1995.

[142] C. M. Jarque and A. K. Bera, "A Test for Normality of Observations and Regression Residuals," *International Statistical Review / Revue Internationale de Statistique*, vol. 55, no. 2, pp. pp. 163–172, 1987.

[143] J. Willems, P. Arnaud, J. van Bemmel, P. Bourdillon, C. Brohet, S. Dalla Volta, J. Andersen, R. Degani, B. Denis, M. Demeester, and et al., "Assessment of the performance of electrocardiographic computer programs with the use of a reference data base.," *Circulation*, vol. 71(3), pp. 523 – 534, 1985.

[144] W. Karlen, M. Turner, E. Cooke, G. Dumont, and J. M. Ansermino, "Capnobase: Signal database and tools to collect, share and annotate respiratory signals," in *Annual Meeting of the Society for Technology in Anesthesia (STA)*, (West Palm Beach), 2010.

[145] B. N. Li, M. C. Dong, and M. I. Vai, "On an automatic delineator for arterial blood pressure waveforms," *Biomedical Signal Processing and Control*, vol. 5, no. 1, pp. 76 – 81, 2010.

[146] C. Orphanidou, O. Brain, J. Feldmar, S. Khan, J. Price, and L. Tarassenko, "Spectral fusion for estimating respiratory rate from the ECG," in *2009 9th International Conference on Information Technology and Applications in Biomedicine*, pp. 1–4, Nov 2009.

[147] L. Mason, *Signal Processing Methods for Non-invasive Respiration Monitoring*. PhD thesis, Department of Engineering Science, University of Oxford, 2002.

[148] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society Series B Statistical Methodology*, vol. 39, no. 1, pp. 1–38, 1977.

[149] M. R. Franz and M. Zabel, "Electrophysiological basis of QT dispersion measurements," *Progress in Cardiovascular Diseases*, vol. 42, no. 5, pp. 311–324, 2000.

[150] R. D. Berger, E. K. Kasper, K. L. Baughman, E. Marban, H. Calkins, and G. F. Tomaselli, "Beat-to-beat QT interval variability: novel evidence for repolarization lability in ischemic and nonischemic dilated cardiomyopathy.," *Circulation*, vol. 96, pp. 1557–1565, Sep 1997.

[151] C. E. Garnett, H. Zhu, M. Malik, A. A. Fossa, J. Zhang, F. Badilini, J. Li, B. Darpö, P. Sager, and I. Rodriguez, "Methodologies to characterize the QT/corrected QT interval in the presence of drug-induced heart rate changes or other autonomic effects," *American Heart Journal*, vol. 163, no. 6, pp. 912 – 930, 2012.

[152] E. Pueyo, P. Smetana, P. Laguna, and M. Malik, "Estimation of the QT/RR hysteresis lag," *Journal of Electrocardiology*, vol. 36, pp. 187–190, 2003.

[153] C. P. Lau, A. R. Freedman, S. Fleming, M. Malik, A. J. Camm, and D. E. Ward, "Hysteresis of the ventricular paced QT interval in response to abrupt changes in pacing rate," *Cardiovascular Research*, vol. 22, pp. 67–72, Jan 1988.

[154] E. Pueyo, M. Malik, and P. Laguna, "A dynamic model to characterize beat-to-beat adaptation of repolarization to heart rate changes," *Biomedical Signal Processing and Control*, vol. 3, no. 1, pp. 29 – 43, 2008.

[155] P. S. Hamilton and W. J. Tompkins, "Quantitative Investigation of QRS Detection Rules Using the MIT/BIH Arrhythmia Database," *IEEE Transactions on Biomedical Engineering*, no. 12, pp. 1157–1165, 1986.

[156] G. D. Clifford, J. Behar, Q. Li, and I. Rezek, "Signal quality indices and data fusion for determining clinical acceptability of electrocardiograms," *Physiological Measurement*, vol. 33, no. 9, pp. 1419–1433, 2012.

[157] W. Zong, G. Moody, and D. Jiang, "A robust open-source algorithm to detect onset and duration of QRS complexes," in *Computing in Cardiology Conference*, pp. 737–740, Sept 2003.

[158] G. Friesen, T. Jannett, M. Jadallah, S. Yates, S. Quint, and H. Nagle, "A comparison of the noise sensitivity of nine QRS detection algorithms," *IEEE Transactions on Biomedical Engineering*, vol. 37, pp. 85–98, Jan 1990.

[159] N. P. Hughes, *Probabalistic Models for Automated ECG Interval Analysis*. PhD thesis, University of Oxford, 2006.

[160] J. P. Couderc, C. Garnett, M. Li, R. Handzel, S. McNitt, X. Xia, S. Polonsky, and W. Zareba, "Highly Automated QT Measurement Techniques in 7 Thorough QT Studies Implemented under ICH E14 Guidelines," *Annals of Noninvasive Electrocardiology*, vol. 16, no. 1, pp. 13–24, 2011.

[161] W. Andrew, V. Michael, D. Jeff, G. M. Nair, C. Plater-Zyberk, L. Griffith, J. Ma, C. Zachos, and M. L. Sivilotti, "Variability of QT Interval Measurements in Opioid-Dependent Patients on Methadone," *Canadian Journal of Addiction Medicine*, vol. 2, pp. 10–16, 2014.

[162] M. Malik, P. Fãrbom, V. Batchvarov, K. Hnatkova, and A. J. Camm, "Relation between QT and RR intervals is highly individual among healthy subjects: implications for heart rate correction of the QT interval," *Heart*, vol. 87, no. 3, pp. 220–228, 2002.

[163] I. Goldenberg, A. J. Moss, W. Zareba, *et al.*, "QT interval: how to measure it and what is "normal"," *Journal of Cardiovascular Electrophysiology*, vol. 17, no. 3, pp. 333–336, 2006.

[164] R. A. Fisher and L. H. C. Tippett, "Limiting forms of the frequency distribution of the largest or smallest member of a sample," in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 24, pp. 180–190, Cambridge Univ Press, 1928.

[165] D. A. Clifton, *Novelty Detection with Extreme Value Theory in Jet Engine Vibration Data*. PhD thesis, University of Oxford, 2009.

[166] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. Boca Raton, FL, USA: Chapman & Hall/CRC, 1996.

[167] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 6, pp. 721–741, 1984.

[168] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*. Englewood Cliffs, NJ, USA: Prentice Hall, 3rd ed., 1994.

[169] J. Geweke, "Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments," in *Bayesian Statistics*, pp. 169–193, University Press, 1992.

[170] A. E. Raftery and S. Lewis, "How Many Iterations in the Gibbs Sampler?," in *Bayesian Statistics 4*, pp. 763–773, Oxford University Press, 1992.

[171] A. E. Raftery and S. M. Lewis, "The Number of Iterations, Convergence Diagnostics and Generic Metropolis Algorithms," in *Practical Markov Chain Monte Carlo* (W. Gilks, D. Spiegelhalter, and S. Richardson, eds.), pp. 115–130, Chapman and Hall, 1995.

[172] A. Gelman and D. B. Rubin, "Inference from iterative simulation using multiple sequences," *Statistical Science*, pp. 457–472, 1992.

[173] A. Gelman, "Inference and Monitoring Covergence," in *Markov Chain Monte Carlo in Practice* (W. Gilks, S. Richardson, and D. Spiegelhalter, eds.), ch. 8, pp. 131–143, Boca Raton, FL, USA: Chapman & Hall/CRC, 1996.

[174] M. K. Cowles and B. P. Carlin, "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review," *Journal of the American Statistical Association*, vol. 91, no. 434, pp. 883–904, 1996.

[175] J. P. LeSage, "Applied econometrics using MATLAB," *Manuscript, Department of Economics, University of Toronto*, 1999.

[176] W. L. Martinez and A. R. Martinez, *Computational statistics handbook with MATLAB*. Boca Raton, FL, USA: Chapman and Hall/CRC, 2007.

[177] M. K. Cowles, G. O. Roberts, and J. S. Rosenthal, "Possible biases induced by MCMC convergence diagnostics," *Journal of Statistical Computation and Simulation*, vol. 64, no. 1, pp. 87–104, 1999.

[178] K. Sahlin, "Estimating convergence of Markov chain Monte Carlo simulations," *Stockholm University, MSc Thesis*, 2011.

[179] A. Breakell and C. Townsend-Rose, "The clinical evaluation of the respi-check mask: a new oxygen mask incorporating a breathing indicator," *Emergency Medicine Journal*, vol. 18, no. 5, pp. 366–369, 2001.

[180] G. Drummond, A. Bates, J. Mann, and D. Arvind, "Validation of a new non-invasive automatic monitor of respiratory rate for postoperative subjects," *British Journal of Anaesthesia*, p. aer153, 2011.

[181] F. L. Wauthier and M. I. Jordan, "Bayesian Bias Mitigation for Crowdsourcing," in *Advances in Neural Information Processing Systems 24* (J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, eds.), pp. 1800–1808, 2011.

[182] P. Wais, S. Lingamneni, D. Cook, J. Fennell, B. Goldenberg, D. Lubarov, D. Marin, and H. Simons, "Towards Building a High-Quality Workforce with Mechanical Turk," in *NIPS Workshop on Computational Social Science and the Wisdom of Crowds*, Dec. 2010.

[183] M. A. Pimentel, P. H. Charlton, and D. A. Clifton, "Probabilistic Estimation of Respiratory Rate from Wearable Sensors," in *Wearable Electronics Sensors: For Safe and Healthy Living* (S. C. Mukhopadhyay, ed.), vol. 15 of *Smart Sensors, Measurement and Instrumentation*, pp. 241–262, Switzerland: Springer International Publishing, 2015.

[184] A. Aboukhalil, L. Nielsen, M. Saeed, R. G. Mark, and G. D. Clifford, "Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform ," *Journal of Biomedical Informatics*, vol. 41, no. 3, pp. 442 – 451, 2008. Computerized Decision Support for Critical and Emergency Care.

[185] Q. Li, R. G. Mark, and G. D. Clifford, "Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter," *Physiological Measurement*, vol. 29, pp. 15–32, Jan 2008.

[186] Tsanas, Athanasios and Zañartu, Matías and Little, Max A. and Fox, Cynthia and Ramig, Lorraine O. and Clifford, G. D., "Robust fundamental frequency estimation in sustained vowels: Detailed algorithmic comparisons and information fusion with adaptive Kalman filtering," *The Journal of the Acoustical Society of America*, vol. 135, no. 5, pp. 2885–2901, 2014.

[187] M. Pimentel, M. Santos, D. Springer, and G. D. Clifford, "Hidden semi-Markov model-based heartbeat detection using multimodal data and signal quality indices," in *Computing in Cardiology Conference*, pp. 553–556, Sept 2014.

[188] M. A. F. Pimentel, M. D. Santos, D. B. Springer, and G. D. Clifford, "Heart beat detection in multimodal physiological data using a hidden semi-Markov model and signal quality indices," *Physiological Measurement*, vol. 36, no. 8, pp. 1717–1727, 2015.

[189] M. Vollmer, "Robust detection of heart beats using dynamic thresholds and moving windows," in *Computing in Cardiology Conference*, pp. 569–572, Sept 2014.

[190] L. Johannesen, J. Vicente, C. Scully, L. Galeotti, and D. Strauss, "Robust algorithm to locate heart beats from multiple physiological waveforms," in *Computing in Cardiology Conference*, pp. 277–280, Sept 2014.

[191] A. E. Johnson, J. Behar, F. Andreotti, G. D. Clifford, and J. Oster, "Multimodal heart beat detection using signal quality indices," *Physiological Measurement*, vol. 36, no. 8, pp. 1665–1677, 2015.

[192] J. Behar, J. Oster, Q. Li, and G. D. Clifford, "ECG Signal Quality During Arrhythmia and Its Application to False Alarm Reduction," *IEEE Transactions on Biomedical Engineering*, vol. 60, pp. 1660–1666, June 2013.

[193] J. Behar, A. Johnson, G. D. Clifford, and J. Oster, "A Comparison of Single Channel Fetal ECG Extraction Methods," *Annals of Biomedical Engineering*, vol. 42, no. 6, pp. 1340–1353, 2014.

[194] J. Behar, J. Oster, and G. D. Clifford, "Combining and benchmarking methods of foetal ECG extraction without maternal or scalp electrode data," *Physiological Measurement*, vol. 35, no. 8, pp. 1569–1589, 2014.

[195] W. Zong, T. Heldt, G. Moody, and R. Mark, "An open-source algorithm to detect onset of arterial blood pressure pulses," in *Computing in Cardiology Conference*, pp. 259–262, Sept 2003.

[196] D. A. Birrenkott, M. A. F. Pimentel, P. J. Watkinson, and D. A. Clifton, "Robust estimation of respiratory rate via ECG- and PPG-derived respiratory quality indices," 2016.

[197] J. Barnard, R. McCulloch, and X. L. Meng, "Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage," *Statistica Sinica*, vol. 10, no. 4, pp. 1281–1312, 2000.

[198] K. Pearson, "Note on regression and inheritance in the case of two parents," *Proceedings of the Royal Society of London*, pp. 240–242, 1895.

[199] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*. Boca Raton, FL, USA: Chapman & Hall/CRC, third ed., 2013.

[200] D. Lewandowski, D. Kurowicka, and H. Joe, "Generating random correlation matrices based on vines and extended onion method," *Journal of Multivariate Analysis*, vol. 100, no. 9, pp. 1989 – 2001, 2009.

[201] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[202] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of State Calculations by Fast Computing Machines," *The Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[203] G. O. Roberts, A. Gelman, W. R. Gilks, *et al.*, "Weak convergence and optimal scaling of random walk Metropolis algorithms," *The Annals of Applied Probability*, vol. 7, no. 1, pp. 110–120, 1997.

[204] F. Dominici, G. Parmigiani, and M. Clyde, "Conjugate analysis of multivariate normal data with incomplete observations," *Canadian Journal of Statistics*, vol. 28, no. 3, pp. 533–550, 2000.

[205] J. Behar, T. Zhu, J. Oster, A. Niksch, D. Mah, T. Chun, J. Greenberg, C. Tanner, J. Harrop, A. Van Esbroeck, *et al.*, "Evaluation of the fetal qt interval using non-invasive foetal ecg technology," in *SMFM-34th Annual Meeting-The Pregnancy Meeting*, Feb 2014.

[206] G. D. Clifford, C. Arteta, T. Zhu, M. Pimentel, M. Santos, J. Domingos, M. Maraci, J. Behar, and J. Oster, "A scalable mhealth system for noncommunicable disease management," in *IEEE Global Humanitarian Technology Conference (GHTC)*, pp. 41–48, Oct 2014.

[207] I. Silva, J. Behar, R. Sameni, T. Zhu, J. Oster, G. D. Clifford, and G. B. Moody, "Noninvasive fetal ECG: The PhysioNet/Computing in Cardiology Challenge 2013," in *Computing in Cardiology Conference*, pp. 149–152, Sept 2013.

[208] T. Zhu, A. E. W. Johnson, J. Behar, and G. Clifford, "Bayesian voting of multiple annotators for improved QT interval estimation," in *Computing in Cardiology Conference*, pp. 659–662, Sept 2013.

[209] T. Zhu, M. Osipov, T. Papastylianou, J. Oster, D. A. Clifton, and G. D. Clifford, "An intelligent cardiac health monitoring and review system," in *Appropriate Healthcare Technologies for Low Resource Settings (AHT 2014)*, pp. 1–4, Sept 2014.

[210] T. Zhu, N. Dunkley, J. Behar, D. A. Clifton, and G. D. Clifford, "Fusing Continuous-Valued Medical Labels Using a Bayesian Model," *Annals of Biomedical Engineering*, vol. 43, no. 12, pp. 2892–2902, 2015.

[211] T. Zhu, M. A. Pimentel, G. D. Clifford, and D. A. Clifton, "Bayesian fusion of algorithms for the robust estimation of respiratory rate from the photoplethysmogram," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pp. 6138–6141, IEEE, 2015.

[212] J. Behar, T. Zhu, J. Oster, A. Niksch, D. Y. Mah, T. Chun, J. Greenberg, C. Tanner, J. Harrop, R. Sameni, J. Ward, A. Wolfberg, and G. D. Clifford, "Evaluation of the fetal QT interval using non-invasive fetal ECG technology," *Physiological Measurement*, 2016.

[213] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Computation*, vol. 3, no. 1, pp. 79–87, 1991.

[214] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural Computation*, vol. 6, no. 2, pp. 181–214, 1994.

[215] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, pp. 174–188, Feb 2002.

[216] J. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," *International Computer Science Institute*, vol. 4, no. 510, p. 126, 1998.

[217] K. P. Murphy, "Conjugate Bayesian analysis of the Gaussian distribution," vol. 1, no. $2\sigma 2$, p. 16, 2007.

[218] K. P. Murphy, *Machine Learning: a Probabilistic Perspective*. 2012.

[219] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," tech. rep., Technical University of Denmark, 2012.

Appendix A

Expectation-Maximisation

A.1 Expectation-Maximisation of Maximum Likelihood

The Expectation-Maximisation (EM) method is an iterative algorithm for parameter estimation by finding a local maximum of the likelihood of the observed data. Denote the set of observed data by $\mathbf{D}$, the latent data or missing values by $\mathbf{Z}$, and a set of model parameters as $\boldsymbol{\theta}$. The complete dataset refers to $\{\mathbf{D}, \mathbf{Z}\}$ while the actual observed $\mathbf{D}$ is considered as the *incomplete* dataset. The likelihood function of the complete dataset is as follows:

$$\mathscr{L}(\boldsymbol{\theta}) = p(\mathbf{D}, \mathbf{Z} \mid \boldsymbol{\theta}). \tag{A.1}$$

The goal is to solve the parameters in $\boldsymbol{\theta}$ by maximising the log-likelihood of the complete dataset, i.e., $\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \{\log p(\mathbf{D}, \mathbf{Z} \mid \boldsymbol{\theta})\}$. This is equivalent to finding the expectation of the $\log p(\mathbf{D}, \mathbf{Z} \mid \boldsymbol{\theta})$ with respect to the unknown $\mathbf{Z}$, given the current estimation of the parameters denoted as $\boldsymbol{\theta}_{old}$ and the observed dataset $\mathbf{D}$:

$$\mathscr{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}_{old}) = \mathbb{E}[\log p(\mathbf{D}, \mathbf{Z} \mid \boldsymbol{\theta}) \mid \mathbf{D}, \boldsymbol{\theta}_{old}]. \tag{A.2}$$

where $\mathscr{Q}(\cdot)$ refers to a function of variable $\boldsymbol{\theta}$ given some fixed values of $\boldsymbol{\theta}_{old}$ and $\mathbf{D}$. The expectation can also been seen as the marginalisation over all possible values in $\mathbf{Z}$ as:

$$\mathbb{E}\left[\log p\left(\mathbf{D},\mathbf{Z}\mid\boldsymbol{\theta}\right)\mid\mathbf{D},\boldsymbol{\theta}_{old}\right]=\int_{\mathbf{Z}}p\left(\mathbf{Z}\mid\mathbf{D},\boldsymbol{\theta}_{old}\right)\log p\left(\mathbf{D},\mathbf{Z}\mid\boldsymbol{\theta}\right)d\mathbf{Z}. \tag{A.3}$$

Furthermore, the joint probability between $\mathbf{D}$ and $\mathbf{Z}$ can be expressed as [216]:

$$p\left(\mathbf{D},\mathbf{Z}\mid\boldsymbol{\theta}\right)=p\left(\mathbf{D}\mid\mathbf{Z},\boldsymbol{\theta}\right)p\left(\mathbf{Z}\mid\boldsymbol{\theta}\right). \tag{A.4}$$

By substituting equation (A.4) into the expectation of the complete data log-likelihood, we have

$$\begin{aligned}\mathscr{Q}(\boldsymbol{\theta},\boldsymbol{\theta}_{old})&=\int_{\mathbf{Z}}p\left(\mathbf{Z}\mid\mathbf{D},\boldsymbol{\theta}_{old}\right)\log p\left(\mathbf{D},\mathbf{Z}\mid\boldsymbol{\theta}\right)d\mathbf{Z}\\&=\int_{\mathbf{Z}}p\left(\mathbf{Z}\mid\mathbf{D},\boldsymbol{\theta}_{old}\right)\log\left\{p\left(\mathbf{D}\mid\mathbf{Z},\boldsymbol{\theta}\right)p\left(\mathbf{Z}\mid\boldsymbol{\theta}\right)\right\}d\mathbf{Z}\\&=\int_{\mathbf{Z}}p\left(\mathbf{Z}\mid\mathbf{D},\boldsymbol{\theta}_{old}\right)\left\{\log p\left(\mathbf{D}\mid\mathbf{Z},\boldsymbol{\theta}\right)+\log p\left(\mathbf{Z}\mid\boldsymbol{\theta}\right)\right\}d\mathbf{Z}.\end{aligned} \tag{A.5}$$

However, the values of $\mathbf{Z}$ are not observed *a priori* and can only be obtained from its posterior distribution $p\left(\mathbf{Z}\mid\mathbf{D},\boldsymbol{\theta}\right)$. Therefore, the EM algorithm proposed by Dempster *et al.* [148] can be used to evaluate the expectation of the complete data log-likelihood in two steps:
(i) The Expectation (E) step evaluates the $p\left(\mathbf{Z}\mid\mathbf{D},\boldsymbol{\theta}_{old}\right)$ given some initialised values of $\boldsymbol{\theta}_{old}$; Note that $p\left(\mathbf{Z}\mid\mathbf{D},\boldsymbol{\theta}\right)$ can be estimated using Bayes' Theorem:

$$p\left(\mathbf{Z}\mid\mathbf{D},\boldsymbol{\theta}\right)\propto p\left(\mathbf{D}\mid\mathbf{Z},\boldsymbol{\theta}\right)p\left(\mathbf{Z}\mid\boldsymbol{\theta}\right). \tag{A.6}$$

(ii) The Maximisation (M) step determines the updated parameter values in $\boldsymbol{\theta}$ that maximise $\mathscr{Q}(\cdot)$, given our belief of $\mathbf{Z}$ in the E step as:

$$\boldsymbol{\theta}_{new}=\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\mathscr{Q}\left(\boldsymbol{\theta},\boldsymbol{\theta}_{old}\right). \tag{A.7}$$

Both E and M steps iterates by updating $\boldsymbol{\theta}_{old}\leftarrow\boldsymbol{\theta}_{new}$ until convergence.

In some case where we would like to compute the most probable estimate of the posterior of $\mathbf{Z}$, a *Hard EM* can be considered by maximising the expectation of the complete log-likelihood on the most probable belief of $\mathbf{Z}$, i.e. $\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\{\log p\left(\mathbf{Z}\mid\mathbf{D},\boldsymbol{\theta}\right)\}$. The log-likelihood

of $\mathbf{Z}$ can now be written as:

$$\log p\left(\mathbf{Z} \mid \mathbf{D}, \boldsymbol{\theta}\right) \propto \log p\left(\mathbf{D} \mid \mathbf{Z}, \boldsymbol{\theta}\right) + \log p\left(\mathbf{Z} \mid \boldsymbol{\theta}\right). \tag{A.8}$$

The most probable latent variables in $\mathbf{Z}$ (denoted as $\hat{\mathbf{z}}$) thereby can be derived by estimating the gradient of the above log-likelihood. Given the estimate of $\hat{\mathbf{z}}$, the expectation of the complete data log-likelihood can be written as:

$$\mathscr{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}_{old}) = \log p\left(\mathbf{D} \mid \hat{\mathbf{z}}, \boldsymbol{\theta}\right) + \log p\left(\hat{\mathbf{z}} \mid \boldsymbol{\theta}\right). \tag{A.9}$$

In the equation (4.10) provided by the EM-R approach, the ground truth, $\mathbf{z}$, for N records, are the hidden variables and its log-posterior can be expressed as:

$$\begin{aligned}
\log p\left(\mathbf{z} \mid \mathbf{D}, \boldsymbol{\theta}\right) &\propto \log p\left(\mathbf{D} \mid \mathbf{z}, \boldsymbol{\theta}\right) + \log p\left(\mathbf{z} \mid \boldsymbol{\theta}\right) \\
&= \log \left\{ \prod_{i=1}^{N} \prod_{j=1}^{R} \mathscr{N}\left(y_i^j \mid z_i, 1/\lambda^j\right) \right\} + \log(1) \\
&= -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{R} \log\left(\frac{2\pi}{\lambda^j}\right) + \left(y_i^j - \mathbb{E}[z_i]\right)^2 \lambda^j.
\end{aligned} \tag{A.10}$$

Let $\hat{z}_i = \mathbb{E}[z_i]$ and it can be derived by estimating the gradient of the log-posterior to zero as:

$$\frac{\mathrm{d} \log p\left(\mathbf{z} \mid \mathbf{D}\right)}{\mathrm{d}\hat{z}_i} = \sum_{j=1}^{R} \lambda^j \left(y_i^j - \hat{z}_i\right) = 0.$$

to obtain the estimate of $\hat{\mathbf{z}}$ for N recordings in the E step:

$$\hat{\mathbf{z}} = \frac{\sum_{j=1}^{R} \lambda^j \mathbf{y}^j}{\sum_{j=1}^{R} \lambda^j}. \tag{A.11}$$

The M step then maximises the $\mathscr{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}_{old})$ with respect to $\mathbf{w}$ and λ^j given the $\hat{z}_i$ respectively:

$$\frac{\mathrm{d}\mathscr{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}_{old})}{\mathrm{d}\mathbf{w}} = \sum_{i=1}^{N} \sum_{j=1}^{R} \lambda^j \left(y_i^j \mathbf{x}_i - \mathbf{x}_i^\intercal \mathbf{w} \mathbf{x}_i\right), \qquad \frac{\mathrm{d}\mathscr{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}_{old})}{\mathrm{d}\lambda^j} = -\frac{1}{2} \sum_{i=1}^{N} \left[\left(y_i^j - \mathbf{x}_i^\intercal \mathbf{w}\right)^2 - \frac{1}{\lambda^j} \right].$$

By setting the derivatives to zero, we have:

$$\mathbf{w} = \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^{\intercal} \right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i \hat{z}_i. \tag{A.12}$$

$$\frac{1}{\lambda^j} = \frac{1}{N} \sum_{i=1}^{N} \left(y_i^j - \mathbf{x}_i^{\intercal} \mathbf{w} \right)^2. \tag{A.13}$$

Note that the values in $\mathbf{z}$ are assumed to be drawn from a univariate normal distribution throughout the thesis, thereby solving the expectation of the complete data log-likelihood using the *Hard EM* seems to be the most plausible and effective approach to ensure fast convergence rate.

A.2 Expectation-Maximisation of Maximum-a-Posteriori

The EM algorithm can be applied to maximum-a-posteriori (MAP) in a similar manner with an additional prior of $\boldsymbol{\theta}$ (denoted as $p(\boldsymbol{\theta})$) that is applied over the parameters. When solving MAP, the E step remains the same as in the maximum likelihood case, whereas the M step would now maximise both the $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}_{old})$ and the $p(\boldsymbol{\theta})$, i.e., $\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \{ \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}_{old}) + \log p(\boldsymbol{\theta}) \}$.

In the equation (5.15), the latent values are the ground truth (i.e., $\mathbf{z}$) for all recordings in the BCLA model, and its log-posterior can be expressed as:

$$\begin{aligned}
&\log p(\mathbf{z} \mid \mathbf{D}, \boldsymbol{\theta}) \\
&\propto \log p(\mathbf{D} \mid \mathbf{z}, \boldsymbol{\theta}) + \log p(\mathbf{z} \mid \boldsymbol{\theta}) \\
&= \log \left\{ \prod_{i=1}^{N} \prod_{j=1}^{R} \mathcal{N}\left(y_i^j \mid z_i + \phi^j, 1/\lambda^j \right) \right\} + \log \left\{ \prod_{i=1}^{N} \mathcal{N}\left(z_i \mid \mathbf{x}_i^{\intercal} \mathbf{w}, 1/b \right) \right\} \\
&= -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{R} \left[\log\left(\frac{2\pi}{\lambda^j} \right) + \left(y_i^j - \phi^j - \mathbb{E}[z_i] \right)^2 \lambda^j \right] - \frac{1}{2} \sum_{i=1}^{N} \left[\log\left(\frac{2\pi}{b} \right) + \left(\mathbb{E}[z_i] - \mathbf{x}_i^{\intercal} \mathbf{w} \right)^2 b \right].
\end{aligned} \tag{A.14}$$

Again let $\hat{z}_i = \mathbb{E}[z_i]$, and it can be estimated using the *Hard EM* by maximising its log-posterior as:

$$\frac{\mathrm{d}\log p\left(\boldsymbol{\theta} \mid \mathbf{D}\right)}{\mathrm{d}\hat{z}_i} = \sum_{j=1}^{R} \left(y_i^j - \phi^j - \hat{z}_i\right) \lambda^j - b\left(\hat{z}_i - \mathbf{x}_i^\intercal \mathbf{w}\right). \tag{A.15}$$

By equating the derivative to zero, we get

$$\hat{z}_i = \frac{\sum_{j=1}^{R} \left(y_i^j - \phi^j\right) \lambda^j + \left(\mathbf{x}_i^\intercal \mathbf{w}\right) b}{\sum_{j=1}^{R} \lambda^j + b}. \tag{A.16}$$

Given the estimate of $\hat{\mathbf{z}}$, the expectation of the complete data log-likelihood can be written as:

$$\mathscr{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}_{old}) = \log p\left(\mathbf{D} \mid \hat{\mathbf{z}}, \boldsymbol{\theta}\right) + \log p\left(\hat{\mathbf{z}} \mid \boldsymbol{\theta}\right) + \log p\left(\boldsymbol{\theta}\right) \tag{A.17}$$

The model parameters, $\mathbf{w}$, $\boldsymbol{\phi}$, α_ϕ, b, and $\boldsymbol{\lambda}$ can be estimated by maximising the $\mathscr{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}_{old})$ as:

$$\frac{1}{\lambda^j} = \frac{1}{N + 2\left(k_\lambda - 1\right)} \left[\sum_{i=1}^{N} \left(y_i^j - \phi^j - \hat{z}_i\right)^2 + \frac{2}{\vartheta_\lambda}\right]. \tag{A.18}$$

$$\mathbf{w} = \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^\intercal\right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i \hat{z}_i. \tag{A.19}$$

$$\phi^j = \frac{1}{N + \frac{\alpha_\phi}{\lambda^j}} \left[\sum_{i=1}^{N} \left(y_i^j - \hat{z}_i\right) + \mu_\phi \left(\frac{\alpha_\phi}{\lambda^j}\right)\right]. \tag{A.20}$$

$$\frac{1}{\alpha_\phi} = \frac{1}{R + 2\left(k_\alpha - 1\right)} \left[\sum_{j=1}^{R} \left(\phi^j - \mu_\phi\right)^2 + \frac{2}{\vartheta_\alpha}\right]. \tag{A.21}$$

$$\frac{1}{b} = \frac{1}{N + 2(k_b - 1)} \left[\sum_{i=1}^{N} \left(\hat{z}_i - \mathbf{x}_i^\intercal \mathbf{w}\right)^2 + \frac{2}{\vartheta_b}\right]. \tag{A.22}$$

Appendix B

Hyperparameter of the Conjugate Prior

B.1 Univariate Gaussian

In this thesis, a univariate Gaussian distribution describes the relationship between the annotator and the ground truth models. The parameters are estimated using Bayesian inference and conjugate priors. Here the conjugate priors are considered for data drawn from a Gaussian distribution, $\mathcal{N}(\mu,\lambda)$ as described in [95, 217] and their hyperparameters are estimated in the following sections. Suppose that there is $D=\{x_1,...,x_N\}$, the dataset being observed from N records. Assuming that the observations are independent, and the likelihood of the dataset given the parameters, mean (denoted as μ) and precision (i.e., inverse variance as denoted as λ), can be expressed as:

$$\begin{aligned} p(D \mid \mu,\lambda) &= \prod_{i=1}^{N} p(x_i \mid \mu,\lambda) \\ &= \prod_{i=1}^{N} \left(\frac{\lambda}{2\pi}\right)^{1/2} \exp\left\{-\frac{\lambda}{2}(x_i-\mu)^2\right\} \\ &= \left(\frac{\lambda}{2\pi}\right)^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{i=1}^{N}(x_i-\mu)^2\right\}. \end{aligned} \tag{B.1}$$

Further assume the empirical mean and precision of the dataset are as the following:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad \frac{1}{\gamma} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2.$$

The terms within the exponent in equation (B.1) can be re-written by substituting the above equations as:

$$\begin{aligned}
\sum_{i=1}^{N}(x_i-\mu)^2 &= \sum_{i=1}^{N}[(x_i-\bar{x})-(\mu-\bar{x})]^2 \\
&= \sum_{i=1}^{N}(x_i-\bar{x})^2 + \sum_{i=1}^{N}(\bar{x}-\mu)^2 - 2\sum_{i=1}^{N}(x_i-\bar{x})(\mu-\bar{x}) \\
&= \frac{N}{\gamma} + N(\bar{x}-\mu)^2 - 2(\mu-\bar{x})\left(\sum_{i=1}^{N}x_i - N\bar{x}\right) \\
&= \frac{N}{\gamma} + N(\bar{x}-\mu)^2.
\end{aligned}$$

The likelihood in equation (B.1) can be described using the the empirical mean and precision as:

$$\begin{aligned}
p(D \mid \mu, \lambda) &= \left(\frac{\lambda}{2\pi}\right)^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{i=1}^{N}(x_i-\mu)^2\right\} \\
&= \left(\frac{1}{2\pi}\right)^{N/2} (\lambda)^{N/2} \exp\left\{-\frac{\lambda}{2}\left[\frac{N}{\gamma} + N(\bar{x}-\mu)^2\right]\right\} \\
&\propto (\lambda)^{N/2} \exp\left(-\frac{N\lambda}{2\gamma}\right) \exp\left\{-\frac{N\lambda}{2}(\bar{x}-\mu)^2\right\}. \qquad \text{(B.2)}
\end{aligned}$$

B.1.1 Inferring the Posterior distribution of the Mean with Known Precision

Assuming that the precision is known and it is a constant, the likelihood of the dataset given μ can be written as:

$$p(D \mid \mu) \propto \exp\left\{-\frac{N\lambda}{2}(\bar{x}-\mu)^2\right\} \propto \mathcal{N}(\bar{x} \mid \mu, N\lambda), \qquad \text{(B.3)}$$

where the likelihood function takes the form of the exponential of a quadratic form in μ. The conjugate prior of μ also has the same form as:

$$p(\mu) \propto \exp\left\{-\frac{\lambda_0}{2}(\mu-\mu_0)^2\right\} \propto \mathcal{N}(\mu \mid \mu_0, \lambda_0). \tag{B.4}$$

The conjugate prior is chosen to have the same form as the likelihood function to ensure that the posterior will be a product of the two exponentials of quadratic functions of μ and therefore will be a Gaussian. The posterior of μ in equation (B.1) can be inferred using Bayes' theorem as:

$$\begin{aligned}
p(\mu \mid D) \propto & p(D \mid \mu)\, p(\mu) \\
\propto & \exp\left\{-\frac{N\lambda}{2}(\bar{x}-\mu)^2\right\} \exp\left\{-\frac{\lambda_0}{2}(\mu-\mu_0)^2\right\} \\
= & \exp\left\{-\frac{N\lambda}{2}\left(\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)\right)^2\right\} \exp\left\{-\frac{\lambda_0}{2}(\mu-\mu_0)^2\right\} \\
= & \exp\left\{-\frac{\lambda\left(\sum_{i=1}^{N} x_i\right)^2}{2N} + \mu\lambda\sum_{i=1}^{N} x_i - \frac{N\lambda\mu^2}{2} - \frac{\lambda_0\mu^2}{2} + \lambda_0\mu\mu_0 - \frac{\lambda_0{\mu_0}^2}{2}\right\} \\
= & \exp\left\{-\frac{\mu^2}{2}(\lambda_0+N\lambda) + \mu\left(\mu_0\lambda_0 + \lambda\sum_{i=1}^{N} x_i\right) - \left(\frac{\lambda_0\mu_0^2}{2} + \frac{\lambda\left(\sum_{i=1}^{N} x_i\right)^2}{2N}\right)\right\}.
\end{aligned} \tag{B.5}$$

As the product of two Gaussian produces a Gaussian, the posterior of μ also has a Gaussian form as:

$$\begin{aligned}
p(\mu \mid D) = & \mathcal{N}(\mu_N, \lambda_N) \\
\propto & \exp\left\{-\frac{\lambda_N}{2}(\mu-\mu_N)^2\right\} \\
= & \exp\left\{-\frac{\lambda_N}{2}(\mu^2 - 2\mu\mu_N + \mu_N^2)\right\}.
\end{aligned} \tag{B.6}$$

By matching the coefficients of μ^2 and μ from equation (B.5) and (B.6), the parameters (i.e., μ_N and λ_N) of the posterior of μ can be derived as:

$$\begin{aligned}
-\frac{\lambda_N\mu^2}{2} &= -\frac{\mu^2}{2}(\lambda_0+N\lambda) \\
\lambda_N &= \lambda_0 + N\lambda.
\end{aligned} \tag{B.7}$$

$$\mu(\lambda_N \mu_N) = \mu\left(\lambda_0\mu_0 + \lambda\sum_{i=1}^{N} x_i\right)$$
$$\mu_N = \frac{1}{\lambda_N}\left(\lambda_0\mu_0 + \lambda\sum_{i=1}^{N} x_i\right). \tag{B.8}$$

B.1.2 Inferring the Posterior distribution of the Precision with Known Mean

Here we assume that the mean of the Gaussian distribution over the dataset in equation (B.1) is known, and would like to infer the precision value. The likelihood of the dataset given λ can be written as:

$$p(D \mid \lambda) = \left(\frac{\lambda}{2\pi}\right)^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{i=1}^{N}(x_i-\mu)^2\right\}$$
$$\propto (\lambda)^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{i=1}^{N}(x_i-\mu)^2\right\}. \tag{B.9}$$

The choice of the conjugate prior is considered to be a Gamma distribution (denoted as Gamma) [95] as it is proportional to the λ in the equation (B.9) and can be described as:

$$p(\lambda) = \text{Gamma}(\lambda \mid k_0, \vartheta_0) = \frac{\lambda^{(k_0-1)}}{\Gamma(k_0)\vartheta^{k_0}} \exp\left(-\frac{\lambda}{\vartheta_0}\right), \tag{B.10}$$

where k_0 is the shape of the distribution and ϑ_0 is the scale of the distribution, and the $\Gamma(\cdot)$ is a gamma function. The posterior of λ in equation (B.1) can be inferred using Bayes' theorem as:

$$p(\lambda \mid D) \propto p(D \mid \lambda)\, p(\lambda)$$
$$\propto (\lambda)^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{i=1}^{N}(x_i-\mu)^2\right\} \frac{\lambda^{(k_0-1)}}{\Gamma(k_0)\vartheta^{k_0}} \exp\left(-\frac{\lambda}{\vartheta_0}\right)$$
$$\propto (\lambda)^{N/2}\, \lambda^{(k_0-1)} \exp\left\{-\frac{\lambda}{2}\sum_{i=1}^{N}(x_i-\mu)^2 - \frac{\lambda}{\vartheta_0}\right\}. \tag{B.11}$$

The posterior of λ is also a Gamma distribution in the form as Gamma$(\lambda \mid k_N, \vartheta_N)$. By matching the power and the coefficients of λ in the exponent term from equation (B.10) and (B.11), the parameters of the posterior of λ can be derived as:

$$\begin{aligned} \lambda^{(k_N-1)} &= (\lambda)^{N/2}\lambda^{(k_0-1)} \\ k_N &= k_0 + \frac{N}{2}. \end{aligned} \tag{B.12}$$

$$\begin{aligned} -\frac{\lambda}{\vartheta_N} &= -\frac{\lambda}{2}\sum_{i=1}^{N}(x_i-\mu)^2 - \frac{\lambda}{\vartheta_0} \\ \frac{1}{\vartheta_N} &= \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^2 + \frac{1}{\vartheta_0}. \end{aligned} \tag{B.13}$$

B.2 Multivariate Gaussian

B.2.1 Inferring the Posterior distribution with Unknown Covariance and Mean

Suppose that there is $\mathbf{D} = \{\mathbf{x}_{i=1}^{\intercal}, ..., \mathbf{x}_{i=N}^{\intercal}\}$, the dataset being observed from N records, where $\mathbf{x}_i$ is a column feature vector for the ith record containing d values. Assuming that the observations are independent, and the likelihood of the dataset given the parameters, D dimensional mean vector (denoted as $\boldsymbol{\mu}$) and covariance matrix (i.e., inverse precision matrix with D by D dimension as denoted as $\boldsymbol{\Sigma}$), can be expressed as:

$$\begin{aligned} p(\mathbf{D} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \prod_{i=1}^{N} p(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= (2\pi)^{\frac{-ND}{2}} |\boldsymbol{\Sigma}|^{\frac{-N}{2}} \exp\left\{ -\frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) \right\}. \end{aligned} \tag{B.14}$$

It can be re-written in the following form as [218]:

$$
\begin{aligned}
&p(\mathbf{D} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&\quad = (2\pi)^{\frac{-ND}{2}} |\boldsymbol{\Sigma}|^{\frac{-N}{2}} \exp\left\{ -\frac{N}{2} (\boldsymbol{\mu} - \bar{\mathbf{x}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \bar{\mathbf{x}}) \right\} \exp\left\{ -\frac{1}{2} tr\left(\boldsymbol{\Sigma}^{-1} \mathbf{S}_{\bar{x}} \right) \right\}, \quad \text{(B.15)}
\end{aligned}
$$

where

$$
\begin{aligned}
&\sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) = tr\left(\boldsymbol{\Sigma}^{-1} \mathbf{S}_{\bar{x}} \right) + (\boldsymbol{\mu} - \bar{\mathbf{x}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \bar{\mathbf{x}}). \\
&\mathbf{S}_{\bar{x}} = \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T.
\end{aligned}
$$

As both the covariance and the mean are unknown, the conjugate prior for modelling the multivariate normal distribution takes the form of a normal-inverse-wishart (denoted as NIW) distribution, which can be written by definition as follows:

$$
\begin{aligned}
&\text{NIW}(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \mathbf{m}_0, k_0, v_0, \mathbf{S}_0) \\
&\triangleq \text{N}\left(\boldsymbol{\mu} \mid \mathbf{m}_0, \frac{\boldsymbol{\Sigma}}{k_0} \right) \times \text{IW}(\boldsymbol{\Sigma} \mid \mathbf{S}_0, v_0) \\
&= \frac{1}{Z_{NIW}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{ -\frac{k_0}{2} (\boldsymbol{\mu} - \mathbf{m}_0)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \mathbf{m}_0) \right\} \\
&\quad \times |\boldsymbol{\Sigma}|^{-\frac{v_0 + D + 1}{2}} \exp\left\{ -\frac{1}{2} tr(\boldsymbol{\Sigma}^{-1} \mathbf{S}_0) \right\} \quad \text{(B.16)} \\
&= \frac{1}{Z_{NIW}} |\boldsymbol{\Sigma}|^{-\frac{v_0 + D + 2}{2}} \\
&\quad \times \exp\left\{ -\frac{k_0}{2} (\boldsymbol{\mu} - \mathbf{m}_0)^T \Sigma^{-1} (\boldsymbol{\mu} - \mathbf{m}_0) - \frac{1}{2} tr(\boldsymbol{\Sigma}^{-1} \mathbf{S}_0) \right\}, \quad \text{(B.17)}
\end{aligned}
$$

where $Z_{NIW} = 2^{v_0 D/2} \Gamma_D(v_0/2) (2\pi/k_0)^{D/2} |\mathbf{S}_0|^{-v_0/2}$ and $\Gamma_D(\cdot)$ is the multivariate Gamma function. $\mathbf{m}_0$ and $\mathbf{S}_0$ are the prior mean for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ respectively, whereas k_0 and v_0 define the strength of our belief in $\mathbf{m}_0$ and $\mathbf{S}_0$.

The posterior distribution obtained after observing N observations thus take the same form as the prior with hyperparameters $\{\mathbf{m}_N, k_N, v_N, \mathbf{S}_N\}$. The posterior distribution of NIW can be inferred using Bayes' theorem as:

$$
\begin{aligned}
p(\boldsymbol{\mu},\boldsymbol{\Sigma} \mid \mathbf{D}) =& \text{NIW}(\boldsymbol{\mu},\boldsymbol{\Sigma} \mid \mathbf{m}_N, k_N, v_N, \mathbf{S}_N) \\
\propto& p(\mathbf{D} \mid \boldsymbol{\mu},\boldsymbol{\Sigma}) \times \text{NIW}(\boldsymbol{\mu},\boldsymbol{\Sigma} \mid \mathbf{m}_0, k_0, v_0, \mathbf{S}_0) \\
\propto& (2\pi)^{\frac{-ND}{2}} |\boldsymbol{\Sigma}|^{\frac{-N}{2}} \exp\left\{ -\frac{N}{2}(\boldsymbol{\mu} - \overline{\mathbf{x}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \overline{\mathbf{x}}) \right\} \\
& \times \exp\left\{ -\frac{1}{2} tr\left(\boldsymbol{\Sigma}^{-1} \mathbf{S}_{\bar{x}}\right) \right\} \frac{1}{Z_{NIW}} |\boldsymbol{\Sigma}|^{-\frac{v_0+D+2}{2}} \\
& \times \exp\left\{ -\frac{k_0}{2}(\boldsymbol{\mu} - \mathbf{m}_0)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \mathbf{m}_0) - \frac{1}{2} tr\left(\boldsymbol{\Sigma}^{-1} \mathbf{S}_0\right) \right\}. \qquad \text{(B.18)}
\end{aligned}
$$

In order to derive the hyperparameters for the posterior NIW distribution, one can compare its similar terms with those in Equation (B.18):

- Comparing the power terms of $|\boldsymbol{\Sigma}|$, one gets:

$$
\begin{aligned}
|\boldsymbol{\Sigma}|^{-\frac{v_N+D+2}{2}} &= |\boldsymbol{\Sigma}|^{\frac{-N}{2}} |\boldsymbol{\Sigma}|^{-\frac{v_0+D+2}{2}} \\
v_N &= v_0 + N. \qquad \text{(B.19)}
\end{aligned}
$$

- When comparing the exponential terms:

$$
\begin{aligned}
& \exp\left\{ -\frac{k_N}{2}(\boldsymbol{\mu} - \mathbf{m}_N)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \mathbf{m}_N) - \frac{1}{2} tr\left(\boldsymbol{\Sigma}^{-1} \mathbf{S}_N\right) \right\} \\
=& \exp\left\{ -\frac{k_0}{2}(\boldsymbol{\mu} - \mathbf{m}_0)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \mathbf{m}_0) - \frac{1}{2} tr\left(\boldsymbol{\Sigma}^{-1} \mathbf{S}_0\right) \right\} \\
& \times \exp\left\{ -\frac{N}{2}(\boldsymbol{\mu} - \overline{\mathbf{x}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \overline{\mathbf{x}}) \right\} \exp\left\{ -\frac{1}{2} tr\left(\boldsymbol{\Sigma}^{-1} \mathbf{S}_{\bar{x}}\right) \right\}. \qquad \text{(B.20)}
\end{aligned}
$$

Equation (B.20) can be re-written as:

$$
\begin{aligned}
tr\left(\boldsymbol{\Sigma}^{-1} \mathbf{S}_N\right) =& tr\left(\boldsymbol{\Sigma}^{-1} \mathbf{S}_0\right) + tr(\boldsymbol{\Sigma}^{-1} \mathbf{S}_{\bar{x}}) \\
& + k_0(\boldsymbol{\mu} - \mathbf{m}_0)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \mathbf{m}_0) + N(\boldsymbol{\mu} - \overline{\mathbf{x}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \overline{\mathbf{x}}) \\
& - k_N(\boldsymbol{\mu} - \mathbf{m}_N)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \mathbf{m}_N).
\end{aligned}
$$

Following the properties of matrix trace [219]: $tr(A) + tr(B) = tr(A+B)$, and $tr(ABC) = tr(BCA) = tr(CAB)$, the above equation can be formulated as:

$$
\begin{aligned}
tr\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}_N\right) =& tr\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}_0+\boldsymbol{\Sigma}^{-1}\mathbf{S}_{\bar{x}}\right) \\
&+tr\Big[\boldsymbol{\Sigma}^{-1}k_0(\boldsymbol{\mu}-\mathbf{m}_0)(\boldsymbol{\mu}-\mathbf{m}_0)^T+\boldsymbol{\Sigma}^{-1}N(\boldsymbol{\mu}-\overline{\mathbf{x}})(\boldsymbol{\mu}-\overline{\mathbf{x}})^T \\
&-\boldsymbol{\Sigma}^{-1}k_N(\boldsymbol{\mu}-\mathbf{m}_N)(\boldsymbol{\mu}-\mathbf{m}_N)^T\Big]. \qquad \text{(B.21)}
\end{aligned}
$$

Some terms in the right-hand side of Equation (B.21) can be written as the following by completing the squares [218]:

$$
\begin{aligned}
&k_0(\boldsymbol{\mu}-\mathbf{m}_0)(\boldsymbol{\mu}-\mathbf{m}_0)^T+N(\overline{\mathbf{x}}-\boldsymbol{\mu})(\overline{\mathbf{x}}-\boldsymbol{\mu})^T \\
=&k_N(\boldsymbol{\mu}-\mathbf{m}_N)(\boldsymbol{\mu}-\mathbf{m}_N)^T+\frac{k_0N}{k_0+N}(\overline{\mathbf{x}}-\mathbf{m}_0)(\overline{\mathbf{x}}-\mathbf{m}_0)^T. \qquad \text{(B.22)}
\end{aligned}
$$

To obtain k_N and m_N, one can extract relevant terms in $\boldsymbol{\mu}^2$ and $\boldsymbol{\mu}$ in Equation (B.22) as:

$$
\begin{aligned}
k_N\boldsymbol{\mu}^2 &= N\boldsymbol{\mu}^2+k_0\boldsymbol{\mu}^2 \\
k_N &= k_0+N. \qquad \text{(B.23)}
\end{aligned}
$$

$$
\begin{aligned}
-2k_N\boldsymbol{\mu}\mathbf{m}_N &= -2N\boldsymbol{\mu}\overline{\mathbf{x}}-2k_0\boldsymbol{\mu}\mathbf{m}_0 \\
m_N &= \frac{N\overline{\mathbf{x}}+k_0\mathbf{m}_0}{k_N} \\
&= \frac{k_0}{k_0+N}\mathbf{m}_0+\frac{N}{k_0+N}\overline{\mathbf{x}}. \qquad \text{(B.24)}
\end{aligned}
$$

Similarly, $\mathbf{S}_N$ can be obtained by substituting Equation (B.22) into Equation (B.21), and compare the remaining terms as:

$$
\begin{aligned}
tr\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}_N\right) &= tr\left[\boldsymbol{\Sigma}^{-1}\mathbf{S}_0+\boldsymbol{\Sigma}^{-1}\mathbf{S}_{\bar{x}}+\frac{k_0N}{k_0+N)}\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}}-\mathbf{m}_0)(\overline{\mathbf{x}}-\mathbf{m}_0)^T\right] \\
\mathbf{S}_N &= \mathbf{S}_0+\mathbf{S}_{\bar{x}}+\frac{k_0N}{k_0+N}(\overline{\mathbf{x}}-\mathbf{m}_0)(\overline{\mathbf{x}}-\mathbf{m}_0)^T. \qquad \text{(B.25)}
\end{aligned}
$$

Appendix C

Additional Results

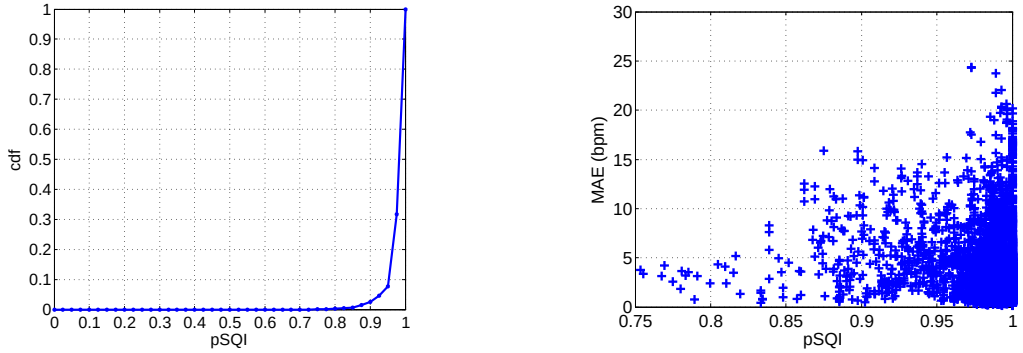
C.1 Results of the BCLAs Model

Capnobase RR results

The cumulative density function (cdf) of the signal quality (i.e., pSQI) of the PPG signals for 150 windows across 42 subjects is shown in Fig. C.1.

PCinC QT results

The 548 recordings of the PCinC QT dataset Division 4 was bootstrapped 1,000 times and the MAE results are shown in Table C.1.



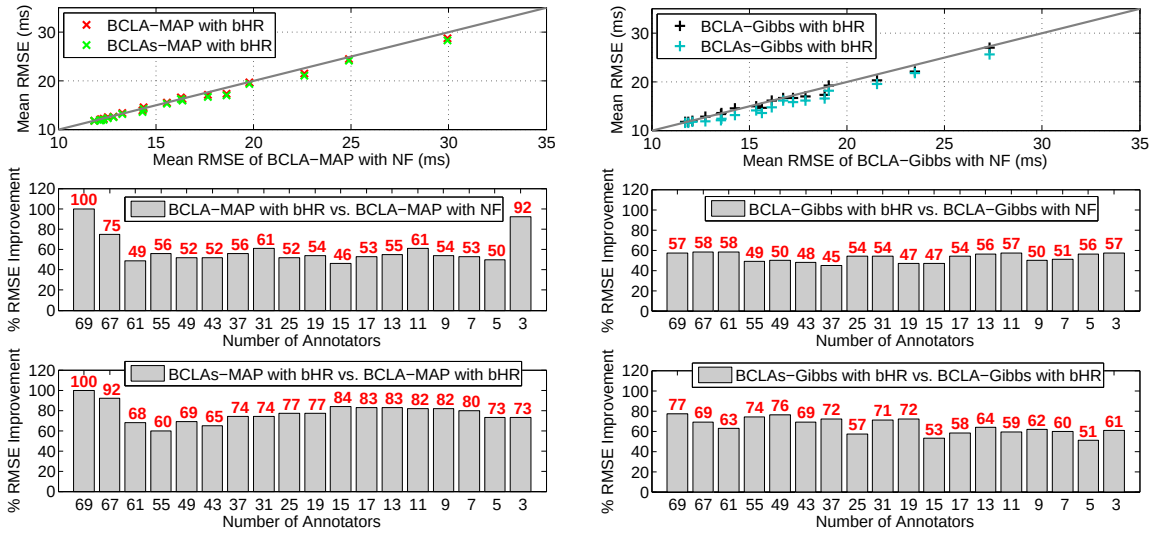
(a) (b)

Fig. C.1 (a) cdf of the pSQI values for different windows across 42 subjects in the Capnobase RR dataset; (b) MAEs across annotators versus pSQI values for all windows.

Table C.1 MAEs of inferred labels using different voting approaches with and without features in the PCinC QT dataset.

MAE (ms)										
Division (number)	sSTAPLE	Mean	Median	EM-R with bHRSQIs	MAP Approach			Gibbs Approach		
					BCLA with		BCLAs with bHR	BCLA with		BCLAs with bHR
					NF	bHR		NF	bHR	
2 (48)	11.64±0.65†‡	12.64±0.55†‡	11.75±0.42†‡	11.32±0.45†‡	9.34±0.43†‡	9.28±0.44*†‡	9.12±0.40†	8.96±0.43†‡	8.94±0.43†‡	8.83±0.39‡
3 (21)	19.15±1.78†‡	22.99±0.83†‡	14.05±0.55†‡	12.32±0.61†‡	10.60±0.69†	10.75±0.71*†‡	10.59±0.73†	11.07±0.66†‡	10.16±0.61*†‡	10.39±0.65‡
4 (69)	13.23±0.49†‡	14.22±0.45†‡	11.23±0.39†‡	11.28±0.43†‡	8.60±0.44†‡	8.54±0.43*†‡	8.49±0.39†	8.51±0.39†	8.52±0.40†	8.70±0.47‡

Note: the above are the mean ± standard deviation of the RMSE results obtained from 1,000 subsamples. Results of BCLAs were obtained using T_{min} as stated in Table 7.1. The † indicates the BCLAs-Gibbs results with the bHR feature were significantly different (for columns 2 to 10 only) from the rest of the voting strategies ($p < 0.05$) using the two-tailed Wilcoxon rank-sum test. The * (columns 6 versus 7, and columns 9 versus 10 only) indicates the significance ($p < 0.05$) of introducing the bHR feature to the BCLA-MAP and BCLA-Gibbs model respectively. The ‡ (columns 2 to 7, and columns 9 to 11) indicates the BCLAs-MAP results with the bHR feature were significantly different from others ($p < 0.05$).



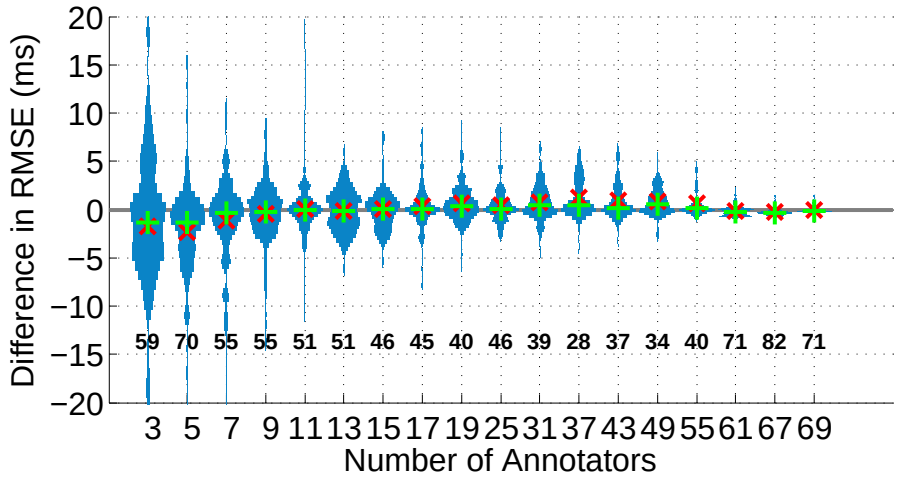
(a) **The MAP approach** (b) **The Gibbs Sampling Approach**

Fig. C.2 The difference in the RMSE results between BCLA and BCLAs as a function of the number of automated annotators: (a) shows the mean RMSEs using the MAP approach, and (b) shows the results using the Gibbs sampling approach. Each top plot in (a) and (b) shows the mean RMSEs of BCLA with no features (NF) versus those of BCLA or BCLAs with the bHR feature, when selecting number of annotators ranged from three to 69. The leading diagonal line of each plot indicates a perfect matched in RMSE between BCLA with NF and BCLA with bHR or BCLAs with bHR. The selected numbers of annotators were sorted in ascending order according to the mean of RMSEs of BCLA with NF. The % RMSE improvement indicates the number of percentage counts (from a total of 100 sub-sampled RMSEs) that BCLA with the bHR feature outperformed (i.e., produced smaller RMSE) BCLA with NF or BCLAs with the bHR feature outperformed BCLA with the bHR feature.

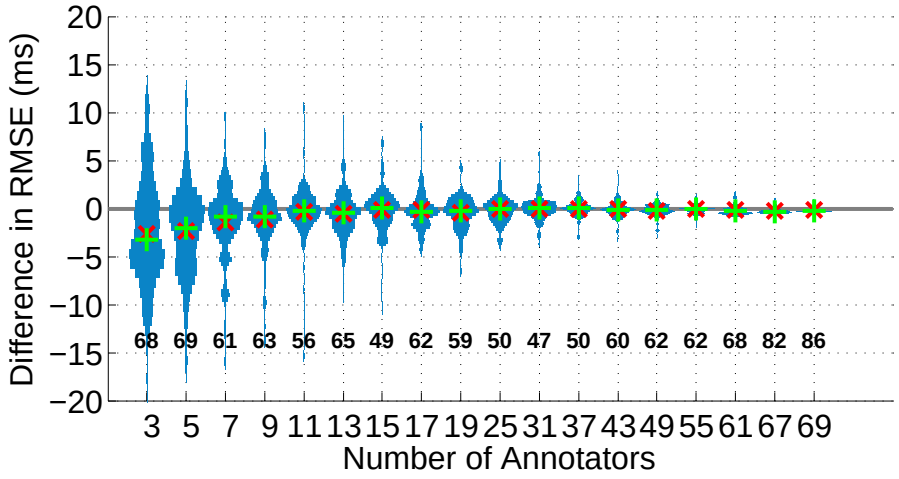
Effect of the HR Feature versus No Features A further analysis was conducted to investigate the effect of introducing the bHR feature into the BCLA and BCLAs models as a function of the number of automated annotators for Division 4 (see Fig. C.2). 100 RMSEs were computed as a result of sub-sampling the number of annotators 100 times ranged from three to 69. For each selected number of annotators (i.e., three), the mean of the 100 RMSEs estimated using BCLA with NF, was plotted against that estimated using BCLA with the bHR feature, and BCLAs with the bHR feature. Fig. C.2 (a) shows the results using the MAP approach, and those using the Gibbs sampling approach are shown in Fig. C.2 (b). In the MAP results, BCLA-MAP with the bHR feature has a slight improvement (i.e., it produced

smaller RMSEs at least over 50% of the time out of a total of 100 sub-sampled RMSEs) over BCLA without features except when annotator number is five, 15, and 61 (see middle bar plot in Fig. C.2 (a)). BCLA-Gibbs RMSE improvement was also relatively small when the bHR feature was included: less than 60% improvement was observed when selecting different numbers of annotators (see middle bar plot in Fig. C.2 (b)). In the case when the BCLAs model with bHR was included, both the MAP and Gibbs sampling approaches have small but consistent superior performance over their corresponding BCLA models with bHR respectively when selecting any number of annotators (see bottom bar plots in Fig. C.2 (a) and (b)).

Fig. C.3 shows the mean and median as well as the probability density distribution of the difference in RMSE out of the 100 runs for selecting number of annotators from three to 69. The frequency of superior performance (i.e., with smaller RMSE) is shown as a percentage in the bottom of each plot for different numbers of annotators. BCLA-Gibbs with bHR outperformed BCLA-MAP with bHR when up to 13 annotators were considered as well as when annotator number is greater and equals to 61 (see Fig. C.3 (a)). This was also observed in Fig. 6.9 when the comparison was with no features. In Fig. C.3 (b), BCLAs-Gibbs with bHR outperformed BCLAs-MAP with bHR for most number of annotators except it was 15, 25, 31, and 37. Looking at the results of the BCLA model with bHR, introducing the signal quality (i.e., BCLAs with bHR) has more effect on improving the performance of the Gibbs sampling approach than the MAP approach (see Fig. C.3): the probability density distribution for each selected number annotators has a shorter tail towards the positive difference in RMSE (i.e., where MAP outperformed Gibbs), and its mode has moved further to the negative difference in RMSE, indicating the superior performance of Gibbs over MAP when considering the signal quality extension of BCLA (i.e., the BCLAs model).

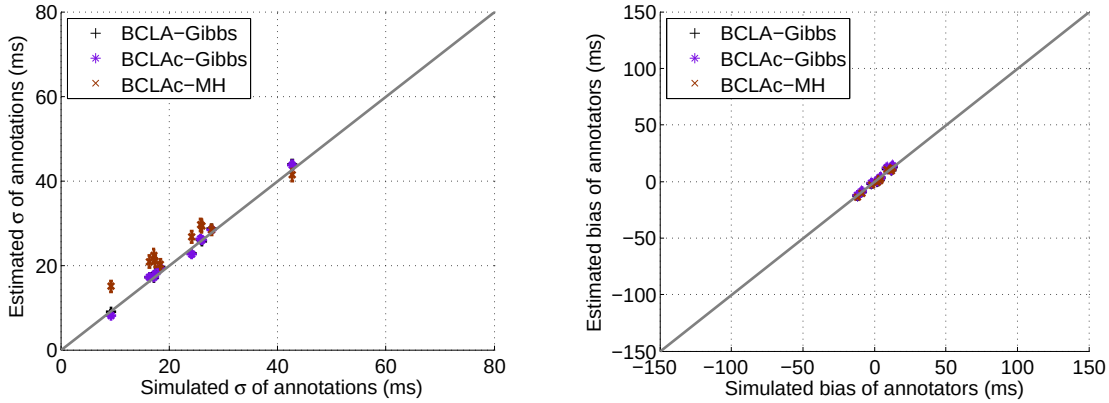


(a) **BCLA-Gibbs with bHR versus BCLA-MAP with bHR**

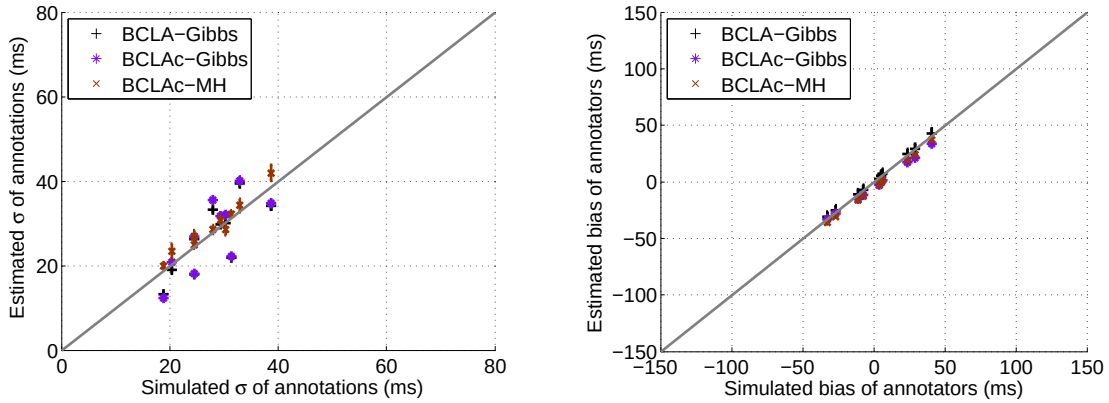


(b) **BCLAs-Gibbs with bHR versus BCLAs-MAP with bHR**

Fig. C.3 The difference in the RMSE results between BCLA and BCLAs with the bHR feature as a function of the number of automated annotators: Each plot has a probability density estimate of the sub-sampled 100 RMSE differences calculated for different numbers of annotators ranged from three to 69. The mean and the median of the density estimate are labelled as $\times$ in red and $+$ in green, respectively. The numbers listed in each plot indicate the percentage counts (from a total of 100 sub-sampled RMSEs) that BCLA-Gibbs with bHR outperformed BCLA-MAP with bHR in (a), and BCLAs-Gibbs with bHR outperformed BCLAs-MAP with bHR in (b).



(a) the $\mathscr{M}_\Sigma$ model: : 1000 annotations x 10 annotators



(b) the $\mathscr{M}_\rho$ model: 1000 annotations x 10 annotators

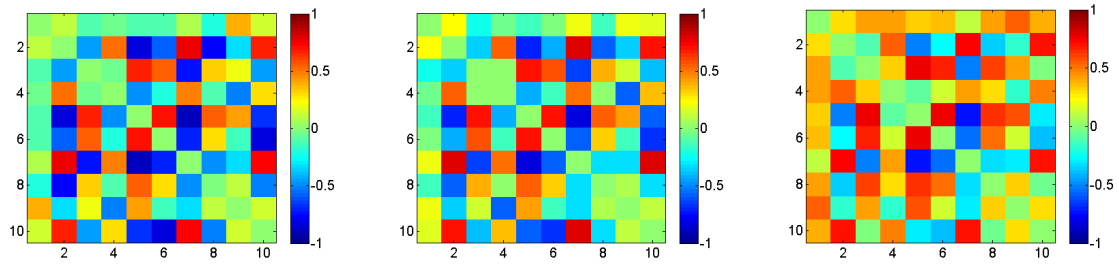
Fig. C.4 A comparison of the simulated and inferred σ and bias values of 10 annotators in the simulated datasets generated from (a) BCLAc with $\mathscr{M}_\Sigma$, and (b) BCLAc with $\mathscr{M}_\rho$. The diagonal (grey) line indicates a perfect match between the simulated and estimated results. The results are plotted as with one standard deviation from the mean.

C.2 Results of the BCLAc Model

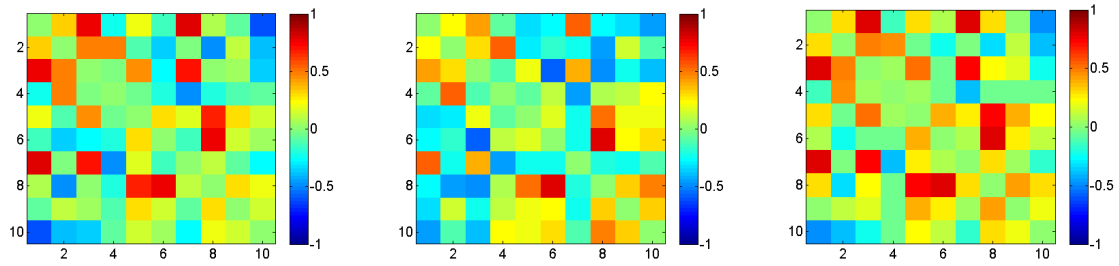
Simulated Datasets with 10 Annotators

When 10 annotators were considered for datasets generated using both models, BCLA-Gibbs alone was sufficient to provide reliable estimation of the bias values for various number of annotations (see examples in Fig. C.4 for datasets of 10 annotators with 1,000 annotations).

The inferred correlated errors of annotators using the BCLAc-Gibbs and BCLAc-MH approaches are shown in Fig. C.5. When the number of annotators increased to 10, it was harder to predict accurately the precise values of the correlated errors of annotators. In the



(a) The $\mathscr{M}_{\Sigma}$ model: 1000 annotations x 10 annotators



(b) The $\mathscr{M}_{\rho}$ model: 1000 annotations x 10 annotators

Fig. C.5 A comparison of the simulated and inferred correlation of errors among 10 annotators in the simulated datasets generated from (a) the $\mathscr{M}_{\Sigma}$ model, and (d) the $\mathscr{M}_{\rho}$ model. In each row, the *true* $\boldsymbol{\rho}$ (left), is compared with its inferred values using BCLAc-Gibbs (middle), and BCLAc-MH (right). Note that each correlated matrix is subtracted from an identity matrix to remove the correlation of an annotator with itself.

simulated dataset with the $\mathscr{M}_{\Sigma}$ model, the inferred precision values (see Fig. C.4 (a)) were well predicted when using both the BCLA-Gibbs and BCLAc-Gibbs approaches. BCLAc-MH was slightly worse than others with an over-estimation of the majority of the σ values. However in the case of the dataset with the $\mathscr{M}_{\rho}$ model (see Fig. C.4 (b)), BCLAc-MH was capable of inferring more accurate precision values over the other two approaches.