

PAPER • OPEN ACCESS

Bayesian fusion of physiological measurements using a signal quality extension

To cite this article: Tingting Zhu *et al* 2018 *Physiol. Meas.* **39** 065008

View the [article online](#) for updates and enhancements.

Related content

- [Multimodal heart beat detection using signal quality indices](#)
Alistair E W Johnson, Joachim Behar, Fernando Andreotti et al.
- [Heart beat detection in multimodal physiological data using a hidden semi-Markov model and signal quality indices](#)
Marco A F Pimentel, Mauro D Santos, David B Springer et al.
- [False alarm reduction in critical care](#)
Gari D Clifford, Ikaro Silva, Benjamin Moody et al.

OPEN ACCESS



PAPER

Bayesian fusion of physiological measurements using a signal quality extension

RECEIVED
12 February 2018REVISED
27 April 2018ACCEPTED FOR PUBLICATION
29 May 2018PUBLISHED
26 June 2018

Original content from
this work may be used
under the terms of the
[Creative Commons
Attribution 3.0 licence](#).

Any further distribution
of this work must
maintain attribution
to the author(s) and the
title of the work, journal
citation and DOI.

Tingting Zhu¹ , Alistair E W Johnson², Yang Yang¹, Gari D Clifford^{3,4} and David A Clifton¹¹ Department of Engineering Science, Institute of Biomedical Engineering, University of Oxford, Oxford, United Kingdom² Institute for Medical Engineering & Science, Massachusetts Institute of Technology, Boston, MA, United States of America³ Department of Biomedical Informatics, Emory University, Georgia Institute of Technology, Atlanta, GA, United States of America⁴ Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA, United States of AmericaE-mail: tingting.zhu@eng.ox.ac.uk**Keywords:** Bayesian models, unsupervised learning, signal quality**Abstract**

Objective: The fusion of multiple noisy labels for biomedical data (such as ECG annotations, which may be obtained from human experts or from automated systems) into a single robust annotation has many applications in physiologic monitoring. Directly modelling the difficulty of the task has the potential to improve the fusion of such labels. This paper proposes a means for the incorporation of task difficulty, as quantified by ‘signal quality’, into the fusion process. **Approach:** We propose a Bayesian fusion model to infer a consensus through aggregating labels, where the labels are provided by multiple imperfect automated algorithms (or ‘annotators’). Our model incorporates the signal quality of the underlying recording when fusing labels. We compare our proposed model with previously published approaches. Two publicly available datasets were used to demonstrate the feasibility of our proposed model: one focused on QT interval estimation in the ECG and the other focused on respiratory rate (RR) estimation from the photoplethysmogram (PPG). We inferred the hyperparameters of our model using maximum- *a posteriori* inference and Gibbs sampling. **Main results:** For the QT dataset, our model significantly outperformed the previously published models (root-mean-square error of 25.61 ± 8.68 ms for our model versus 30.79 ± 13.16 ms from the best existing model) when fusing labels from only three annotators. For the RR dataset, no improvement was observed compared to the same model without signal quality modelling, where our model outperformed existing models (mean-absolute error of 1.89 ± 0.36 bpm for our model versus 2.22 ± 0.41 bpm from the best existing model). We conclude that our approach demonstrates the feasibility of using a signal quality metric as a confidence measure to improve label fusion. **Significance:** Our Bayesian learning model provides an extension over existing work to incorporate signal quality as a confidence measure to improve the reliability of fusing labels from biomedical datasets.

1. Introduction

In the scenario where only subjective predicted estimates from multiple annotators (i.e. algorithms or humans) are provided, where they are highly variable and the ground truth is not readily available, a voting model can be considered to aggregate predictions from these annotators to infer a single ground truth. A classical approach is the use of EM to jointly infer the precision of each annotator with the inferred ground truth (Dawid and Skene 1979). More recently, Welinder and Perona (2010) 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. In medical imaging, Warfield *et al* (2008) proposed a method for validating segmentation by estimating the bias and variance of each annotator. A limitation of previous approaches was that they do not cater for missing annotations or the incorporation of physiological features into the model. Our previously proposed Bayesian continuous-valued label aggregator (BCLA) model

(Zhu *et al* 2015) addressed this by jointly inferring the ground truth and estimating the bias and precision of annotators in an unsupervised manner. However, our model assumed that the precision of each annotator was not affected by underlying task difficulty, an assumption which is infrequently true in practice, particularly in medical applications. Large amounts of noise can increase the false-positive alarm rate for automated annotators in electrocardiogram (ECG) arrhythmia analysis (Aboukhalil *et al* 2008) or photoplethysmogram analysis (Pettersson *et al* 2007).

The level of noise in a signal can be quantified by one or more derived measures referred to as signal quality indices (SQIs). Signal quality, as quantified by SQIs, can be seen as a measure of task difficulty: noisy records are harder to label due to noise contamination of the signals, while clean records are easier to label. Most studies have used signal quality as a filter for removing artefactual segments in the pre-processing stage (Karlen *et al* 2013), or considered it as a feature component for automated prediction (Behar *et al* 2013, Zhu *et al* 2014). Other studies have used the signal quality in a similar context to estimate confidence for a given segment of medical data (Li *et al* 2008, Tsanas *et al* 2014, Vollmer 2014, Johnson *et al* 2015, Pimentel *et al* 2015).

Instead of using the signal quality as an indicator for removing noisy segments which might cause large sections of data to be discarded, we propose to incorporate signal quality as an independent variable in a Bayesian model. Expert annotators 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. Our proposed model incorporates a probabilistic measure of signal quality to capture task difficulty and thus better infer the underlying ground truth of a physiological signal.

2. Novelty

This article introduces the use of signal quality of the physiological measurements to describe the uncertainty of the labels provided by automated algorithms. Our proposed Bayesian unsupervised model, BCLAs, is able to fuse the predicted labels from algorithms by incorporating the underlying quality of a physiology, to further improve the estimation of consensus, and guarantee the output is better than any of the individual algorithms considered independently.

3. Methods

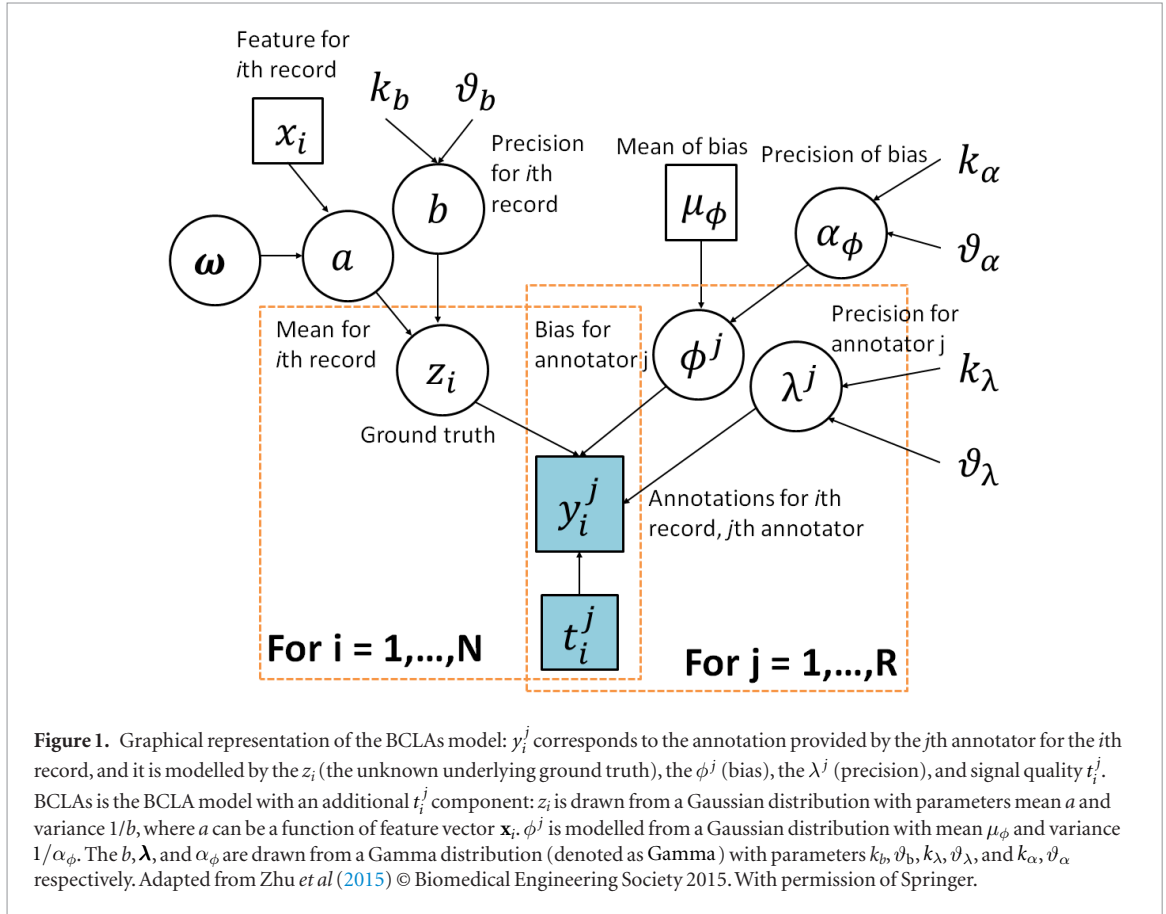
Suppose that there are N recordings of time-series data annotated by R annotators. Let $\mathbf{D} = [\mathbf{x}_i^T, y_i^{j=1}, \dots, y_i^{j=R}]_{i=1}^N$, where \mathbf{x}_i is a column feature vector for the i th record containing d features, and y_i^j corresponds to the annotation provided by the j th annotator for the i th record. A novel incorporation of the signal quality is proposed in our model: we assume that y_i^j is a noisy version of z_i , the unknown underlying ground truth for the i th record, and that y_i^j is drawn from a Gaussian distribution as $\mathcal{N}(z_i + \phi^j, (t_i^j \lambda^j)^{-1})$ ⁵. The bias for annotator j , denoted as ϕ^j , is defined as the averaged estimation error, and λ^j is the precision of the j th annotator, defined as the estimated inverse-variance of annotator j . The signal quality for the i th record from a j th annotator is denoted as t_i^j (note the dependence on j as not all annotators provide labels on the same segment). The signal quality is assumed to be within the range of $(0, 1]$, where 1 indicates a good signal quality (the segment is clean) while any value close to 0 indicates a noisy segment. Essentially, t_i^j acts as a scaling factor for λ^j : as $t_i^j \rightarrow 0$ annotator j is less precise in labelling the i th 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 dataset specific lower bound of t_i^j needs to be obtained to define the range of ‘good quality’ data.

The bias for j th annotator, ϕ^j , is also assumed to be drawn from a Gaussian distribution as $\mathcal{N}(\mu_\phi, 1/\alpha_\phi)$ ⁶. Furthermore, the ground truth, z_i , has its own distribution as $\mathcal{N}(a, 1/b)$, where a can be expressed as a linear regression function $f(\mathbf{w}, \mathbf{x})$, with \mathbf{w} being the coefficients of the regression as previously described in the BCLA model (Zhu *et al* 2015). An intercept that models the overall offset predicted in the regression is included, and is different from the annotator-specific bias in this model. The graphical representation of BCLAs is illustrated in figure 1.

Under the assumption that records are independent and y_i^1, \dots, y_i^R are conditionally independent given the feature \mathbf{x}_i , the posterior probability density function with signal quality extension of the parameter $\theta = \{\mathbf{w}, \lambda, \phi, \alpha_\phi, b, z_i\}$ for a given dataset \mathbf{D} can be written using Bayes’ theorem as

⁵ In the absence of prior knowledge of the underlying distribution, the simplest is that of the Gaussian distribution.

⁶ Although the biases of the annotators might be assumed to have other distributions, such choices are likely to be dataset-dependent. In the absence of any knowledge of the underlying distribution of biases, we adopt the strategy of assuming them to be drawn from a Gaussian distribution. The same assumption was applied to the ground truth.



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

3.1. The maximum-*a posteriori* (BCLAs-MAP) approach

Estimation of $\boldsymbol{\theta}$ can be performed using the maximum-*a posteriori* (MAP) approach, which maximises the log-posterior of the parameters, i.e. $\arg\max_{\boldsymbol{\theta}} \{\log p(\boldsymbol{\theta} | \mathbf{D})\}$. The parameters in $\boldsymbol{\theta}$ can be derived by equating the gradient of the log-posterior to zero. In the case where there exist missing labels from annotators, only the available annotations should be considered for inferring the ground truth. Let U_j be the set of records/segments with annotations provided by the j th annotator, and V_i be the set of annotators that provided annotations for the i th record. Then, we can derive equations for each parameter in $\boldsymbol{\theta}$ as follows:

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

$$\mathbf{w} = \left(\sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \sum_{i=1}^N \mathbf{x}_i z_i. \tag{3}$$

$$\phi^j = \frac{\sum_{i \in U_j} t_i^j (y_i^j - z_i) + \mu_\phi (\frac{\alpha_\phi}{\lambda^j})}{\sum_{i \in U_j} t_i^j + \frac{\alpha_\phi}{\lambda^j}}. \tag{4}$$

$$\frac{1}{\alpha_\phi} = \frac{1}{R + 2(k_\alpha - 1)} \left[\sum_{j=1}^R (\phi^j - \mu_\phi)^2 + \frac{2}{\vartheta_\alpha} \right]. \quad (5)$$

$$z_i = \frac{\sum_{j \in V_i} t_i^j \lambda^j (y_i^j - \phi^j) + (\mathbf{x}_i^T \mathbf{w}) b}{\sum_{j \in V_i} t_i^j \lambda^j + b}. \quad (6)$$

$$\frac{1}{b} = \frac{1}{N + 2(k_b - 1)} \left[\sum_{i=1}^N (z_i - \mathbf{x}_i^T \mathbf{w})^2 + \frac{2}{\vartheta_b} \right]. \quad (7)$$

The parameters can be solved using a two step expectation maximization (EM) algorithm: (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 (6); (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} , ϕ , α_ϕ , b , and λ can be updated using equations (2)–(5) and (7) accordingly in a sequential order until convergence.

3.2. The Gibbs sampling (BCLAs-Gibbs) approach

The Gibbs sampling approach can be considered as a stochastic process of the EM algorithm. Unlike the MAP approach, which is a point estimation, the Gibbs sampler not only provides estimates but also produces confidence in its estimation—this is a key advantage of fully-Bayesian inference. The random samples of the parameters in θ are drawn from the posterior distribution while one parameter is assigned to a fixed value at each sampling process. Each parameter in θ described in equation (1) can therefore be updated by sampling from its posterior distribution with its hyperparameters (denoted by $*$) as

$$z_i \sim \mathcal{N} \left(a_i^*, \frac{1}{b^*} \right), \quad (8)$$

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

$$\phi^j \sim \mathcal{N} \left(\mu_\phi^{j*}, \frac{1}{\alpha_\phi^{j*}} \right), \quad (9)$$

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

$$\lambda^j \sim \text{Gamma} \left(k_\lambda^{j*}, \vartheta_\lambda^{j*} \right), \quad (10)$$

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

$$b \sim \text{Gamma} (k_b^*, \vartheta_b^*), \quad (11)$$

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

$$\alpha_\phi \sim \text{Gamma}(k_\alpha^*, \vartheta_\alpha^*), \quad (12)$$

$$k_\alpha^* = k_\alpha + \frac{R}{2}, \quad \frac{1}{\vartheta_\alpha^*} = \frac{\sum_{j=1}^R (\phi^j - \bar{\phi})^2}{2} + \frac{1}{\vartheta_\alpha}. \quad (13)$$

Note that \bar{z} and $\bar{\phi}$ are the averaged value of \mathbf{z} and ϕ respectively. The Gamma distributions (as conjugate priors), are considered as the prior probabilities for the parameter set θ . After M sequences of the Gibbs sampler, the expectation of each parameter can be approximated by calculating the mean across samples after the burn-in period (i.e. $\frac{M}{2}$).

4. Data description

4.1. Capnabase RR dataset

The Capnabase dataset (Karlen *et al* 2010) was collected during elective surgery and routine anaesthesia of 59 children (median age: 9, range: 1–17 years) and 35 adults (median age: 52, range: 26–76 years). Photoplethysmogram (PPG) recordings and reference capnometry data were collected at a sampling frequency of 300 Hz. The dataset is useful for the development of algorithms to estimate respiratory rate from the PPG as the simultaneous capnometry data can be used as a ground truth. Karlen *et al* (2013) defined a test set of 42 recordings lasting 8 min each from 42 distinct subjects for estimating respiratory rate from the PPG. All experiments performed in this paper use this set.

Three respiratory-induced modulation signals are derived from the PPG: amplitude modulation (AM), baseline wander (BW), and frequency modulation (FM). These modulation signals are created using 32 s windows, with successive windows having 29 s overlap. To extract these modulations, PPG beat detection was performed using a segmentation algorithm (Li *et al* 2010), where it produced a series of maximum and minimum intensities for each PPG pulse. The series of maximum intensities were used for extracting the BW signals; the (max–min) amplitude was used to derive the AM signal; the intervals between successive beats was used to extract the FM signal. Respiratory rate (RR) was then estimated using two different spectral approaches: Fourier analysis (FFT) and autoregressive (AR) modelling. Spectral analysis requires evenly-sampled data, and so each time-series (corresponding to BW, AM and FM) was first re-sampled at 4 Hz using linear interpolation. For the FFT method, the frequency at which the maximum intensity of each spectrum is obtained within the frequency range of interest (corresponding to 3–60 breath-per-minute (denoted as bpm)) to derive RR. For the AR method (Orphanidou *et al* 2009), a AR model with order of 7 was fitted to each modulation signal. The ideal respiratory frequency was identified as that corresponding to the pole with the greatest magnitude within the plausible range of frequencies to derive RR. A maximum of 900 RRs were extracted for each subject for 150 windows of data. Each window has six RR estimates calculated using the FFT- and AR-based methods applied to the three modulation signals.

4.1.1. Signal quality in the Capnabase RR dataset

A pulse signal quality index (denoted as pSQI) was adapted from Pimentel *et al* (2015) to measure signal quality. pSQI is defined as the percentage of the agreement between two PPG peak detectors, *wabp* and *PUD*, calculated using 5 s windows with 50% overlap. To derive a single pSQI for each 32 s window described earlier, we average all the pSQI values calculated within the 32 s window. A total of 150 pSQI values were obtained for the 8 min 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* (2003)⁷. 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 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* (2010), 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.

4.2. PCinC QT dataset

The 2006 PhysioNet/Computing in Cardiology (PCinC) Challenge QT dataset (Moody *et al* 2006) is publicly-available and provides ECG QT interval annotations from both human and algorithmic annotations. Each participant in the Challenge was required to submit a Q onset with accompanying T offset for one ‘representative’ beat in lead II of each of the 548 recordings in the Physikalisch-Technische Bundesanstalt Diagnostic ECG Database (PTBDB) (Bousseljot 1995). The QT length estimated from each participant represents the QT interval of a particular recording. The records were obtained from 290 subjects (209 men with mean age of 55.5 and 81 women with mean age of 61.6) with a variety of ECG morphologies having different QT intervals ranging from 256 ms to 529 ms, each represented by between one and five recordings. About 20% of the subjects were healthy controls.

A total of 38 621 annotations sourced from: 20 human annotators (grouped into Division 1 and used to derive the reference annotations); 48 automated algorithms (grouped into Division 2 as algorithms were closed-source); and 21 automated algorithms (grouped into Division 3 as algorithms were open-source). An additional

⁷<https://physionet.org/physiotools/matlab/wfdb-app-matlab/html/wabp.html>

division, Division 4, was created so as to combine all automated algorithms from Divisions 2 and 3, and to infer a potentially better estimation of QT intervals. The competition score for each entry was calculated from the root-mean-square-error (RMSE) between the submitted and the reference QT intervals. To reduce any possible inter-beat variations, the first 5 s segment of each record was considered as a short segment where the QT interval was not changing dramatically (with respect to any particular beat chosen by an annotator), 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.

4.2.1. Signal quality in the QT dataset

In the context of the QT dataset considered by this study, it is known that an underlying *true* QT interval can be modulated by heart rate variation (International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use 2014) and signal quality is recording-dependent. As a proof-of-concept, a beat signal quality (denoted as bSQI) as described by Johnson *et al* (2015) was used in the work described by this study. The bSQI variable measures the percentage of the agreement of two R-peak detectors, naming *gqrs* and *jqrs*, within a 5 s segment, evaluated every 1 s. $\text{bSQI} = 1$ 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 $\text{bSQI} < 1$ indicates the disagreement between detectors which is assumed to be due to the presence of noise and artefact. Description of each detector is as follows (Johnson *et al* 2015):

- (i) *gqrs* is available in PhysioNet as a Matlab function of the ‘wfdb’ toolbox⁸. 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 described in Behar *et al* (2013) and was modified: (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.

5. Methodology of comparison

5.1. Capnobase RR dataset

The BCLAs model was applied for fusing RR estimates for 42 subjects individually. It was evaluated in the same parameter setting as the BCLA model (see table 1). The mean of the mean-absolute-error (MAE) and its standard error of the inferred RR estimates from the PPG across all subjects using BCLAs were compared to the reference RR values from capnography. The MAE was chosen as the method of comparison as it allows us to compare our results with those that were published in literature. BCLAs was applied to the RR estimates extracted using the FFT-based, AR-based, and all (FFT + AR) algorithms. In addition, BCLAs was compared to BCLA, ‘smart fusion’ (i.e. the benchmarking algorithm proposed by authors of the dataset (Karlen *et al* 2013)), the best-performing (lowest-MAE) single algorithm (denoted ‘Best’), the EM algorithm proposed by Raykar *et al* (2010) (denoted as EM-R), the scalar Simultaneous Truth and Performance Level Estimation (denoted as sSTAPLE) proposed by Warfield *et al* (2008), and also the mean and median voting approaches.

5.2. PCinCQT dataset

As mentioned previously, the value of the lower bound threshold (denoted as T) of bSQI needs to be initialised to be greater than zero to avoid assigning a zero precision to each annotator (which results variance of their annotations to be infinite). The evaluation of T is dataset-specific, and can be estimated as follows: 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 $\text{bSQI} \leq 1$ are mapped to 1. 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 obtained.

To test the feasibility of introducing bSQI into the BCLAs model, the same parameter setting (see table 1) was provided as in the BCLA model for latter comparison. The feature matrix of the beat-specific heart rate (bHR) $\mathbf{X}_{\text{bHR}} = [\mathbf{x}_1^T, \dots, \mathbf{x}_N^T]$ for N records was also considered, where $\mathbf{x}_i^T = [\sqrt{\tilde{RR}_i^{j=1}}, \dots, \sqrt{\tilde{RR}_i^{j=R}}]$ from R annotators for the i th record, and \tilde{RR} is the median of the R–R intervals within the segment.

In assessing the performance of BCLAs as a function of the number of annotators, a random number of annotators was selected 100 times with replacement. This was repeated with the annotator numbers varied from three to the maximum number of annotators in Division 4. RMSEs using BCLAs with the bHR feature were computed and compared to (1) BCLA with no features (denoted as NF); (2) BCLA with the bHR feature; (3) EM-R with the beat-specific heart rate and signal quality features described in a previous work (Zhu *et al* 2014)

⁸<http://physionet.org/physiotools/matlab/wfdb-app-matlab/html/gqrs.html>

Table 1. The parameters for modelling RR and QT datasets.

Symbol	Definition	Value	
		RR	QT
k_b	Shape of gamma distribution for b	3	3 ^e
ϑ_b	Scale of gamma distribution for b	0.006	0.0002 ^e
μ_ϕ	Mean of the bias distribution	Variable ^b	10 ^d
k_λ	Shape of gamma distribution for λ	3 ^a	4 ^c
ϑ_λ	Scale of gamma distribution for λ	0.02 ^a	0.003 ^c
k_α	Shape of gamma distribution for α_ϕ	5	3 ^d
ϑ_α	Scale of gamma distribution for α_ϕ	0.1	0.0005 ^d

N.B. b is the precision for the estimate of the ground truth, λ refers to annotator/signal-specific precisions, and α_ϕ is the precision for the estimate of the bias from ground truth.

RR dataset: the values with ^a are determined with the assumption that the RR estimates provided by the best modulation signal is ± 2 bpm away from the Breakell and Townsend-Rose (2001) and Drummond *et al* (2011). The values with ^b are estimated from the median RR estimates provided by the algorithms. *QT dataset:* the values with ^c are determined with the assumption that the annotations provided by the best performing algorithm is ± 5 ms away from the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (2014). The values with ^d are derived from Hughes (2006) and Couderc *et al* (2011). The values with ^e are derived from Malik *et al* (2002) and Goldenberg *et al* (2006).

Table 2. Comparison of MAE and standard error (SE) across 42 subjects in the Capnabase RR dataset.

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

Note: The smart fusion (Karlen *et al* 2013) is denoted as Mean* in the table as it is a variation of the mean voting, and it discards 44.6% windows of RR estimates as they had a standard deviation greater than four bpm.

(denoted as EM-R with bHRSQIs); (4) sSTAPLE. The RMSE was chosen as the method of comparison in this dataset as it allows us to compare our results with those that were published in 2006 PCinC Challenge.

6. Results and discussion

6.1. Capnabase RR dataset

MAEs of the RR estimates using BCLAs were similar to those using BCLA when fusing FFT-based, AR-based, and all (FFT + AR) algorithms (see table 2). Noting in figure 2(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 previously 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, but extraction of RRs was performed on the modulation signals, making the pSQI values from the PPG less predictable of the underlying signal quality. As a matter of fact, the pSQI values were not associated with MAEs of the RR estimates (see figure 2(b)). Although the BCLAs approaches did not improve upon the BCLA approaches, they outperformed other voting methods as well as the best-performing single algorithm. In particular, the BCLAs approaches (using all windows available) had smaller MAEs than those of the smart fusion which discarded 44.6% of windows that had a large variance in the estimates. Our proposed model is therefore deemed to have robust performance for noisy physiological data.

6.2. PCinC QT dataset

Figure 3(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 had little associated with the physical length of the reference QT intervals. A similar observation is shown in figure 3(b), where there is no linear relationship between the MAEs across annotators and the averaged bSQI extracted from all recordings. The

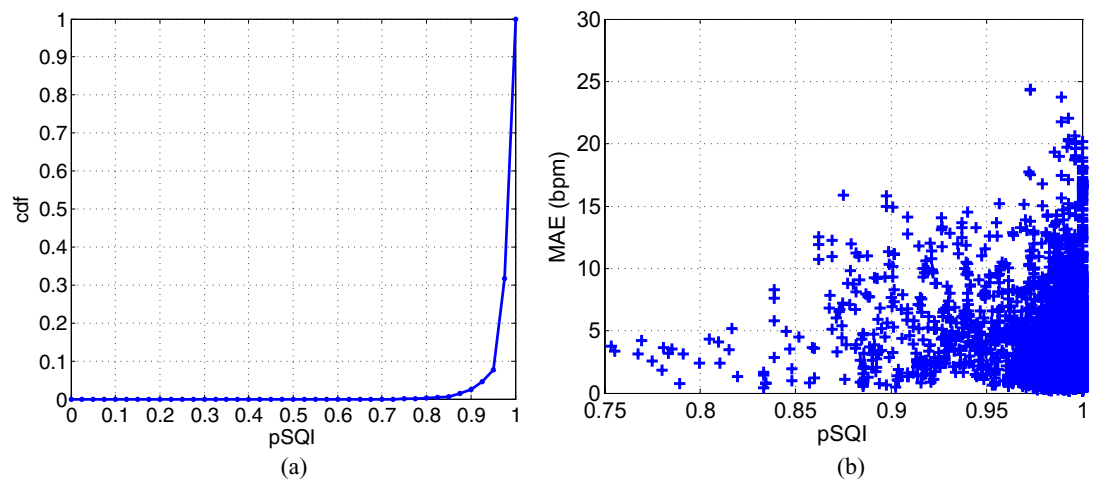


Figure 2. (a) The cumulative distribution function (cdf) of pSQI values for different windows across 42 subjects in the Capnobase RR dataset; (b) MAEs across annotators versus pSQIs values for all windows.

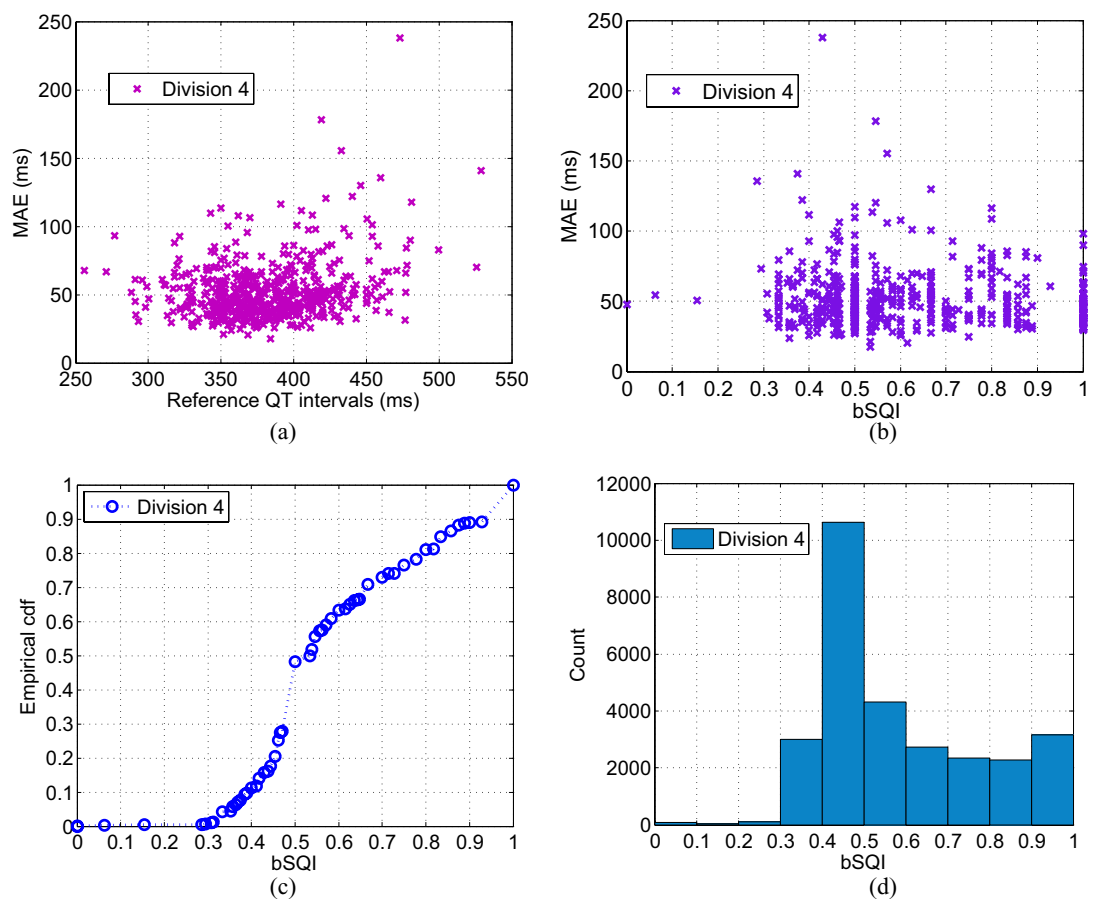


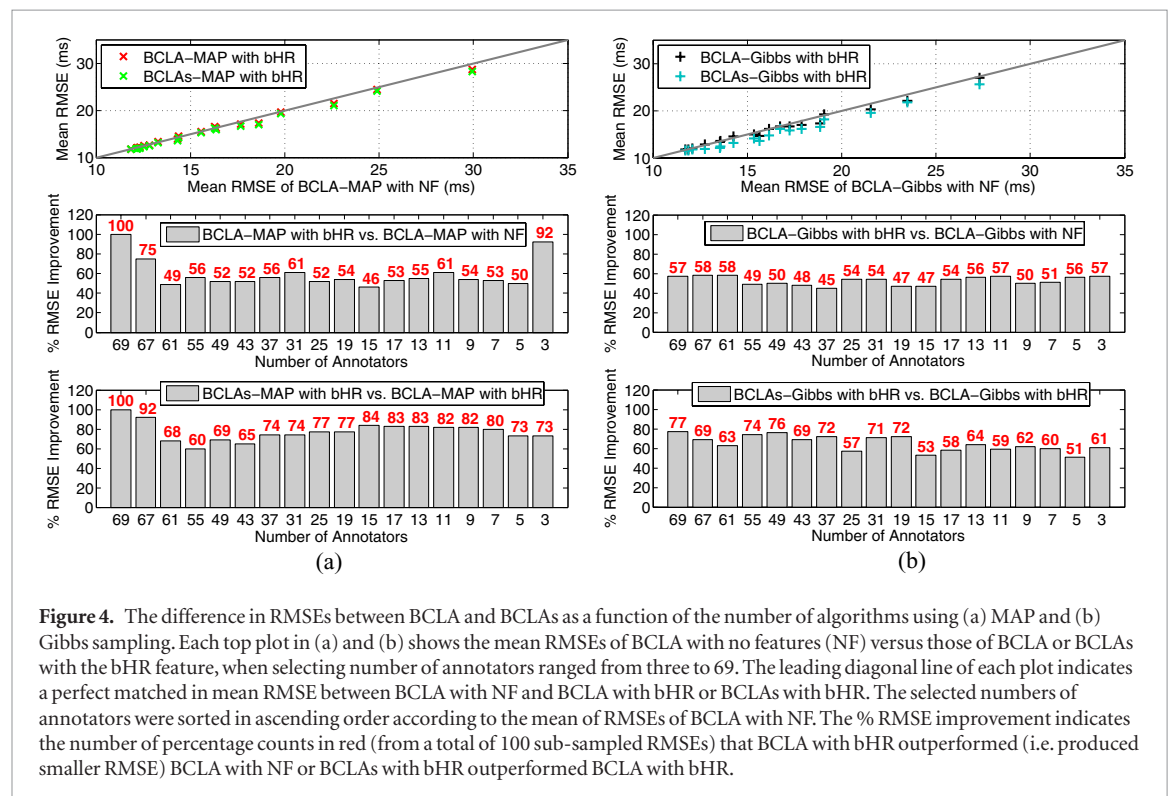
Figure 3. Example of errors and distribution plots for the PCinC QT dataset Division 4: (a) MAEs across annotators for all recordings were plotted against the reference QT intervals; (b) MAEs across annotators versus bSQI values for all recordings; (c) empirical cdf of bSQI values; (d) histogram of the distribution of bSQI values.

cumulative distribution function of bSQI values for Division 4 is shown in figure 3(c). 1.08% of the annotations have $\text{bSQI} \leq 0.3125$, which amounted to 260, 103, and 363 annotations for Divisions 2, 3, and 4, respectively. However, only 10.90% annotations had $\text{bSQI} = 1$ in Division 4, the remaining are distributed across different values of bSQI (see figure 3(d)). Hence a lower bound of bSQI should be obtained to define the range of ‘good quality’ data, and avoid discarding a large amount of annotations which have $\text{bSQI} < 1$.

Table 3. Optimisation of the lower bound T of the bSQI values for each division of the PCinC QT dataset.

Division (number)	BCLA-MAP	BCLA-Gibbs	BCLAs-MAP		BCLAs-Gibbs	
	RMSE ($T_{min} = 1$)		RMSE (T_{min})	T range	RMSE (T_{min})	T range
2 (48)	12.65 ms	12.07 ms	12.32 ms (0.33)	0.05–0.46	11.98 ms (0.83)	0.73–0.94
3 (21)	14.36 ms	14.70 ms	14.19 ms (0.41)	0.35–0.45	14.27 ms (0.95)	0.90–1.00
4 (69)	11.76 ms	11.47 ms	11.71 ms (0.67)	0.60–0.74	11.43 ms (0.42)	0.05–0.48

Note: 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.



6.2.1. Lower bound optimisation of bSQI

As described 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. Table 3 shows the optimised lower bound T of the bSQI values for different divisions of the QT dataset. Across all divisions, the BCLAs approaches had smaller RMSEs than the BCLA approaches. Note that a range of T values (see T range in table 3) produced the same minimum RMSE (rounded to two decimal places). As algorithms considered in Division 3 were different from those in Division 2, it was not trivial to perform direct comparison between Divisions 2 and 3 concerning the lower bound on bSQI. 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. Moreover, 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. For the remaining analysis, the lower bound on bSQI was chosen from Division 4 to provide a generalised approach to test the feasibility of the BCLAs methods.

6.2.2. Comparison of results

A first 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 figure 4). 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. Figure 4(a) shows the

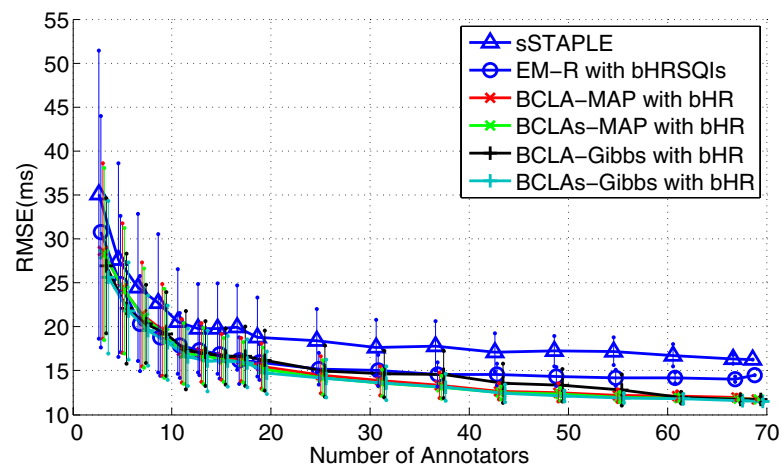


Figure 5. The mean and standard deviation of the RMSE results as a function of the number of annotators for Division 4 in the PCinCQT dataset.

Table 4. RMSEs of using three to 11 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 ^b	26.92 ± 7.66	28.28 ± 9.78 ^b	25.61 ± 8.68 ^a
5	24.86 ± 7.79	24.37 ± 7.37 ^b	22.06 ± 6.28 ^a	24.09 ± 7.13 ^b	21.79 ± 5.56 ^a
7	20.34 ± 5.42	21.41 ± 5.92 ^b	20.31 ± 4.45	20.99 ± 5.63	19.56 ± 4.02
9	18.82 ± 4.42	19.61 ± 5.22 ^b	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 ^a	16.55 ± 3.28 ^a

The ^a indicates results of using BCLA and BCLAs with bHR were significantly different from EM-R with bHRSQIs ($p < 0.05$) using a two-sided Wilcoxon rank-sum test. The ^b shows results where BCLAs-Gibbs with bHR is significantly different from both BCLAs-MAP and BCLA-MAP with bHR using the same test ($p < 0.05$).

results using the MAP approach, and those using the Gibbs sampling approach are shown in figure 4(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 figure 4(a)). BCLA-Gibbs RMSE improvement was also relatively small when the bHR feature was included: less than 60% RMSE improvement was observed when selecting different numbers of annotators (see middle bar plot in figure 4(b)). In the case when the BCLAs model with bHR was included, both the MAP and Gibbs sampling approaches have small but consistent improved performance over their corresponding BCLA models with bHR respectively when selecting any number of annotators (see bottom bar plots in figure 4). This has demonstrated the feasibility of incorporating the signal quality extension, where it has further reduced the overall RMSE errors when selecting any number of annotators.

A comparison of the performance of BCLAs with state-of-art benchmarking approaches like EM-R and sSTAPLE is shown in figure 5. 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. For numbers of annotators varied from three to 11, the RMSE values are shown in table 4. 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 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 bHR provides the best estimation of the inferred ground truth with the lowest RMSE when compared to other methods. Although no signifi-

cant difference was observed when comparing BCLAs with bHR to BCLA with bHR for both MAP and Gibbs approaches, respectively, figure 4 demonstrates that the BCLAs approaches consistently outperform the BCLA approaches. Moreover, the nature of the PCinC QT dataset is that the recordings are of high quality with very little noise. BCLAs is therefore expected to have an even larger benefit for noisier datasets.

7. Conclusion

In this study, a signal quality extension was introduced in Bayesian continuous-valued label aggregator (i.e. BCLAs) to reduce the error of the inferred ground truth when aggregating predictions from algorithms. In the PPG application, the record-specific pSQI as a signal quality metric was applied to the respiratory rate estimation. Although no improvement was observed as the dataset was with little noise, BCLAs guaranteed its performance better than the three bench-marking algorithms (i.e. EM-R, sSTAPLE, and smart fusion), best-performing single algorithm, as well as mean and median voting approaches. 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 estimates. A fully-Bayesian approach (i.e. BCLAs-Gibbs) was shown to be more robust across two datasets and produced the smallest errors. A significant improvement of error using BCLAs-Gibbs was also observed based on a selection of using three algorithms. In summary, incorporating signal quality in the BCLAs model better informs its belief in the annotations provided and hence further improves the estimation of the inferred ground truth.

Acknowledgments

TZ is supported by a Research Fellowship with St. Hildas College, Oxford. DAC is supported by the Royal Academy of Engineering and an EPSRC Challenge Award. YY is supported by the Oxford-Suzhou Centre for Advanced Research. This paper is compliant with RCUK 'open data' policy because the underlying data are all publicly available datasets.

ORCID iDs

Tingting Zhu  <https://orcid.org/0000-0002-1552-5630>

References

- Aboukhalil A, Nielsen L, Saeed M, Mark R G and Clifford G D 2008 Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform *J. Biomed. Inform.* **41** 442–51
- Behar J, Oster J, Li Q and Clifford G D 2013 ECG signal quality during arrhythmia and its application to false alarm reduction *IEEE Trans. Biomed. Eng.* **60** 1660–6
- Bousseljot R and Kreiseler D S A 1995 Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet *Biomed. Tech.* **40** 317–8
- Breakell A and Townsend-Rose C 2001 The clinical evaluation of the Respi-Check mask: a new oxygen mask incorporating a breathing indicator *Emergency Med. J.* **18** 366–9
- Couderc J P, Garnett C, Li M, Handzel R, McNitt S, Xia X, Polonsky S and Zareba W 2011 Highly automated QT measurement techniques in 7 thorough QT studies implemented under ICH E14 guidelines *Ann. Noninvasive Electrocardiol.* **16** 13–24
- Dawid A P and Skene A M 1979 Maximum likelihood estimation of observer error-rates using the EM algorithm *J. Royal Statistical Society C (Applied Statistics)* **28** 20–8
- Drummond G, Bates A, Mann J and Arvind D 2011 Validation of a new non-invasive automatic monitor of respiratory rate for postoperative subjects *Br. J. Anaesthesia* **107** 462–9
- Goldenberg I et al 2006 QT interval: how to measure it and what is 'normal' *J. Cardiovasc. Electrophysiol.* **17** 333–6
- Hughes N P 2006 Probabilistic models for automated ECG interval analysis *PhD Thesis* University of Oxford
- International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use 2014 Guidance for industry E14: clinical evaluation of QT/QTc interval prolongation and proarrhythmic potential for non-antiarrhythmic drugs (<https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm073153.pdf>)
- Johnson A E, Behar J, Andreotti F, Clifford G D and Oster J 2015 Multimodal heart beat detection using signal quality indices *Physiol. Meas.* **36** 1665–77
- Karlen W, Raman S, Ansermino J and Dumont G 2013 Multiparameter respiratory rate estimation from the photoplethysmogram *IEEE Trans. Biomed. Eng.* **60** 1946–53
- Karlen W, Turner M, Cooke E, Dumont G and Ansermino J M 2010 Capnobase: signal database and tools to collect, share and annotate respiratory signals *Annual Meeting of the Society for Technology in Anesthesia (West Palm Beach)* (<https://www.research-collection.ethz.ch/handle/20.500.11850/87887>)
- Li B N, Dong M C and Vai M I 2010 On an automatic delineator for arterial blood pressure waveforms *Biomed. Signal Process. Control* **5** 76–81
- Li Q, Mark R G and Clifford G D 2008 Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter *Physiol. Meas.* **29** 15–32
- Malik M, Färbon P, Batchvarov V, Hnatkova K and Camm A J 2002 Relation between QT and RR intervals is highly individual among healthy subjects: implications for heart rate correction of the QT interval *Heart* **87** 220–8

- Moody G B, Koch H and Steinhoff U 2006 The PhysioNet/Computers in Cardiology Challenge 2006: QT interval measurement *Computing in Cardiology Conf.* (<https://ieeexplore.ieee.org/document/4511851/>)
- Orphanidou C, Brain O, Feldmar J, Khan S, Price J and Tarassenko L 2009 Spectral fusion for estimating respiratory rate from the ECG 2009 9th Int. Conf. on Information Technology and Applications in Biomedicine (<https://ieeexplore.ieee.org/document/5394435/>)
- Petterson M T, Begnoche V L and Graybeal J M 2007 The effect of motion on pulse oximetry and its clinical significance *Anesthesia Analgesia* **105** S78–84
- Pimentel M A F, Santos M D, Springer D B and Clifford G D 2015 Heart beat detection in multimodal physiological data using a hidden semi-Markov model and signal quality indices *Physiol. Meas.* **36** 1717–27
- Raykar V C, Yu S, Zhao L H, Valadez G H, Florin C, Bogoni L and Moy L 2010 Learning from crowds *J. Mach. Learn. Res.* **11** 1297–322
- Tsanas A, Zañartu M, Little M A, Fox C, Ramig L O and Clifford G D 2014 Robust fundamental frequency estimation in sustained vowels: detailed algorithmic comparisons and information fusion with adaptive Kalman filtering *J. Acoust. Soc. Am.* **135** 2885–901
- Vollmer M 2014 Robust detection of heart beats using dynamic thresholds and moving windows *Computing in Cardiology Conf.* (<https://ieeexplore.ieee.org/abstract/document/7043106/>)
- Warfield S K, Zou K H and Wells W M 2008 Validation of image segmentation by estimating rater bias and variance *Phil. Trans. R. Soc. A* **366** 2361–75
- Welinder P and Perona P 2010 Online crowdsourcing: rating annotators and obtaining cost-effective labels *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)* (<https://ieeexplore.ieee.org/document/5543189/>)
- Zhu T, Dunkley N, Behar J, Clifton D A and Clifford G D 2015 Fusing continuous-valued medical labels using a Bayesian model *Ann. Biomed. Eng.* **43** 2892–902
- Zhu T, Johnson A E, Behar J and Clifford G D 2014 Crowd-sourced annotation of ECG signals using contextual information *Ann. Biomed. Eng.* **42** 871–84
- Zong W, Heldt T, Moody G and Mark R 2003 An open-source algorithm to detect onset of arterial blood pressure pulses *Computing in Cardiology Conf.* (<https://ieeexplore.ieee.org/document/1291261/>)