Comparison of Parametric and Non-Parametric Bayesian Inference for Fusing Sensory Estimates in Physiological Time-series Analysis

Tingting Zhu, Hamza Javed, and David A. Clifton

The rapid proliferation of wearable devices for medical applications has necessitated the need for automated algorithms to provide labelling of physiological time-series data to identify abnormal morphology. However, such algorithms are less reliable than gold-standard human expert labels (where the latter are typically difficult and expensive to obtain), due to their large inter- and intra- subject variabilities. Actions taken in response to these algorithms can therefore result in sub-optimal patient care. In a typical scenario where only unevenly-sampled continuous or numeric estimates are provided, without access to the "ground truth", it is challenging to choose which algorithms to trust and which to ignore, or even how to merge the outputs from multiple algorithms to form a more precise final estimate for individual patients. In this work, we demonstrate the novel application of two previously proposed parametric fully-Bayesian graphical models for fusing labels from (i) independent and (ii) potentially-correlated algorithms, validated on two publicly available datasets for the task of respiratory rate (RR) estimation. These unsupervised models aggregate RR labels and estimate jointly the assumed bias and precision of each algorithm. Fusing estimates in this way may then be used to infer the underlying ground truth for individual patients. We show that modelling the latent correlations between algorithms improves the RR estimates, when compared to commonly-employed strategies in the literature. Finally, we demonstrate that the adoption of a strongly-Bayesian approach to inference using Gibbs sampling results in improved estimation over the current state-of-the-art (e.g. hierarchical Gaussian Processes) in physiological time-series modelling.

***Introduction and Related Work**:* With the rapid increase in the volume and variety of wearable devices now routinely in use for healthcare applications, there exists the possibility of personalising the care patients receive based on their individual physiologies. This "personalised" and patient-centric approach to healthcare is built on the assumption that physiological data collected from patient-worn sensors can be reliably utilised, for diagnostic and prognostic applications, in clinical practice. However, with very large quantities of sensor data being accumulated over time, there is an urgent need for algorithms capable of automatically labelling the collected physiological time series data (e.g., abnormal respiratory rate readings) without the need for human input.

Yet to date, automated algorithms remain less reliable in practice than labelling from human experts. The latter is often the accepted gold-standard but is typically expensive, difficult or even unfeasible to obtain for the majority of applications, such as labeling of data arising from patients in real-time. In these cases, and many other real-life clinical applications, automated algorithms have to be relied on to process and label sensor data. Additionally, when there is no knowledge of the "ground-truth" in the form of expert labelling, it is a challenge to know which algorithms to trust and which to ignore at any given point in time. Particularly as different algorithms may be optimal for different patient subsets, or even optimal for the same patient at different points in time. Often, naive methods are used to combine the recommendations of various algorithms to form a final estimate that is intended to have maximum precision for an individual.

Modelling continuous-valued labels in addition to the biases and expertise of each annotator producing those labels, remains an active area of research, with key contributions outlined as follows. In the context of medical imaging [1], the use of an expectation maximisation (EM) method was demonstrated to fuse labels from different annotators estimating the diameter of lesions from images. A method for validating medical image segmentation, which estimated both the bias and variance of annotators, was proposed in the work of [2]. Similar to this approach, [3] more recently presented a model that estimated the ground truth in the form of count and percentage estimation, in a "crowd sensing" setting. A Bayesian EM framework fused binary, multi-valued and continuous-valued labels was proposed in [4]. This method described explicitly modelled the precision (but not bias) of individual annotators by taking into account their different skill levels. By contrast, [5] used a Gaussian prior on the bias parameter of annotators attempting to produce cardiac landmark labels in 2D images. However, it is worth noting that physiological features were not incorporated into the models of [5] as a means of further improving estimation of the ground truth label.

In all of the aforementioned studies, the proposed models did not include a principled way to take into account the quality of data or how to cater for missing labels. Moreover, in these studies it was assumed that all annotators are independent, which may not always be the case when labels are produced by slightly different implementations of the same underlying algorithm. Previous work by the author tackled these issues by first proposing a Bayesian framework to jointly model both annotator bias and precision using [6]. This work was then extended in [7], in which the author proposed a fully-Bayesian approach through Gibbs sampling for fusing continuous valued labels, from both independent and/or partially correlated annotators, as a means of arriving at a consensus in an unsupervised manner.

In this letter, we present a novel application of the methodologies proposed by the author in [7], on two publicly available datasets. The task considered is estimation of the underlying respiratory rate (RR) from photoplethysmogram (PPG) recordings contained within the CapnoBase and BIDMC datasets [8] [9]. Robust estimation of RR is a practically well motivated task, as accurate monitoring of the vital sign can facilitate improved diagnosis and patient care. We demonstrate improved estimation of RR is possible using our approach of fusing labels from different annotators, when compared with existing methods presented in the literature; namely two EM models by [1] and [2], as well as a Hierarchical Gaussian process approach [10].

The remainder of this letter is organised as follows. First we outline the methodology proposed by the authors in [7], briefly describing the formulation of the two models considered. The experiments used to validate and compare the methods with selected baselines, along with the results obtained, are then detailed before concluding remarks are presented.

***Problem Formulation**:* Consider the case where we have N samples of physiological time-series data, with N corresponding continuous-valued labels (e.g. RR labels from PPG time-series samples). We can assume that the underlying ground truth for the ith sample, z_i, can be drawn from a Gaussian distribution with mean a_i and variance $1/b$. We can express a_i as a linear regression function $f(\mathbf{w}, \mathbf{x}_i)$ with an intercept w_0. In this formulation $\mathbf{w}$ are the coefficients of the regression (which includes w_0[1]). While $\mathbf{x}_i$ is a column feature vector for the ith record containing d features (i.e., we have an $(N \times d)$-dimensional design matrix, $\mathbf{X} = \left[\mathbf{x}_1^\intercal; ...; \mathbf{x}_N^\intercal\right]$). Note that, a scalar value of one was added to the feature matrix (i.e., $\mathbf{x}_i := [1, \mathbf{x}_i]$)) to cater for the w_0 intercept. Finally, the precision of the ground truth (defined as the inverse-variance b) is assumed to be modelled from a gamma distribution where k_b is the shape parameter and ϑ_b is the scale parameter. It therefore follows that the conditional probability density function (pdf) of $\mathbf{z}$ as a vector of labels can be written as $\prod_{i=1}^{N} \mathcal{N}\left(z_i \mid \mathbf{x}_i^\intercal \mathbf{w}, 1/b\right)$.

The Independent Annotator Model (IAM): Assuming once again the presence of N samples, we have a dataset, $\mathbf{D} = \left[\mathbf{x}_i^\intercal, y_i^{j=1}, \cdots, y_i^{j=R}\right]_{i=1}^{N}$, where y_i^j corresponds to the label estimate provided by the jth annotator for the ith sample, with a total of R annotators. This model assumes that y_i^j is a noisy version of z_i, with a Gaussian distribution $\mathcal{N}\left(y_i^j \mid z_i, 1/\lambda^j\right)$, where λ^j is the precision of the jth annotator, defined as the estimated inverse-variance for annotator j. Furthermore, the bias of each annotator, which measures the average difference between the estimation and the ground truth, can be modelled as an additional term, denoted as ϕ^j. The pdf of estimating y_i^j can thus be written as $\mathcal{N}\left(y_i^j \mid z_i + \phi^j, 1/\lambda^j\right)$, It is assumed that $y_i^1, \cdots, y_i^R$ are conditionally independent given the ground truth z_i; assuming samples are independent, it follows that the conditional pdf of $\mathbf{y}$ can be expressed as :

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

However, as noted earlier, conditional independence between annotators may not always be the case as labels may be produced by variants of the same underlying algorithmic approach. That is annotators that differ only in, for example, operational parameter settings. Nevertheless, this assumption can be made to simplify the model and subsequent derivation of the likelihood. Relaxation of this independence assumption will be explored in the second proposed model, the correlated annotator model (described in the proceeding section). The pdf of the bias for annotator j, ϕ^j, is assumed to once again be drawn from a Guassian this time with mean μ_ϕ and variance $1/\alpha_\phi$ [5]:

$$p\left(\phi^j \mid \mu_\phi, \alpha_\phi\right) = \mathcal{N}\left(\phi^j \mid \mu_\phi, 1/\alpha_\phi\right). \quad (2)$$

Although the biases of the annotators could very well be assumed to follow other distributions, such choices are likely to be dataset-dependent. In the absence of any knowledge of the underlying distribution of biases, we choose to assume they are drawn from a Gaussian distribution. The precision values, such as λ^j and α_ϕ, by contrast are assumed to be drawn from a gamma distribution, with parameters k_λ, ϑ_λ, and k_α, ϑ_α, respectively:

$$p\left(\lambda^j \mid k_\lambda, \vartheta_\lambda\right) = \mathrm{Gamma}\left(\lambda^j \mid k_\lambda, \vartheta_\lambda\right). \quad (3)$$

$$p\left(\alpha_\phi \mid k_\alpha, \vartheta_\alpha\right) = \mathrm{Gamma}\left(\alpha_\phi \mid k_\alpha, \vartheta_\alpha\right). \quad (4)$$

It follows that for a given dataset $\mathbf{D}$, the likelihood of the parameters $\boldsymbol{\theta} = \{\mathbf{w}, \boldsymbol{\lambda}, \boldsymbol{\phi}, \alpha_\phi, b, z_i\}$, can be formulated as :

$$p\left(\mathbf{D} \mid \boldsymbol{\theta}\right) = \prod_{i=1}^{N} p\left(y_i^1, \cdots, y_i^R \mid \mathbf{x}_i, \boldsymbol{\theta}\right). \quad (5)$$

Bayes' theorem can then be used to determine the posterior probability of the parameters $\boldsymbol{\theta}$, for a given dataset $\mathbf{D}$, as

$$p\left(\boldsymbol{\theta} \mid \mathbf{D}\right) = \frac{p\left(\mathbf{D} \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right)}{\int_{\boldsymbol{\theta}} p\left(\mathbf{D} \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta}}, \quad (6)$$

[1] w_0 models the overall offset predicted in the regression, and is therefore different from the bias ϕ specific to each annotator in the proposed models, which will be described later.

where

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

Obtaining the posterior probability of the parameters $\boldsymbol{\theta}$ essentially allows us to learn the latent ground truth for the ith sample z_i, and jointly predict the bias ϕ^j and precision λ^j of the jth annotator simultaneously.

Learning from Incomplete Data using Gibbs Sampling: An important practical scenario to consider is the case that arises when there are missing labels from different annotators (i.e., not all R algorithms provide N estimates for all samples). To account for this, the posterior distribution hyperparameters of the IAM can be re-written using Gibbs sampling as follows (see graphical model in Figure 1(a)):

$$z_i \sim \mathcal{N}\left(z_i \middle| a_i^*, \frac{1}{b_i^*}\right), \phi^j \sim \mathcal{N}\left(\phi^j \middle| \mu_\phi^{j*}, \frac{1}{\alpha_\phi^{j*}}\right),$$
$$\lambda^j \sim \mathrm{Gamma}\left(\lambda^j \middle| k_\lambda^{j*}, \vartheta_\lambda^{j*}\right),$$
$$b \sim \mathrm{Gamma}\left(b | k_b^*, \vartheta_b^*\right), \alpha_\phi \sim \mathrm{Gamma}\left(\alpha_\phi \mid k_\alpha^*, \vartheta_\alpha^*\right).$$

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

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

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

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

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

Note that U_j is the set of samples with labels provided by the jth annotator whilst V_i is the set of annotators that provided labels for the ith sample, and N_j is the number of samples annotated by the jth annotator. Finally, $\mathbf{w}$ can be learnt by finding the zero gradient of the expectation of the complete data log-likelihood as $\mathbf{w} = \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^\intercal\right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i z_i$. The above formulation allows us to cope robustly with the commonly-encountered difficulties arising from incomplete (or even sparse) labelling, in a principled and probabilistic manner.

The Correlated Annotator Model (CAM): As noted previously, annotator independence may not always be an accurate assumption to make in reality. To account for this, we can incorporate a correlation measure into the annotator model described in the preceding section. This would facilitate an improved aggregation of the different annotator labels, and thus a better inferred ground truth estimate. In this formulation, annotators are considered to be anomalous when they are highly correlated to other annotators but possess relatively large variances and biases. These anomalous annotators are penalised with lower weighting for their labels. Expert annotators, defined as those that are highly correlated to other annotators but which have relatively small variances and biases, on the other hand have their labels weighted more heavily in the model.

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

$$p(\mathbf{y} \mid z_i, \boldsymbol{\phi}, \boldsymbol{\Sigma}) = \prod_{i=1}^{N} \mathcal{N}(z_i + \boldsymbol{\phi}, \boldsymbol{\Sigma}), \tag{7}$$

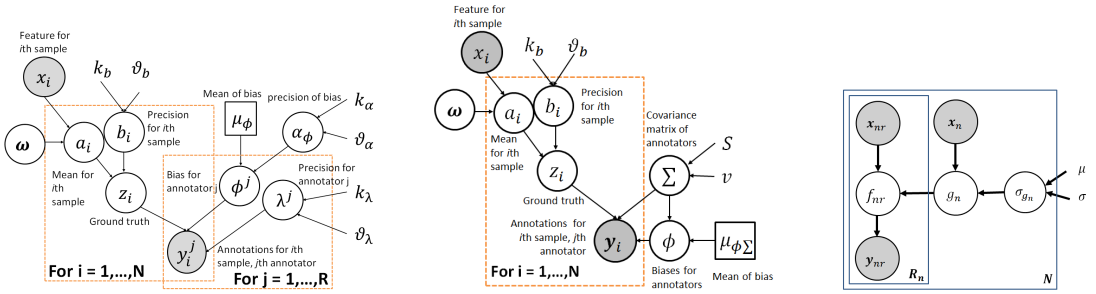
where $\boldsymbol{\Sigma}$ is the covariance matrix of the R annotators and where there are N samples.

Matrix $\boldsymbol{\Sigma}$ can be further decomposed into a correlation matrix and the precision values of the annotators. Using the separation strategy proposed by [11], $\boldsymbol{\Sigma}$ is formulated as $\boldsymbol{\Sigma} = \mathbf{Q}\boldsymbol{\rho}\mathbf{Q}$, where $\mathbf{Q}$ is an R-by-R diagonal matrix with entries being $\frac{1}{\sqrt{\lambda^{j=1}}}, ..., \frac{1}{\sqrt{\lambda^{j=R}}}$. Here, λ^j is the precision value for the jth annotator, and $\boldsymbol{\rho}$ is the latent correlation matrix of the annotation errors among R annotators. The biases of individual annotators are now assumed to be drawn from a MVN constrained by $\boldsymbol{\Sigma}$, with conditional pdf:

$$p\left(\boldsymbol{\phi} \mid \boldsymbol{\mu}_{\phi\Sigma}, \boldsymbol{\Sigma}\right) = \mathcal{N}\left(\boldsymbol{\phi} \mid \boldsymbol{\mu}_{\phi\Sigma}, \boldsymbol{\Sigma}/k_0\right), \tag{8}$$

where $\boldsymbol{\mu}_{\phi\Sigma}$ is the prior mean for $\boldsymbol{\phi}$, and k_0 is a positive scalar that expresses our belief on $\boldsymbol{\mu}_{\phi\Sigma}$. The posterior of the parameter $\boldsymbol{\theta}_c = \{\boldsymbol{\phi}, \boldsymbol{\Sigma}, b, z_i\}$ for a given dataset $\mathbf{D}$ can be written using Bayes' theorem as

$$p(\boldsymbol{\theta}_c \mid \mathbf{D}) = \frac{p(\mathbf{D} \mid \boldsymbol{\theta}_c)\, p(\boldsymbol{\theta}_c)}{\int_{\boldsymbol{\theta}_c} p(\mathbf{D} \mid \boldsymbol{\theta}_c)\, p(\theta)\, d\boldsymbol{\theta}_c}, \tag{9}$$



(a) (b) (c)

Fig. 1: (a) The independent annotator model. (b) The correlated annotator model. (c) Hierarchical Gaussian Processes with an additional prior on the latent ground truth.

where:

$$p\left(\mathbf{D} \mid \boldsymbol{\theta}_c\right) p\left(\boldsymbol{\theta}_c\right)$$

$$= \mathcal{N}\left(\boldsymbol{\phi} \mid \boldsymbol{\mu}_{\phi\Sigma}, \boldsymbol{\Sigma}/k_0\right) \mathrm{IW}\left(\boldsymbol{\Sigma} \mid v, S\right) \times$$

$$Gamma\left(b \mid k_b, \vartheta_b\right) \left[\prod_{i=1}^{N} \mathcal{N}\left(z_i \mid a_i, 1/b\right) \mathcal{N}\left(\mathbf{y_i} \mid z_i + \boldsymbol{\phi}, \boldsymbol{\Sigma}\right)\right]$$

The new parameters are now updated using the Gibbs sampler as follows (see graphical model in Figure 1(b)) :

$$\boldsymbol{\phi} \sim \mathcal{N}\left(\boldsymbol{\phi} \mid \boldsymbol{\mu}^*_{\phi\Sigma}, \boldsymbol{\Sigma}^*_{\phi}\right), \boldsymbol{\Sigma} \sim \mathrm{IW}\left(\boldsymbol{\Sigma} \mid v^*, \mathbf{S}^*\right).$$

$$\boldsymbol{\mu}^*_{\phi\Sigma} = \frac{k_0 \boldsymbol{\mu}_{\phi\Sigma}}{k_0 + N} + \frac{\mathbf{U}\bar{\mathbf{y}}_b}{k_0 + \mathbf{U}}, \boldsymbol{\Sigma}^*_{\phi} = \frac{\boldsymbol{\Sigma}}{k_0 + N}, v^* = v + N,$$

$$\mathbf{S}^* = \mathbf{S} + \sum_{i=1}^{N} \left(\mathbf{y}_i - z_i - \bar{\mathbf{y}}_b\right)^T \left(\mathbf{y}_i - z_i - \bar{\mathbf{y}}_b\right)$$

$$+ \frac{k_0 N}{k_0 + N} \left(\bar{\mathbf{y}}_b - \boldsymbol{\mu}_{\phi\Sigma}\right)^T \left(\bar{\mathbf{y}}_b - \boldsymbol{\mu}_{\phi\Sigma}\right).$$

where $\mathbf{U}$ is a 1-by-R vector, and each of its elements indicates the total number of labels provided by a respective annotator. $\bar{\mathbf{y}}_b = [\bar{y}_b^{j=1}, \cdots, \bar{y}_b^{j=R}]$, where $\bar{y}_b^j = \frac{1}{N_j} \sum_{i=1}^{N} \left(y_i^j - z_i\right)$.

Experiment and Results: We evaluate the efficacy of our proposed models using two publicly-available biomedical datasets: (1) The CapnoBase dataset by [8] which contains 42 PPG recordings (each with 8 minutes duration) of spontaneous or controlled breathing from 42 subjects (29 paediatric and 13 adults); (2) The BIDMC dataset by [9] which comprises PPG recordings with the same duration from 53 adult subjects. The three respiratory-induced modulation time-series (Amplitude Modulation, Baseline Wander, and Frequency Modulation) were extracted from the PPG recordings. To estimate the RR, it was computed for 32-second windows, with successive windows having 29 seconds overlap, using a Fourier spectral approach. For each window and each modality, three RR estimates were calculated from three modulation time-series. The underlying subject-specific latent RR was then estimated by fusing these 6 "algorithms".

We compared our proposed models with two parametric Maximum-likelihood EM models (EM-R by [1]; STAPLE by [2]), as well as the non-parametric Hierarchical Gaussian Processes (HGPs) ([10]) with an additional Bayesian regularisation (i.e., a lognormal prior) on the noise variance of the latent ground truth (see Figure 1(c)). By comparing the gold-stand RR labels for a subject over 150 windows, the mean absolute error (MAE) was computed for each model. The mean MAE and the standard error of the mean (SEM) were also estimated across all subjects. The results are shown in Table 1. The CAM had the least error for CapnoBase, but the IAM model was better for BIDMC. Nevertheless, both proposed models outperformed the state-of-the-art approaches recreated from the literature: a MAE of 1 bpm vs. 1.5 bpm [9] and 1.2 bpm [8] for the CapnoBase dataset, and a MAE of 2.96 bpm vs. 4 bpm [9] and 5.8 bpm [8] for the BIDMC datase using all possible windows. Furthermore, the proposed models provide the Bayesian interpretation of their estimates through 95% confidence intervals (see dashed lines in Figure 2). In comparison, HGPs had a larger noise variance: as 5 out of 6 algorithms (A2 to A6) were biased with smaller estimation of RR, this resulted a large uncertainty in the latent RR estimates when fusing labels using HGPs.

Conclusion: Automated labelling of large volumes of physiological time-series data being collected from wearable sensors, often in real-time, is a prerequisite to being able to provide patients with personalised care. In this work we have applied two parametric unsupervised fully-Bayesian graphical models for fusing labels from (i) independent and (ii) potentially-correlated algorithms, to estimate the underlying RR from PPG signals obtained from the publicly available CapnoBase and BIDMC datasets. Robust estimation

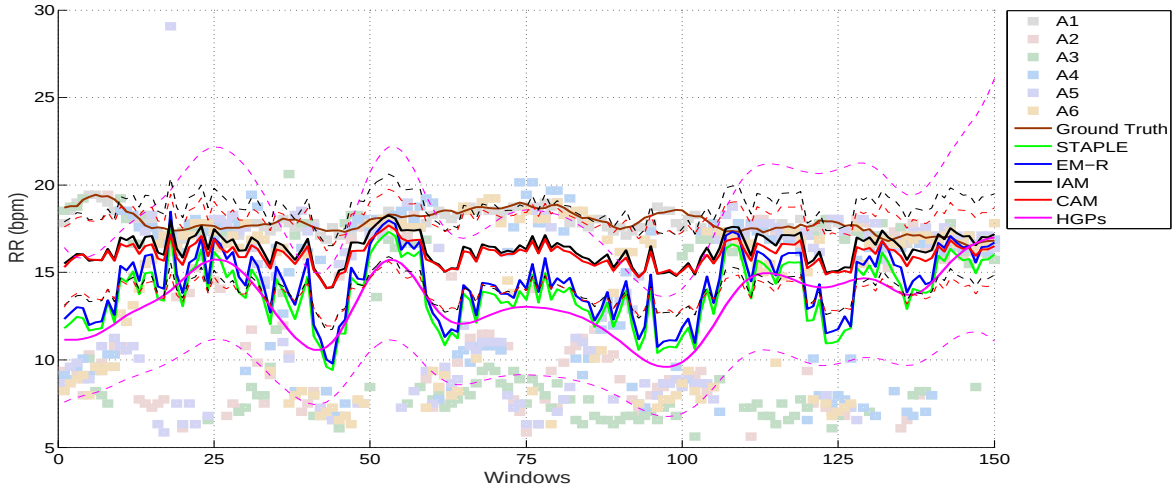


Fig. 2: Example of the RR estimates for a subject.

Table 1: Mean MAE±SEM (bpm) of the inferred RR across subjects using different models for CapnoBase and BIDMC datasets.

Model	CapnoBase	BIDMC
EM-R	1.14±0.22	3.15±0.34
sSTAPLE	1.78±0.34	3.51±0.28
HGPs	1.46±0.32	3.37±0.40
IAM	1.18±0.16	2.96±0.43
CAM	1.00±0.20	3.03±0.33

of RR is of clinical value and could be used to improve patient care. By jointly estimating the assumed bias and precision of each algorithm considered, we have demonstrated that these models are able to infer the underlying ground truth more robustly than existing state of the art methods. In addition to improved performance, we show that the proposed models are robust when dealing with missing values (as often occurs in real-life biomedical applications due to sensor failure), and that they are suitably efficient for use in real-time applications.

References

1 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.

2 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.

3 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: ACM, 2015, pp. 190–201.

4 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*, 2010, pp. 25–32.

5 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*, 2011, pp. 7962 –796 206.

6 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.

7 T. Zhu, M. A. F. Pimentel, G. D. Clifford, and D. A. Clifton, "Unsupervised bayesian inference to fuse biosignal sensory estimates for personalizing care," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 1, pp. 47–58, Jan 2019.

8 W. Karlen, S. Raman, J. Ansermino, and G. Dumont, "Multiparameter Respiratory Rate Estimation From the Photoplethysmogram," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 7, pp. 1946–1953, July 2013.

9 M. A. Pimentel, A. E. Johnson, P. H. Charlton, D. Birrenkott, P. J. Watkinson, L. Tarassenko, and D. A. Clifton, "Toward a robust estimation of respiratory rate from pulse oximeters," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 8, pp. 1914–1923, 2017.

10 T. Zhu, G. W. Colopy, C. Macewen, K. Niehaus, Y. Yang, C. W. Pugh, and D. A. Clifton, "Patient-Specific Physiological Monitoring and Prediction Using Structured

Gaussian Processes," *IEEE Access*, vol. 7, pp. 58 094–58 103, 2019.

11 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.