

Likelihood-based artefact detection in continuously-acquired patient vital signs

Glen Wright Colopy*, Tingting Zhu*, Lei Clifton**, Stephen J. Roberts*, David A. Clifton*

**Department of Engineering Science, University of Oxford, Oxford, UK,*

***Centre for Statistics in Medicine, University of Oxford, Oxford, UK*

Abstract—Robust continuous monitoring of patient vital signs (VS) is limited by artefactual data yielding measurements that are not representative of the patient’s physiology. These artefacts are typified by several distinct “archetypes”. We present several of these archetypal artefacts for heart rate (HR) monitoring, and propose a light weight, real-time algorithm to remove the majority of these artefacts. Most artefacts are not identifiable by their values in absolute terms, but instead by their values relative to other measurements nearby in time. We model temporally-proximate measurements as independent and identically distributed (i.i.d.) samples from a Gamma distribution. Measurements with low likelihood with respect to the distribution are candidates for artefact removal. This light-weight algorithm is important for real-time deployment on wearable sensors, which are becoming increasingly common in hospital and home care. The clinical applicability of artefact-removal is demonstrated in its ability to enhance patient deterioration detection. A Kalman filter-based patient monitoring algorithm is shown to improve early warning of deterioration when the proposed artefact-removal algorithm is used. We demonstrate this real-time system with patient data from a clinical trial that we have undertaken.

I. CLINICAL NEED

The automation of patient vital-sign monitoring has several desirable features, including continuity, the avoidance of human error, and the capacity to estimate patient health states. Fully-automated systems are hampered by inappropriate handling of artefactual data that is acquired by the monitoring devices. Artefactual measurements may result from many causes [1] including probe detachment, algorithmic failure (e.g., missed or extra beats in a heartbeat-detector), or signal interference (e.g., from movement or perspiration). Regardless of origin, artefacts are vexing to statistical inference with vital-sign data because they are not representative of the patient’s physiology at the time in which the patient’s physiology was measured. If not handled, these measurements may interfere with the machine’s inference with regard to statistical descriptions of the data and which, in turn, yield poor estimates the patient’s physiological status.

Common responses to the presence of artefacts are (i) further smoothing of the vital-sign measurements

under consideration, and (ii) upper and lower thresholds to remove physiologically-implausible values. For example, [2] removes HR values outside of the 30-300 bpm range. Both methods have short-comings: common smoothing techniques (e.g., mean/median imputation) introduce bias to statistical descriptions of the time-series by obscuring the imprecision of the monitoring device (or the short-term variance of the time-series itself), which may be clinically informative. Simple upper and lower thresholds are routinely used to remove extremal values, but such measurements are typically in the minority of measurement artefacts. The remaining majority of artefacts would continue to interfere with clinical inference and which result in a lack of robustness for many clinical inference systems.

Likely motivations to use less-principled instead of more principled methods [3], are algorithmic simplicity, experimental design, and the types of data available (e.g., values of HR without the original sensor waveform - such as the ECG). The proposed method is attractive in both its lightweight implementation (making it suitable for inclusion in wearable monitors) and applicability in the absence of waveform data.

II. DATA

We study vital-sign data from a study of 336 patients in a step-down ward at the University of Pittsburgh Medical Center [2]. Patients’ time-series were retrospectively validated by clinical experts for clinical emergency events (which are extreme non-artefactual measurements that would warrant the intervention of an emergency medical team).

To assess the discriminative ability of our approach, 20 patients without validated clinical emergency events were selected for manual annotation of HR artefacts. The annotation process yielded 80 artefacts (<0.01%) exceeding the 30-300 bpm thresholds suggested by [2], a further 20,982 artefacts (14.9%) within 30-300 bpm, and 119,494 non-artefacts (85.1%) from 140,556 total measurements. As a demonstration of the artefact removal method’s clinical utility, 59 patients with emergency events, and 89 patient without emergency events were

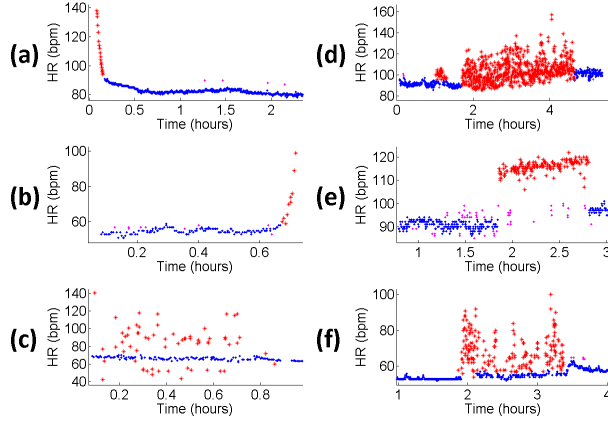


Figure 1. An example time-series containing 6 archetypal HR artefacts. Examples contain non-artefacts (•), example artefacts (*), and plausible artefacts of a different archetype (•). Archetypes (a)-(c) are transient, whereas (d)-(f) are more persistent, suggesting that they likely differ in the appropriate artefact removal method.

used to assess the improvement of an early warning system, before and after artefact removal.

III. ARTEFACT ANNOTATION

The time series of 20 patients were inspected for likely artefacts. Each patient had between 1 and 48 hours of HR recording at acquisition rate of about $\frac{1}{5}$ Hz to $\frac{1}{3}$ Hz.

Figure 1 shows six archetypal artefacts in continuous HR time-series. We propose a method to detect transient artefacts, such as those archetypes shown in 1(a), 1(b), and 1(c). The proposed method is less appropriate for artefacts described in 1(d), 1(e), and 1(f), since the latter are non-transient. Plausible explanations for 1(d)-(f) include partial probe-detachment, algorithmic failure to handle unusual waveforms, or actual arrhythmia. Probe attachment and detachment artefact are shown in 1(a) and 1(b), respectively. In these two cases, HR values show a precipitous rise/fall, which is non-physiological. The artefacts are further contextualised by the absence of measurements before (a) or after (b), indicating that the monitoring device has just been (de-)attached. In 1(c), artefacts are contrasted with the consistent HR around 70 bpm, which are implausible variations over this time-scale. Less apparent artefacts (after $t = 0.8$ hours) are much closer to 70 bpm. This is useful to illustrate (i) the subjectivity, and (ii) the potential source of human error in annotating less-obvious artefacts.

The number and proportion of artefacts and non-artefacts for each patient in our study are shown in figure 2. The large proportion of annotated artefacts emphasise how artefacts are more frequent than generally believed, but many are not so extreme as to be noticed by inspection, such as those shown after $t = 0.8$ hours in 1(c).

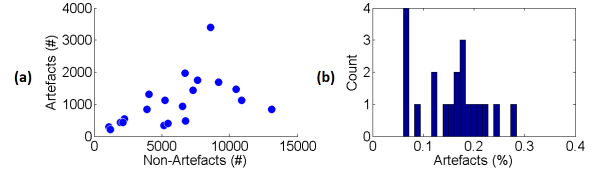


Figure 2. For each patient, the number of annotated artefactual and non-artefactual measurements is shown in (a). The total number of HR measurements per patient ranged from 1,390 to 13,950. The proportion of artefactual measurements as part of total number of HR measurements is shown in (b).

IV. ARTEFACT MODELLING

We observe from the examples shown in figure 1 that artefactual values are outstanding in the context of nearby points (or lack thereof). It is intuitive to detect artefacts with respect to a probability distribution function (PDF) of measurements in close temporal proximity.

We here aim for a computationally-lightweight method that might be embedded within a wearable sensor. A suitable approach is to model all measurements within a short window of duration τ as i.i.d. draws from a gamma distribution:

$$p(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right), \quad x > 0, \quad (1)$$

where the gamma distribution is parameterised by α , a shape parameter, and β , a scale parameter. Incorporation of a priori knowledge for α and β is undesirable given the large variability in the value and volatility of the HR measurements across time and across patients. Therefore, α and β are fit to their maximum likelihood estimate (MLE), that is, the values of α and β that maximise (1) with respect to the data, x , within window τ . The gamma PDF has several desirable attributes, including support only along the real positive line, and smaller cumulative density in the tail than alternative distributions.

We define a window to be of duration $\tau = 5$ minutes and calculate an artefact score, $z(x)$, for each point x as

$$z(x) = \frac{1}{w} \sum_{i=1}^w \log p_i(x) \quad (2)$$

where w is the total number of windows of length τ that include point x , as illustrated in figure 3. The feature that distinguishes artefacts from non-artefacts is that over a small time window, the former will have low values of $z(x)$ compared with those of the latter. Across w small windows, the average log-likelihood, $z(x)$, of an outlying measurement becomes well-separated from its neighbors. Since the score $z(x)$ is on a log-likelihood scale, small increments are sufficient to differentiate between artefactual and non-artefactual data.

V. DISCRIMINATIVE ABILITY OF ARTEFACT SCORE

Figure 4 demonstrates the discriminatory ability of the proposed light-weight method. We examine the trade-off between the proportion of artefacts and non-artefacts that would be discarded at any given threshold k on the artefact score $z(x)$. Results are shown both for inter-patient variability (a) and upon aggregating all artefactual and non-artefactual measurements across all patients (b).

Despite inter-patient variability, it may be seen that the artefact score effectively differentiates between artefacts and non-artefacts. For example, at threshold $k = -4$, half of all patients would have at least 35% of all artefacts removed, and almost no non-artefacts removed. The top 20% of patients would have at least 50% of artefacts removed, and the top 80% of patients would have at least 30% of artefacts removed. On aggregate, shown in figure 4(b), about 40% of artefacts and <1% of non-artefacts had a score below threshold $k = -4$.

These results lead to several conclusions: firstly, it is possible to remove a large portion of artefacts while removing only a negligible proportion of the non-artefactual points (which we would like to keep for clinical inference). Secondly, although there is inter-patient variability in the proportion of artefacts removed at a given threshold, the variability is modest, suggesting that a robust artefact score threshold may be developed for a cohort of patients, suitable (for example) for embedding within a wearable sensor.

The proposed method compares favorably to simple thresholding methods, such as those suggested by [2]. Computationally, both methods clean an hour of data in less than a second. Although the extreme range of 30-300 bmp virtually assures that all artefacts identified are true artefacts, these instances are less than 1% of the total artefact population. Further tuning of the thresholds degenerates quickly since heterogeneous patient populations exhibit a wide range of values.

VI. ARTEFACT REMOVAL IN PATIENT DETERIORATION DETECTION

The utility of artefact removal for patient monitoring is illustrated in figure 5. Time-series modeling (e.g. via Gaussian processes or linear state-space models) provides an effective means by which to identify patient deterioration in the step-down ward. As demonstrated in [4], time-series forecasts may quantify step-changes in vital-sign time-series. This works by training a model using a window representing the patient's current time-series, and then forecasting. By evaluating the likelihood of subsequently-observed vital-sign values with respect to the predictive distribution of the forecast, we may quantify a patient's deviation from expectation. Large deviations suggest a deteriorating physiology.

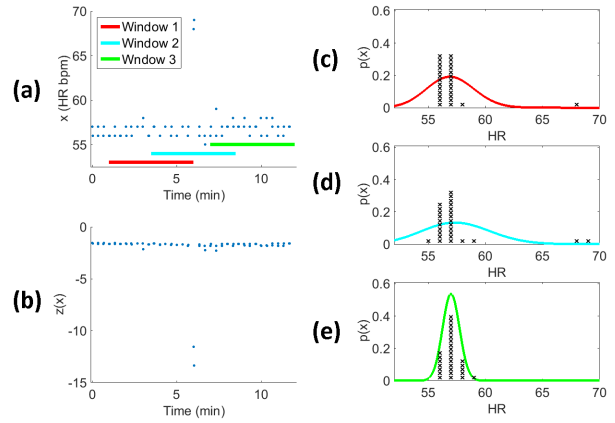


Figure 3. In (a), HR measurements are shown over the course of 12 minutes of a patient's HR time series. Two HR artefact are visible at $t = 6$ minutes. The time-span of three example 5-minute windows are shown, which determine the current sample of measurements. A new window originates at every acquired measurement, so nearby windows have substantial overlap in measurements. For each HR value, x , we would like to calculate $z(x)$ as shown in (b). For the measurements within each window, the gamma distribution is fit using MLE. The corresponding PDFs are shown in (c), (d), and (e), with colors corresponding to the respectively colored windows in (a). Each x in (c-e) is an HR measurement within the window. Note that the artefactual measurements in the tails of (c) and (d) are sufficient to significantly flatten the PDF over the window of measurements, and that these tail data have likelihoods that are significantly lower than the likelihood of the other measurements in the same window. In contrast, the artefact-free data in (e) is well-described by the fitted PDF. As described by equation 2, the average of these logged PDF values, across all windows, yields $z(x)$ in (b).

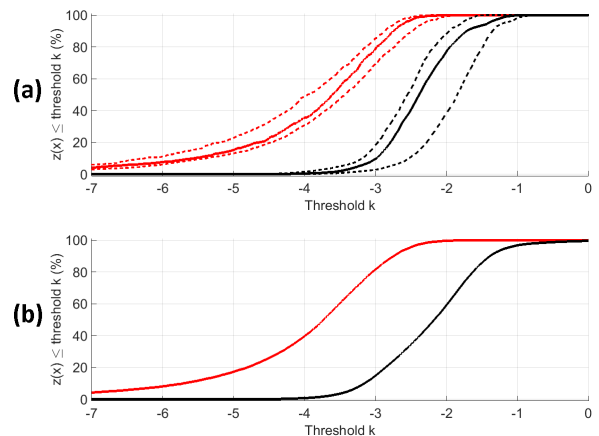


Figure 4. Ability to discriminate between artefact (—) and non-artefact (---), assessed on (a) the per-patient level, and (b) the population level. In (a), the cumulative density below a threshold $z(x) = k$ was calculated for each of the 20 patients under consideration. Due to inter-patient variability, the 20%, 50%, and 80% quantiles across all per-patient results are shown. In (b), the cumulative density is shown for all measurements aggregated across patients. Using such plots, an acceptable trade-off for a particular application may be identified, and the corresponding threshold k may be determined.

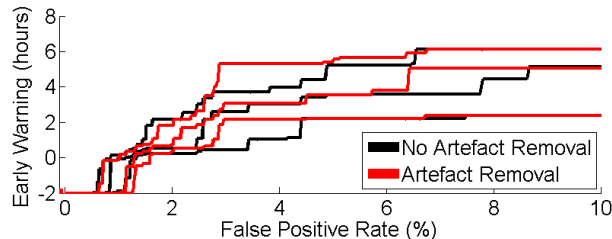


Figure 5. Improving patient monitoring with artefact removal. GP deterioration detection is performed with (—) and without (---) artefact removal, using threshold of $k = -4$. Increased sensitivity improves the hours of early warning in advance of the annotated clinical emergency, but also increases the rate of false positive alarms in patients without clinical emergencies. Lines show the 15%, 20%, and 25% quantiles of early warning among the 59 patients with clinical emergencies.

The presence of artefacts hinders this inference by (i) influencing the probabilistic model fit to the time-series data (typically by making the training data appear more volatile than the patient’s true physiology), and (ii) providing false evidence of deterioration (since artefactual measurements are difficult to predict, and therefore lower the likelihood of the forecast). Artefact removal may improve patient monitoring by removing measurements with which the monitoring system should not perform inference.

The improvement after artefact removal is shown in figure 5, using 59 patients with annotated emergencies and 89 patients with no annotated emergencies. The time of early warning is the warning given in advance of the first annotated clinical emergency for each of those 59 patients with an annotated emergency.

While patient monitoring with artefact removal showed nearly identical performance for higher quantiles (not shown), artefact removal showed marked improvement for patients for the lower quantiles. This is a positive result, since we conclude that artefact removal improves early warning in those patients who are currently receiving the least warning. For example, at a false position rate of 3%, the lowest 15% of patients without artefact removal (shown by the bottom-most lines of figure 5) had an early warning of 0 hours, that is, the warning only occurred at the time of emergency. In contrast, those same patients would have had 2 hours of early warning using artefact removal. The improvement of advanced warning from 0 to 2 hours could be invaluable for clinical staff, and potentially life-saving for a patient.

VII. CONCLUSION

A fast and computationally lightweight method has been proposed. The method is robust between patients, as seen by the large proportion of artefacts that may be removed at the expense of few non-artefactual measurements. Furthermore, artefact removal is a practical

way in which to improve patient monitoring outcomes by increasing early warning of deterioration, and reducing the rate of false alarms, which distract and desensitise clinical staff via “alarm fatigue”.

VIII. FUTURE WORK

The lightweight nature of the proposed technique invites significant possibilities for future work by (i) application across other vital-signs, and (ii) more complex artefact modelling, performed prior to inference in settings where computational resources would permit it.

The proposed method is applicable to vital-signs other than HR, such as breathing rate, blood oxygen saturation, and blood pressure measurements. Different vital-signs could be cleaned via artefact detection on an individual vital-sign basis, or jointly as some sources of artefact (e.g. signal corruption due to movement) may simultaneously affect multiple vital-sign measurements.

Superior artefact modelling should first address the i.i.d. assumption, which underlies the current approach. A time-series model of vital-sign data would explicitly model the temporal correlation between measurements, and incorporate a greater number of measurements to estimate the latent mean and variance at the time of measurement. These models may also incorporate derivative observations [5], change-point detection [6] [7], and regularising priors to incorporate patient physiology.

From a supervised learning perspective, a model-driven approach to artefact detection might define a set of artefact archetypes (e.g., any of those in figure 1) and incorporate a set of derived features from the time series to identify these specific archetypes.

ACKNOWLEDGMENTS

GWC was supported by the Clarendon fund and EPSRC. SJR gratefully acknowledges a Royal Academy of Engineering / Man-AHL Research Chair. DAC was supported by the Royal Academy of Engineering; Balliol College, Oxford; and an EPSRC “Challenge Award”.

REFERENCES

- [1] G. Friesen *et al.*, “A comparison of the noise sensitivity of nine QRS detection algorithms,” *IEEE Transactions on Biomedical Engineering*, vol. 37, no. 1, pp. 85–98, 1990.
- [2] A. Hann, “Multi-parameter monitoring for early warning of patient deterioration,” Ph.D. dissertation, University of Oxford, 2008.
- [3] J. Quinn *et al.*, “Factorial switching linear dynamical systems applied to physiological condition monitoring,” *IEEE TPAMI*, vol. 31, no. 9, pp. 1537–1551, 2009.
- [4] G. W. Colopy *et al.*, “Bayesian Gaussian processes for identifying the deteriorating patient,” *38th Intl. Conf. of the IEEE EMB Society (EMBC), Orlando, FL*, pp. 5311–5314, 2016.
- [5] E. Solak *et al.*, “Derivative observations in Gaussian process models of dynamic systems,” 2003.
- [6] R. Darby Turner, “Gaussian processes for state space models and change point detection,” Ph.D. dissertation, University of Cambridge, 2011.
- [7] S. Reece *et al.*, “Anomaly detection and removal using non-stationary Gaussian processes,” 2015.