Gaussian Process Robust Regression for Noisy Heart Rate Data

Oliver Stegle*, Sebastian V. Fallert, David J. C. MacKay, and Søren Brage

Abstract—**Heart rate data collected during nonlaboratory conditions present several data-modeling challenges. First, the noise in such data is often poorly described by a simple Gaussian; it has outliers and errors come in bursts. Second, in large-scale studies the ECG waveform is usually not recorded in full, so one has to deal with missing information. In this paper, we propose a robust postprocessing model for such applications. Our model to infer the latent heart rate time series consists of two main components: unsupervised clustering followed by Bayesian regression. The clustering component uses auxiliary data to learn the structure of outliers and noise bursts. The subsequent Gaussian process regression model uses the cluster assignments as prior information and incorporates expert knowledge about the physiology of the heart. We apply the method to a wide range of heart rate data and obtain convincing predictions along with uncertainty estimates. In a quantitative comparison with existing postprocessing methodology, our model achieves a significant increase in performance.**

Index Terms—**Gaussian process (GP), heart rate, noise, robust regression.**

I. Introduction

RECORDING of biological data during nonlaboratory conditions can be a challenging task, and the monitoring of heart rate in free-living humans is no exception. The accurate assessment of heart rate depends on the ability to locate consecutive recurring features in cardiac activation cycles, e.g., the QRS complex in the ECG, in order to measure the interbeat intervals (IBIs).

Several authors have considered automated strategies for inferring heart rate and other statistics from recordings of the full ECG waveform [1]–[3] or from the complete sequence of IBI [4]–[6]. These methods rely on the availability of either the complete ECG waveform or the complete IBI time series. However, in most large-scale studies, only a fraction of the full information can be captured, owing to memory and battery limits [7], [8]. These restrictions are rooted in the nature of such studies; for example, when assessing habitual physical activity during free-living conditions, an important determinant for health in affluent societies [9], average heart rate is typically recorded over several days [7], and hence, devices must be small enough to be tolerated by participants for such periods [10].

Such monitors perform some immediate online data processing to obtain an estimate of average heart rate at each sampling point, e.g., every minute. However, processing schemes that are local in time cannot deal sufficiently well with outliers or bursty noise. This motivates the search for a robust postprocessing procedure that can take both short- and long-term correlations into account. Furthermore, newer monitors store additional auxiliary information along with the average heart rate, e.g., the fraction of time the onboard peak detection algorithm could not detect beats within a specified physiological range [8], which can be beneficial for the postprocessing machinery. We propose an inference scheme that aims for robust prediction of the latent heart rate time series using both the noisy average heart rate data and the auxiliary variables. Our full model has two main components: unsupervised clustering of the auxiliary variables followed by Bayesian regression of the heart rate data. The clustering stage is intended to detect outliers and identify noise bursts based on auxiliary variables. The subsequent Gaussian process (GP) regression model uses a soft assignment of points to clusters as prior information about the noise and incorporates beliefs about physiology of the heart. Applying the proposed method to epidemiological heart rate data, we obtain convincing predictions along with uncertainty estimates. A quantitative comparison to existing methodology suggests a significant improvement in performance when using the proposed method. Moreover, all individual components of our approach are found to be relevant, thus justifying the complexity of the full model.

II. Methods And Model

As an illustration of the type of data considered in this investigation, Fig. 1 shows a typical heart rate trace (top panel). From the mean heart rate as reported by the sensor, one can identify typical physiological properties of heart rate such as 24 h periodicity (circadian rhythm). Newer sensors like the one used here can also store additional information extracted from the underlying ECG waveform, e.g., the longest and shortest IBI in each sampling interval and the fraction of time during which the onboard beat detection algorithm returned out-of-range IBI values [8]. Panels 2–4 of Fig. 1 show combinations of these.

The data in Fig. 1 already suggest that the auxiliary data may be helpful in identifying noisy periods. One can by eye identify regions that contain data with moderate levels of noise (e.g., interval $4500 < t < 5500$) as well as regions dominated by heavy distortions (e.g., interval $2000 < t < 2300$). These regions are also found to be characterized by distinctly different

Manuscript received May 16, 2007; revised January 20, 2008. This work was supported by the U.K. Engineering and Physical Sciences Research Council (EPSRC) (OS, SF), the Cambridge Gates Trust (OS), and the Cambridge European Trust (SF). *Asterisk indicates corresponding author.*

*O. Stegle was with the University of Cambridge, Cambridge CB3 0HE, U.K. He is now with the Department of Physics, University of Cambridge, Cambridge CB3 0HE, U.K. (e-mail: os252@cam.ac.uk).

S. V. Fallert is with the Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, U.K. (e-mail: sf287@cam.ac.uk).

D. J. C. MacKay is with the Department of Physics, University of Cambridge, Cambridge CB3 0HE, U.K. (e-mail: mackay@mrao.cam.ac.uk).

S. Brage is with the MRC Epidemiology Unit, Cambridge CB2 0QQ, U.K. (e-mail: sb400@medschl.cam.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBME.2008.923118

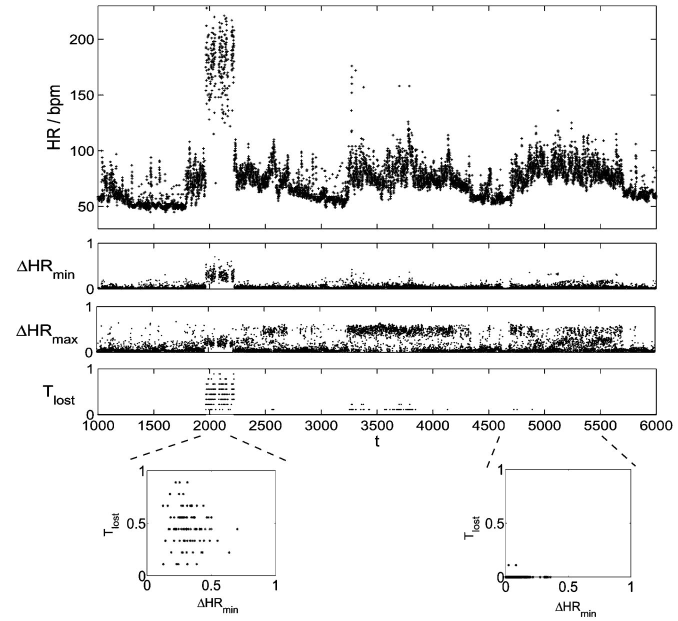


Fig. 1. Typical heart rate time series for one individual over a period of four days. Plots show (from top to bottom) the mean heart rate reported from the sensor, the relative difference between the longest and shortest measured IBI (converted to instantaneous heart rate) and the mean, $\Delta\mathrm{HR}_{\mathrm{min}}$ and $\Delta\mathrm{HR}_{\mathrm{max}}$, and the out-of-range time fraction (T_{lost}). In addition, T_{lost} is plotted against $\Delta\mathrm{HR}_{\mathrm{min}}$ for two different regions of the time series where the left panel presumably shows noisy data while the region corresponding to the right panel is relatively clean.

signatures in the auxiliary information as displayed in panels 2–4 of Fig. 1. This is exemplified by the scatter plots of two of the auxiliary variable combinations ($\Delta\mathrm{HR}_{\mathrm{min}}$ vs. T_{lost}).

This association between the auxiliary variables and the noisiness of the heart rate data motivates the inference model depicted in Fig. 2. The variables $f_1, \ldots, f_n$ represent values of the latent heart-rate function $f(t)$ at time points $t_1, \ldots, t_n$ and the variables $y_1, \ldots, y_n$ correspond to the measured heart rates. The variables $z_1, \ldots, z_n$ describe the associated status of the measurement system. Our model extends common methods for heart rate regression in two ways. First, we use a GP prior [Fig. 2(A)] over latent heart rate functions $f(t)$ that takes physiological properties specific to heart rate into account. Characteristics we model include the different time scales of variation of the beat rate and the daily approximate periodicity of heart rate. Thereby, physiological beliefs can be taken into account. Second, we extract information from the auxiliary variables to identify outliers and noise bursts [Fig. 2(B)]. The main assumption for the clustering module is the existence of latent status variables z_n for each data point that affect both the corresponding auxiliary data $\mathbf{S}_n$ and the noise level. By inferring z_n from the auxiliary variables using Bayesian clustering, the algorithm learns class assignments of data points that can be interpreted as distinct noise classes. The assignments are soft which means that each data point can belong to all noise classes but to varying degrees. A noise class can be thought of as a grouping of points by noise level, for example, “clean,” “noisy,” or “very noisy.” Finally, a heavy-tailed noise model [Fig. 2(C)] links the two components relating experimental observations y_n to the latent heart rate f_n, while making use of the inferred noise class assignments z_n.

A. Gaussian Process Regression

The central component of the proposed heart rate regression model is the GP prior that expresses beliefs about smoothness and lengthscales of heart rate functions. Suppose we want to regress a heart rate dataset $\mathcal{D}_{\mathrm{R}} = \{\mathbf{x}_n, y_n\}_{n=1}^{N}$ where inputs $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^{N}$ correspond to time points and targets $\mathbf{y} = \{y_n\}_{n=1}^{N}$ are measured mean heart rate values. We assume that the observed data have been created by a latent function $f(\mathbf{x})$ mapping from the input space $\mathbf{X}$ to output values $\mathbf{f}$, and that the observed data $\mathbf{y}$ differ from outputs because of observation noise. Following a Bayesian approach for regression, the posterior distribution over these latent functions f is

$$P(f|\mathcal{D}_{\mathrm{R}}) = \frac{1}{Z} P(f) P(\mathcal{D}_{\mathrm{R}}|f). \tag{1}$$

The prior $P(f)$ specifies smoothness assumptions while the likelihood $P(\mathcal{D}_{\mathrm{R}}|f)$ captures how function values and observed data are related (noise model). We choose a GP prior to integrate prior knowledge about typical properties of heart rate into the regression model. GP priors can, for example, embody beliefs about heart rate time series such as approximate 24 h periodicity (circadian rhythm).

The noise (the difference between true and measured heart rate) is not expected to be Gaussian. Rather, we choose a heavy-tailed distribution reflecting the belief that some of the data points are very noisy while others may be measured considerably more precisely. Details on how the clustering assignments are used to adjust this distribution depending on the state of the inferred noise classes will be described later.

Generally, for any independently distributed noise model, the GP-posterior distribution over functions f is

$$P\left(f|\mathcal{D}_{\mathrm{R}}, \theta_{\mathrm{k}}, \theta_{\mathrm{l}}\right) \propto \mathcal{N}(f|0, K\left(\mathbf{X}, \mathbf{X}|\theta_{\mathrm{k}}\right)) \prod_{n=1}^{N} p_{\mathrm{l}}(y_n|f_n, \theta_{\mathrm{l}}) \,. \tag{2}$$

The GP prior has zero mean and covariance structure $K(\mathbf{X}, \mathbf{X}|\theta_{\mathrm{k}})$ derived from the covariance function $k(\mathbf{x}, \mathbf{x}'|\theta_{\mathrm{k}})$, which specifies how function values at two inputs $\mathbf{x}$ and $\mathbf{x}'$ are correlated. The kernel parameters or hyperparameters θ_{k} govern smoothness and lengthscales of the prior process. For any finite representation of the input space of f, the covariance function defines the covariance matrix of the multivariate Gaussian that models the joint distribution of function values at those locations. The likelihood for a single data point $p_{\mathrm{l}}(y_n|f_n, \theta_{\mathrm{l}})$ has parameters θ_{l} that determine the noise level.

1) Covariance Functions and Hyperparameters: A simple and popular choice is the squared exponential covariance function $k_{\mathrm{SE}}(x, x', |A_0, l) = A \exp\{-1/2((x - x')/l))^2\}$, which yields very smooth, infinitely differentiable samples f. The parameter A determines the typical squared amplitude of deviation from the mean and l determines the typical lengthscale on which the function varies. Besides this simple example, there exists a large class of well-studied covariance functions [11], most of them having a similar parametrization by some lengthscale

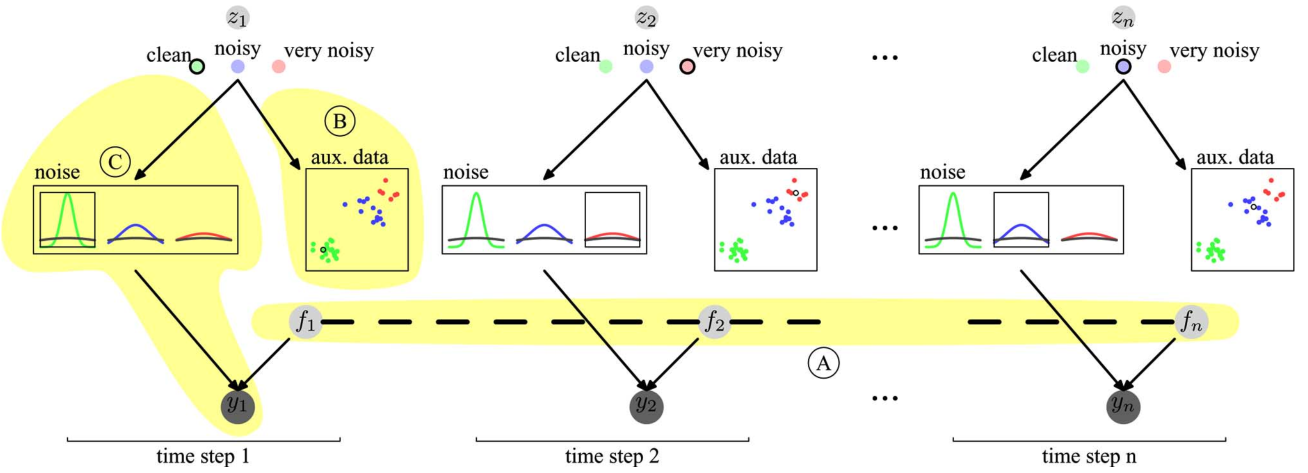


Fig. 2. Heart rate inference model with symbols as defined in the text. (A) Comprising of GP prior. (B) Clustering module for auxiliary variables. (C) Robust noise model. Variables y_n denote observed mean heart rate at time point n, f_n is the true latent heart rate at this time, and z_n the status of this data point, i.e., its noise level.

parameters and an amplitude. Other covariance functions can be created by forming combinations, as used later.

In the following, we will introduce our choice of covariance functions and argue that they promise to be suitable for modeling of heart rate. We study statistical properties of typical heart rate traces (like the one shown in Fig. 1) and identify features with physiological origins.

Examining heart rate data collected over multiple days (Fig. 1), two dominant lengthscales are apparent. Restricting our view to a 1 h window, we observe a very rough process, which shows smoothness only on a short timescale, say on the order of 2–3 min. This smoothness is in line with the physiological understanding that there is inertia in how fast the heart can change its rate in response to both potentiating and inhibiting factors [12]. In addition, viewing an entire week of data reveals periodicity on a much longer timescale, the circadian rhythm, i.e., a distinctive variability between day and nighttimes.

These two effects appear to be additive, and hence, can be modeled by a sum of two covariance functions. We model the short-lengthscale process using a Matern kernel [11]

$$k_{\rm S}(x, x'|A_{\rm S}, \delta_{\rm S}) = A_{\rm S}\ \text{Matern}_{3/2}(x, x', \delta_{\rm S}) \tag{3}$$

which has very rough typical samples. With a lengthscale $\delta_{\rm S} \approx$ 2–3 min and an amplitude $A_{\rm S} \approx 100$, we obtain typical samples, as shown in Fig. 3(a). For the long-term correlations, we choose a covariance function whose typical samples are approximately periodic. The covariance function

$$k_{\rm L}(x, x'|A_{\rm L}, \delta_{\rm L}, P_{\rm L}, V_{\rm L}) = A_{\rm L}\ \exp\left\{-\frac{(x-x')^2}{2\delta_{\rm L}^2}\right\}$$
$$\times \exp\left\{-\frac{\sin^2((2\pi/P_{\rm L})(x-x'))}{2V_{\rm L}^2}\right\} \tag{4}$$

has the required properties; setting the amplitude $A_{\rm L}$ to ≈ 100, the period length $P_{\rm L}$ to 24 h, the scale on which periodicity is expected $\delta_{\rm L}$ to three days, and $V_{\rm L}$ to 1, one obtains typical samples, as depicted in Fig. 3(b). Combining both models using the sum of covariance functions

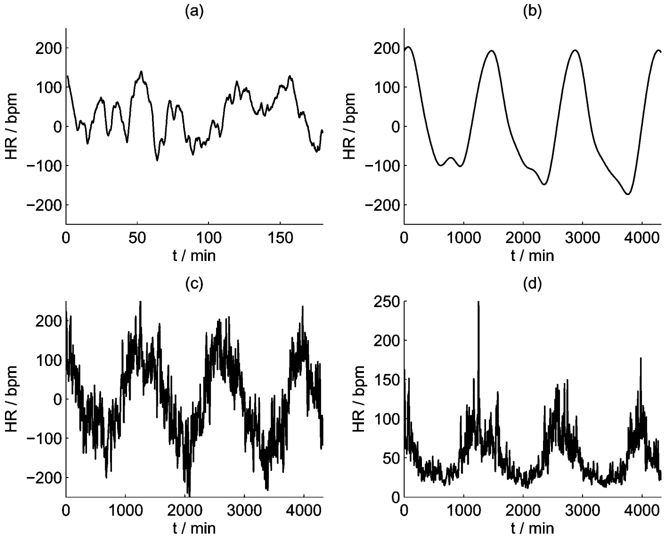


Fig. 3. Samples with hyperparameters as defined in the text. (a) Short-lengthscale kernel $k_{\rm S}$. (b) Long-lengthscale kernel $k_{\rm L}$. (c) Their sum. (d) Log-transformed GP.

$$k(x, x'|\theta_{\rm k}) = k_{\rm S}(x, x'|A_{\rm S}, \delta_{\rm S}) + k_{\rm L}(x, x'|A_{\rm L}, \delta_{\rm L}, P_{\rm L}, V_{\rm L}) \tag{5}$$

we obtain samples that display features similar to real heart rate traces [Fig. 3(c)]. In addition to the design of a suitable covariance function, it is sensible to model the data in a transformed space. While heart rate is defined to be strictly positive, nontransformed GP samples can in principle be negative. Furthermore, samples from a GP are always symmetric around the mean that is unlikely to be true for heart rate. For a typical young adult, a mean daily heart rate would be between 70 and 80 beats/min and the heart rate barely ever drops below 50 beats/min, while occasional peak heart rates of up to 200 beats/min can plausibly occur. In order to address both of these aspects—strict positivity and skewness of samples in the time series—we transform the observed data y using a generalized log-transformation

$$w(y) = \log^{\beta}(y) \tag{6}$$

where the parameter β controls the degree of asymmetry. We assume that the noise model acts in the transformed space, leading to the following generative model for observations $\mathbf{Y}$

$$\mathbf{y}(\mathbf{X}) = \exp[f\,(\mathbf{X}) + \epsilon]^{1/\beta} \tag{7}$$

where $f(\mathbf{x})$ is a GP with the covariance structure introduced earlier and ϵ is the noise process. Fig. 3(d) shows noise-free samples from this model with $\beta = 0.75$.

In summary, we have used prior physiological knowledge and observed statistical properties of real heart rate traces in order to construct a generative model based on a log-transformed GP. We note that the specific parameters for kernel, noise model, and data transformation stated earlier are merely given as an example; in the actual implementation of our scheme, these parameters are learned from the data by optimizing the log marginal likelihood $\log p(\mathbf{y}|\mathbf{X}, \theta_{\mathrm{k}}, \theta_{\mathrm{l}})$ subject to prior distributions on θ_{k} and θ_{l} [11]. The assumptions we made only define the space of alternatives in which we find optimal parameters for each dataset separately in order to accommodate biological individuality.

B. Clustering

As discussed earlier, the clustering module aims to assign the data to noise classes. We define a number of clusters (noise classes) and infer soft assignments, which describe how probable it is that each point belongs to each cluster. We assume that the auxiliary data $\mathcal{D}_{\mathrm{C}} = \{\mathbf{S}_n\}_{n=1}^{N}$ have been drawn independently from a mixture of C Gaussian-distributed clusters with means μ_c, covariance structures Λ_c, and mixing coefficients π_c for each cluster c. The likelihood, i.e., the probability of the data given a specific choice of parameters, can be written as

$$P(\mathcal{D}_{\mathrm{C}}|\{\mu_c, \Lambda_c\}_{c=1}^{C}, \mathbf{Z}) = \prod_{n}^{N} \mathcal{N}(S_n|\mu_{z_n}, \Lambda_{z_n}) \tag{8}$$

where $\mathbf{Z} = \{z_n\}_{n=1}^{N}$ is a set of indices denoting cluster membership for each data point. Introducing priors for means, variances, and mixing coefficients, we can write the posterior distribution as

$$\begin{aligned} &P(\{\mu_c, \Lambda_c\}_{c=1}^{C}, \mathbf{Z}, \pi|\mathcal{D}_{\mathrm{C}}) \propto \\ &\quad P(\mathcal{D}_{\mathrm{C}}|\{\mu_c, \Lambda_c\}_{c=1}^{C}, \mathbf{Z})\, P(\mathbf{Z}|\pi)\, P(\{\mu_c, \Lambda_c\}_{c=1}^{C})\, P(\pi). \end{aligned} \tag{9}$$

Since this exact posterior distribution is intractable, we revert to a variational Bayes approximate solution, as described in Appendix I. The factorized approximate posterior yields a distribution over noise class assignments $Q(\mathbf{Z})$ that means we obtain a soft assignment of each data point to all C clusters, which will be used in the regression stage.

For optimal clustering results, it is important to choose an informative representation of the clustering data, i.e., the auxiliary variables from a heart rate sensor. As an example application of the proposed method, discussed in Section III, we analyzed data recorded with a combined movement and heart rate monitor (Actiheart, Cambridge Neurotechnology, Ltd., Papworth, U.K.) that implements a peak detection algorithm [1]. For long-term recordings (>24 h), the monitor stores a trimmed mean heart rate estimate based on the 16 most recent IBI at each sampling point, say every 30 s. Further, a set of auxiliary variables are stored with each data point. These include the two longest and shortest among these IBI, along with the fraction of time where no detection of IBI within the sensor's range (between 240 and 1992 ms) was possible [8]. These auxiliary variables were combined into quantities that are deemed likely to exhibit characteristic behavior for noisy data points. For such points, the observed IBI distribution is expected to be broader and possibly skewed. Therefore, we convert the four extreme IBI values to instantaneous heart rate ($\mathrm{HR}_{\min 1-2}$ and $\mathrm{HR}_{\max 1-2}$) and include the relative differences between these variables and $\overline{\mathrm{HR}}$ as clustering variables. Also, the out-of-range time fraction (T_{lost}) promises to be a good indicator. The specific combinations of auxiliary variables used for the clustering variables were

$$\begin{aligned}
S(1) &= \left|\frac{(\overline{\mathrm{HR}} - \mathrm{HR}_{\mathrm{min1}})}{(\overline{\mathrm{HR}} + \mathrm{HR}_{\mathrm{min1}})}\right| \\
S(2) &= \left|\frac{(\overline{\mathrm{HR}} - \mathrm{HR}_{\mathrm{min\,2}})}{(\overline{\mathrm{HR}} + \mathrm{HR}_{\mathrm{min\,2}})}\right| \\
S(3) &= \left|\frac{(\overline{\mathrm{HR}} - \mathrm{HR}_{\mathrm{max1}})}{(\overline{\mathrm{HR}} + \mathrm{HR}_{\mathrm{max1}})}\right| \\
S(4) &= \left|\frac{(\overline{\mathrm{HR}} - \mathrm{HR}_{\mathrm{max2}})}{(\overline{\mathrm{HR}} + \mathrm{HR}_{\mathrm{max2}})}\right| \\
S(5) &= \left|\frac{(\mathrm{HR}_{\mathrm{min1}} - \mathrm{HR}_{\mathrm{max1}})}{(\mathrm{HR}_{\mathrm{min1}} + \mathrm{HR}_{\mathrm{max1}})}\right| \\
S(6) &= \left|\frac{(\mathrm{HR}_{\mathrm{min\,2}} - \mathrm{HR}_{\mathrm{max2}})}{(\mathrm{HR}_{\mathrm{min\,2}} + \mathrm{HR}_{\mathrm{max2}})}\right| \\
S(7) &= T_{\mathrm{lost}}.
\end{aligned}$$

C. Noise Model

Having explained how the nature of noise is learned from auxiliary data and how the GP machinery regresses data, we now connect these components using a noise model.

While noise is often modeled as Gaussian for mathematical convenience, the noise encountered in real settings, like the case of heart rate measurement, is rarely Gaussian. Therefore, we consider a heavy-tailed noise model, which acknowledges the existence of outliers and thereby delivers a robust regression method.

In traditional approaches to postprocessing of heart rate data, it is common practice to simply filter out apparent outliers, perhaps supervised by the trained eye of a physiologist, after which all remaining data are treated as noise-free observations [13]. Such an approach will yield results of varying quality depending on the noise characteristics of the sensor system, and whether there exists additional prior knowledge about the phenomenon

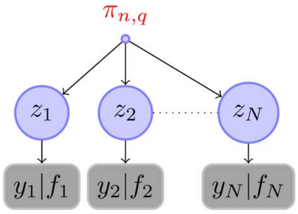


Fig. 4. Likelihood model for robust mixture of Gaussian noise.

that can aid in identifying outliers. However successful such pragmatic approaches may be, they remain an uncontrolled approximation, for which the implicit assumptions are not clear and cannot be formally evaluated. We will return to this point when comparing our results to such approaches in Section III.

The systematic Bayesian framework proposed here takes outliers into account by modeling them explicitly. Consider a simple robust likelihood model, the "two model" by Jaynes and Bretthorst [14]

$$p_{\mathrm{l}}(y_n|f_n,\theta_{\mathrm{l}}) = \pi_0\,\mathcal{N}(y_n|f_n,\sigma_0^2) + (1-\pi_0)\,\mathcal{N}(y_n|f_n,\sigma_1^2) \tag{10}$$

which reflects the belief that a small fraction $(1-\pi_0)$ of the data can be regarded as outliers. If the second mixture component has a variance σ_1^2 that is much larger than σ_0^2, this model can effectively ignore outliers. Generalizing this to the situation where one can identify Q classes of data points with different levels of noise, (10) becomes

$$p_{\mathrm{l}}(y_n|f_n,\theta_{\mathrm{l}}) = \sum_{q=1}^{Q} \pi_{n,q}\,\mathcal{N}(y_n|f_n,\sigma_q^2) \tag{11}$$

which is illustrated in Fig. 4. As explained earlier and illustrated in Fig. 2, our model assumes the existence of latent status variables (noise classes) z_n. These determine both the noise level and the auxiliary variables $\mathbf{S}_n$ for each data point. Under this assumption, the probability for each data point to be assigned to cluster c in the auxiliary variables $Q(z_n = c)$ is used to set corresponding mixing coefficients $\pi_{n,c}$ of the noise model. We further improve the robustness of our model by accounting for outliers that remain unidentified by the clustering. An additional outlier class is introduced to which all data points are assigned with the same small probability regardless of z_n.

D. Implementation of Inference

Having introduced all our assumptions, we now implement inference for this model. Looking back at the posterior distribution over functions f, (2), combined with the robust noise model of the previous section, it takes the form of a mixture of N^K GPs

$$P\left(f|\mathcal{D}_{\mathrm{R}},\theta_{\mathrm{k}},\theta_{\mathrm{l}}\right) \propto \mathcal{N}(f|0, K\left(\mathbf{X},\mathbf{X}|\theta_{\mathrm{k}}\right)) \prod_{n=1}^{N} p_L\left(y_n|f_n,\theta_{\mathrm{l}}\right)$$
$$= \mathcal{N}(f|0, K(\mathbf{X},\mathbf{X}|\theta_{\mathrm{k}})) \prod_{n=1}^{N}\sum_{q=1}^{Q} \pi_{n,q}\,\mathcal{N}\left(y_n|f_n,\sigma_q^2\right). \tag{12}$$

While for a Gaussian noise model there exists a closed-form solution for the posterior mean and variance of a GP (Appendix II), there is no simple solution for the robust mixture noise model. Therefore, we use a deterministic approximate inference technique, expectation propagation (EP) [15], to approximate the posterior distribution (Appendix II). EP yields a GP as approximation to the posterior, from which we can evaluate the predicted mean and variance for test predictions.

As described previously, we link this noise model with the clustering component by setting the mixing coefficients to the inferred soft cluster assignments. In addition to the noise classes inferred by the clustering, we introduce an extra mixing class in the noise model to acknowledge a small probability of outliers that are not detected by the clustering stage. Therefore, the number of classes in the noise model is $Q = C + 1$, and we assign

$$\pi_{n,q} = \begin{cases} \frac{Q(z_n=q)}{1+\pi_{\mathrm{b}}^*}, & \text{for } q = 1,\ldots,C \\ \frac{\pi_{\mathrm{b}}^*}{1+\pi_{\mathrm{b}}^*}, & \text{for } q = Q \end{cases} \tag{13}$$

where π_{b}^* is a parameter that is optimized along with the other hyperparameters. Once the mixing coefficients have been assigned, we learn the GP with the robust noise model and obtain a posterior distribution for the latent heart rate function.

Our current implementation of the full model treats clustering and regression as two separate steps. We first use clustering to infer noise classes and thereby determine the mixing coefficients $\pi_{n,q}$ and then perform robust regression. This approach is similar to the first step of an expectation maximization (EM) approximation to the joint inference model. The GP prior coupling the latent function at various time points f_n makes the model not tree-like, and hence, exact EM inference becomes intractable.

The efficient approximate inference techniques applied to clustering and regression keep the computational cost for our model moderate. On a desktop machine with a 2 GHz CPU it takes approximately 5 min to process a four-day dataset with 6000 time points.

III. Results And Discussion

In what follows, we will present the results of applying our model to a range of datasets and discuss them in relation to previous approaches to the problem. First, we study a free-living dataset containing heart rate data collected over a period of several days in 40 adult men and women of varying age, body composition, and origin[1] (U.K. and Kenya). We discuss results and properties of the algorithm, particularly its robustness. Second, we compare the predictive performance of models of increasing complexity to a standard ("heuristic") baseline model commonly applied in epidemiological studies. Finally, we verify the consistency of our results by investigating a double-sensor dataset.

[1] Ethical approval for data collection was obtained from the Local Research Ethics Committees and all individuals provided informed consent before undergoing any measurement.

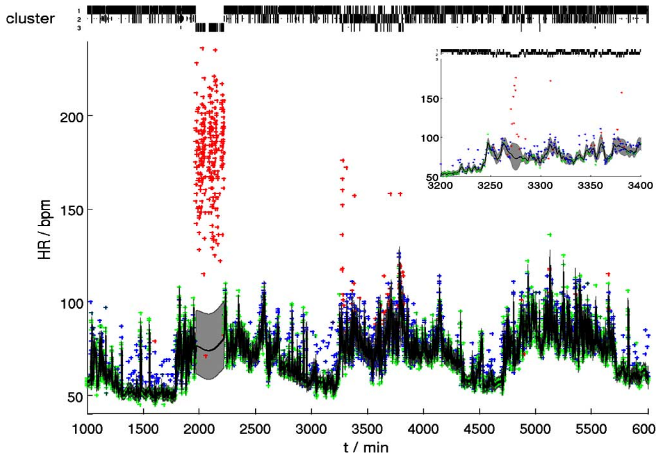


Fig. 5. Mean predicted heart rate from the regression (solid line) with two-sigma error bars (gray-shaded area) for a four-day dataset (+). The Hinton diagram above the plot denotes inferred membership of $C = 3$ clusters where the top row is the cleanest and the last row the noisiest cluster. The size of the black block corresponds to the probability of a data point being in a specific cluster. Cluster membership is also color coded via red, green, and blue (RGB) values where green corresponds to the cleanest and red to the noisiest cluster. Inset shows a zoomed view of the same dataset.

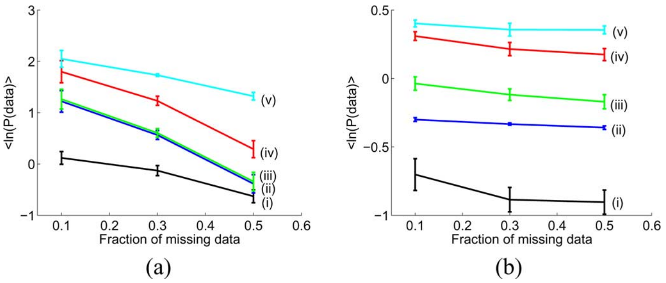


Fig. 6. Predictive performances (mean log probability of filled-in data) of the five models discussed in the text as a function of the fraction of data points removed where (from bottom to top). (i) Baseline. (ii) GP with k_s. (iii) GP with $k_s + k_l$. (iv) As (iii) but with T_{lost} threshold noise model. (v) As (iv) but with robust noise model using clustering (full model). Both removal of single points (a) and removal of 1 h blocks (b) are considered.

A. Free-Living Dataset

Application of the clustering followed by GP regression for one individual's heart rate over a four-day period yielded the results shown in Fig. 5. The most probable number of clusters was found to be $C = 3$. The soft assignments of points to clusters is visualized via Hinton diagrams.

The noise block starting from $t \approx 2000$ is a typical example of the type of noise we set out to model. We note that, encouragingly and in agreement with intuition, all data points of this block have been assigned to the very noisy cluster. By closer inspection of the clustering variables, the cause for these data being classified as structured noise is the existence of larger relative differences between $\text{HR}_{\min 1-2}$ and $\text{HR}_{\max 1-2}$ and $\overline{\text{HR}}$, as well as high values for the out-of-range time fraction (T_{lost}). Subsequently, the trimmed average heart rate values and clustering assignments were used to perform robust GP regression. Hyperparameters, including the noise levels for the individual clusters, were determined from the data by maximizing the log marginal likelihood with respect to all parameters as described in Section II-A. The inferred noise levels for each noise cluster determine the order used in the Hinton diagrams. All plots show GP mean predictions with error bars of ± 2 standard deviations.

Reassuringly, the noise block in Fig. 5 previously identified as very noisy is indeed characterized by large predicted uncertainty. Uncertainty is thus found to naturally increase in areas where data have been identified as noisy by the clustering stage while it is low-to-moderate in regions of low noise such as the zoomed view displayed in the inset of Fig. 5.

The empirical results mentioned earlier were obtained using the full model. We assess the predictive performance of models of increasing complexity in relation to a baseline model, a standard postprocessing method commonly used in epidemiological studies [13]. In this predictive test, a certain fraction of the data are removed from the training set. Each model's predictions for the missing values are then compared with their known observed values. For a single point, we measure the predictive performance by taking the logarithm of the probability assigned by the model to the experimental value. To summarize a model's performance, we average this log probability over all filled-in points. Fig. 6(a) shows the predictive performances of five models as a function of the fraction of data points removed. The baseline model is a heart rate interpolation method based on heuristics [13]. In this approach, physiologically implausible data points (heart rate values outside the range $[30, 200]$) are deleted. Subsequently, missing values are imputed by local regression if the gap is <5 min, otherwise they are replaced by the mean heart rate value of the dataset. This baseline model makes no use of the auxiliary data. Since this approach does not per se yield a predicted uncertainty, error bars are estimated using cross validation. We compare this baseline approach [model (i)] to increasingly complex versions of our full model. The simplest variant of our model uses GP regression with a standard Gaussian noise model (no outliers) and, like the baseline model, ignores the auxiliary data [model (ii)]. This GP model uses the short-lengthscale covariance function described before. This can be readily extended by adding the long-lengthscale covariance function [model (iii)]. Subsequently, we improve the noise model by classifying data points as "clean" or "noisy" based on a threshold rule (>0 %) for the T_{lost} auxiliary variable [model (iv)]. Finally, by adding clustering of all auxiliary variables, we arrive at our full model [model (v)]. Fig. 6(a) shows that the baseline model performs worst, and the more sophisticated the model, the better the performance is. As the fraction of removed data points increases, the performance of all models deteriorates.

In order to investigate the effect of longer periods of removed data points, the predictive performance test was repeated with a removal of 1 h blocks [Fig. 6(b)]. In this case, while the same trends as before are observed, the performance of all models is found to be lower than in Fig. 6(a). Further, the change of performance with increasing fraction of removed data is less significant than in the previous case.

While in the case of single-point removal, the long-lengthscale covariance function does not add any significant gain, it improves performance when removing blocks of data.

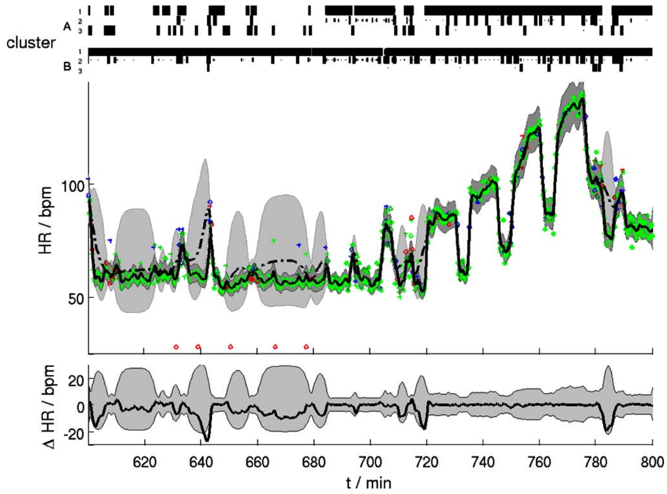


Fig. 7. Double-sensor dataset from calorimetry experiment over 24 h and mean predicted heart rate for upper-placed sensor (◇, dot-dashed line) and lower-placed sensor (+, solid line) with error bars of ±2 standard deviations (light and dark gray areas, respectively). Lower graph shows the difference between the predicted heart rate means from both sensors with ±2 standard deviations. Hinton diagrams denote cluster memberships as before [first upper (A), then lower-placed sensor (B)].

This indicates that the modeling of long-range correlations is essential for dealing with noise in bursts as encountered during measurement in nonlaboratory conditions. Additional comparison to annotated datasets is made available as supplementary material [16].

B. Double Sensor Dataset

In order to examine the consistency of the proposed method, we used a double dataset (27 individuals). This dataset was recorded with two separate sensors on the same individual, one placed at the level of the third intercostal space (upper sensor) and the other placed below the apex of the sternum (lower sensor) [17]. Information from each sensor was processed separately and regressed using two independent GPs. As shown in Fig. 7, the two predictions overlap within uncertainty estimates, thereby giving evidence for the consistency of our model. More formally, a χ^2-test confirms this and indicates that, on average, our model makes conservative predictions.

On average, the uncertainties for the upper-placed sensor were higher than those of the lower-placed sensor (± 7.4 beats/min and ± 4.3 beats/min, respectively). This reproduces an earlier empirical finding [17] and further demonstrates the merit of quantifying uncertainty for heart rate data.

IV. Conclusion

We proposed a novel postprocessing method for heart rate data from measurement devices collecting reduced ECG information. Our method combines noise pattern identification based on clustering with a physiologically motivated GP prior for regression. The proposed model performs well on typical heart rate data collected in large-scale epidemiological studies. Further, a quantitative comparison to existing methodology shows that a significant improvement is achieved.

The source code that was used to obtain our results is available on request.

Appendix I

Variational Mixture of Gaussian Clustering

Applying variational methods, we choose a separable distribution $Q(\{\mu_c, \Lambda_c\}, \mathbf{Z}, \pi|\mathcal{D}_\mathrm{C}) = Q_1(\{\mu_c, \Lambda_c\})Q_2(\mathbf{Z})Q_3(\pi)$, to approximate the exact posterior of the clustering problem (9). Parameters of the approximation are optimized by minimizing the Kullback–Leibler (KL) divergence [19] between this distribution and the exact posterior

$$\mathrm{KL}\left[Q_1(\{\mu_c, \Lambda_c\}_{c=1}^{C})Q_2(\mathbf{Z})Q_3(\pi)||P(\{\mu_c, \Lambda_c\}_{c=1}^{C}, \mathbf{Z}, \pi|\mathcal{D}_\mathrm{C})\right].$$

After choosing conjugate priors for the clustering parameters $P(\{\mu_c, \Lambda_c\}_{c=1}^{C})$ and $P(\pi)$, it is possible to derive a set of equations that allow an iterative update of the parameters of the approximate posterior distribution [18].

We note that the approximate posterior distribution for cluster assignments $Q(\mathbf{Z})$ assigns each data point a soft membership of the C clusters that is used in the noise model (Section II-C). The most probable number of clusters is determined by optimizing the evidence for a given dataset $Z = P(\mathcal{D}_\mathrm{C}) = \int_\theta P(\mathcal{D}_\mathrm{C}|\theta)P(\theta)$ with respect to C.

Appendix II

Implementation Of Inference

As an intermediate step to obtain an approximate solution to the posterior (12), let us first consider GP inference for the Gaussian noise model. In this special case, the joint distribution over training (observed) data and predictions $\mathbf{f}^*$ at test inputs $\mathbf{X}^* = \{x_n^*\}_{n=1}^{N^*}$ can be expressed in closed form

$$P\left([\mathbf{y}, \mathbf{f}^*]|\mathbf{X}, \mathbf{X}^*, \theta_\mathrm{k}, \theta_\mathrm{l}\right) \propto \mathcal{N}\left([\mathbf{y}, \mathbf{f}^*]|\mathbf{0}, \begin{bmatrix} K_{\mathbf{X}\ \mathbf{X}} + \sigma^2 I & K_{\mathbf{X}\ \mathbf{X}^*} \\ K_{\mathbf{X}^*\mathbf{X}} & K_{\mathbf{X}^*\mathbf{X}^*} \end{bmatrix}\right) \qquad (14)$$

where, for brevity, the kernel matrices $K(\mathbf{X}, \mathbf{X}'|\theta_\mathrm{k})$ are abbreviated as $K_{\mathbf{X}\mathbf{X}'}$ omitting the dependency on hyperparameters θ_k.

Since the predictive distribution $P(\mathbf{f}^*|\mathcal{D}_\mathrm{R}, \mathbf{X}^*)$ is proportional to this joint distribution, we can condition on the observed data, complete the square, and obtain the predictive mean and covariance for the test inputs

$$P(\mathbf{f}^*|\mathcal{D}_\mathrm{R}, \mathbf{X}^*) \propto \mathcal{N}\left(\overline{\mathbf{f}^*}, \mathrm{cov}(\mathbf{f}^*)\right) \qquad (15)$$

where

$$\overline{\mathbf{f}^*} = K_{\mathbf{X}^*\mathbf{X}}\left[K_{\mathbf{X}\mathbf{X}} + \sigma^2 I\right]^{-1}\mathbf{y}$$
$$\mathrm{cov}(\mathbf{f}^*) = K_{\mathbf{X}^*\mathbf{X}^*} - K_{\mathbf{X}^*\mathbf{X}}\left[K_{\mathbf{X}\mathbf{X}} + \sigma^2 I\right]^{-1}K_{\mathbf{X}\mathbf{X}^*} \qquad (16)$$

The implementation of GP inference using the mixture of Gaussian noise model (11) is more challenging than for a Gaussian model. The posterior distribution (2) is a mixture of N^K GPs

$$\begin{aligned} P\left(f|\mathcal{D}_{\rm R}, \theta_{\rm k}, \theta_{\rm l}\right) &\propto \mathcal{N}(f|0, K_{\mathbf{XX}}) \prod_{n=1}^{N} p_L(y_n|f_n, \theta_{\rm l}) \\ &= \mathcal{N}(f|0, K_{\mathbf{XX}}) \prod_{n=1}^{N} \sum_{q=1}^{Q} \pi_{n,q}\, \mathcal{N}\big(y_n|f_n, \sigma_q^2\big). \end{aligned} \tag{17}$$

For efficient inference, we use EP [15] to obtain a deterministic approximation to the exact posterior

$$Q(f|\mathcal{D}_{\rm R}, \theta_{\rm k}, \theta_{\rm l}) \propto \mathcal{N}(f|0, K_{\mathbf{XX}}) \prod_{n=1}^{N} t_n(f_n|C_n, \mu_n, \nu_n) \tag{18}$$

where $t_n(f_n|C_n, \mu_n, \nu_n)$ denote approximate factors. If we choose nonnormalized Gaussians for the approximate factors

$$t_n(f_n|C_n, \mu_n, \nu_n) = C_n \exp\left(-\frac{1}{2\nu_n^2}(f_n - \mu_n)^2\right) \tag{19}$$

we obtain a GP for the approximate distribution again. The idea of EP is to iteratively update one approximate factor leaving all other factors fixed. This is achieved by minimizing the KL divergence, a distance measure for distributions [19]. The update for one approximate factor i is performed by minimizing

$$\begin{aligned} \mathrm{KL}\Bigg[&\mathcal{N}(f|0, K_{\mathbf{XX}}) \prod_{n\neq i} t_n(f_n|C_n, \mu_n, \nu_n)\, \overbrace{p_L(y_i|f_i, \theta_{\rm l})}^{\text{exact factor}}\, || \\ &\mathcal{N}(f|0, K_{\mathbf{XX}}) \prod_{n\neq i} t_n(f_n|C_n, \mu_n, \nu_n)\, \underbrace{t_i\left(f_i|C_i, \mu_i, \nu_i\right)}_{\text{approximation}}\Bigg] \end{aligned} \tag{20}$$

with respect to the ith factor's parameters μ_i, ν_i, and C_i. This is done by matching the moments between the two arguments of the KL divergence that can then be translated back into an update for factor parameters. It is convenient to work in the natural parameter representation of the distributions where multiplication and division of factors are equivalent to addition and subtraction of the parameters.

There is no convergence guarantee for EP but in practice it is found to converge for the likelihood model we consider [20]. The fact that the mixture of Gaussian likelihood is not log-concave represents a problem as it may cause invalid EP updates, leading to a covariance matrix that is not positive definite. We avoid this problem by damping the updates [20], [21].

By capturing the zeroth moment of the exact distribution with the explicit normalization constant C_n, we obtain an approximation to the log marginal likelihood

$$\begin{aligned} \log P(\mathcal{D}_{\rm R}|\theta_{\rm k}, \theta_{\rm l}) &\approx \ln \int df\, \mathcal{N}(f|0, K_{\mathbf{XX}}) \prod_{n=1}^{N} t_n(f_n) \\ &= \frac{1}{2}\sum_{n=1}^{N}\left(\ln \nu_n^2 + \ln C_n\right) - \frac{1}{2}\ln|K_{\mathbf{XX}} + \Sigma| \\ &\quad - \frac{1}{2}\mathbf{y}^T\left(K_{\mathbf{XX}} + \Sigma\right)\mathbf{y} \end{aligned} \tag{21}$$

where $\Sigma = \mathrm{diag}(\{\nu_n\}_{n=1}^N)$. This log marginal likelihood approximation enables us to optimize the kernel and likelihood parameters $\theta_{\rm k}$ and $\theta_{\rm l}$.

After EP converged, we obtain a GP as approximate posterior distribution again, and hence, can evaluate a predicted mean and variance as for the Gaussian noise model (15).

ACKNOWLEDGMENT

The authors would like to thank D. L. Christensen and K. B. Johnsen for providing additional data. N. Wareham and C. Neugebauer are acknowledged for valuable discussions in the early stages of this research.

REFERENCES

[1] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Trans. Biomed. Eng.*, vol. BME-32, no. 3, pp. 230–236, Mar. 1985.

[2] G. M. Friesen, T. C. Jannett, M. A. Jadallah, S. L. Yates, S. R. Quint, and H. T. Nagle, "A comparison of the noise sensitivity of nine QRS detection algorithms," *IEEE Trans. Biomed. Eng.*, vol. 37, no. 1, pp. 85–98, Jan. 1990.

[3] I. I. Christov, "Real time electrocardiogram QRS detection using combined adaptive threshold," *Biomed. Eng. Online*, vol. 3, no. 28, pp. 3–28, 2004.

[4] G. G. Berntson, K. S. Quigley, J. F. Jang, and S. T. Boysen, "An approach to artifact identification: Application to heart period data,," *Psychophysiology*, vol. 27, no. 5, pp. 586–598, 1990.

[5] A. Machado, E. Migliaro, P. Contreras, and F. Coro, "Automatic filtering of R–R intervals for heart rate variability analysis," *Ann. Noninvasive Electrocardiol.*, vol. 5, no. 3, pp. 255–261, 2000.

[6] G. G. Berntson and J. R. Stowell, "ECG artifacts and heart period variability: Don't miss a beat!,," *Psychophysiology*, vol. 35, no. 1, pp. 127–132, 2001.

[7] N. J. Wareham, S. J. Hennings, A. M. Prentice, and N. E. Day, "Feasibility of heart-rate monitoring to estimate total level and pattern of energy expenditure in a population-based epidemiological study: The Ely young cohort feasibility study 1994–1995," *Br. J. Nutrition*, vol. 78, pp. 889–900, 1997.

[8] S. Brage, N. Brage, P. Franks, U. Ekelund, and N. Wareham, "Reliability and validity of the combined heart rate and movement sensor Actiheart," *Eur. J. Clin. Nutrition*, vol. 59, pp. 561–570, 2005.

[9] U. Ekelund, P. W. Franks, S. Sharp, S. Brage, and N. J. Wareham, "Increase in physical activity energy expenditure is associated with reduced metabolic risk independent of change in fatness and fitness," *Diab. Care*, vol. 30, no. 8, pp. 2101–2106, Aug. 2007.

[10] G. Wickström and T. Bendix, "The "hawthorne effect"—What did the original hawthorne studies actually show?," *Scand. J. Work Environ. Health*, vol. 26, no. 4, pp. 363–367, Aug. 2000.

[11] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press, 2006.

[12] P. Åstrand, *Textbook of Work Physiology: Physiological Bases of Exercise*. Champaign, IL: Human Kinetics, 2003.

[13] L. Davidson, G. McNeill, P. Haggarty, J. Smith, and M. Franklin, "Free-living energy expenditure of adult men assessed by continuous heart-rate monitoring and doubly-labelled water," *Br. J. Nutrition*, vol. 78, pp. 695–708, 1997.

[14] E. Jaynes and G. Bretthorst, *Probability Theory: The Logic of Science*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[15] T. Minka, "Divergence measures and message passing," Microsoft Research, Cambridge, U.K.Tech. Rep., 2005.

[16] O. Stegle, S. Fallert, D. Mackay, and S. Brage. Robust Gaussian process regression on annotated datasets, Tech. Rep. [Online]. Available: www.inference.phy.cam.ac.uk/os252/papers/IEEEhrSuppl.pdf.

[17] S. Brage, N. Brage, U. Ekelund, J. Luan, P. Franks, K. Froberg, and N. Wareham, "Effect of combined movement and heart rate monitor placement on physical activity estimates during treadmill locomotion and free-living," *Eur. J. Appl. Physiol.*, vol. 96, no. 5, pp. 517–524, 2006.

[18] C. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006.

[19] S. Kullback and R. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, no. 1, pp. 79–86, 1951.

[20] M. Kuss, T. Pfingsten, L. Csato, and C. Rasmussen, "Approximate inference for robust Gaussian process regression," Max Planck Inst. Biological Cybern., Tubingen, GermanyTech. Rep. 136, 2005.

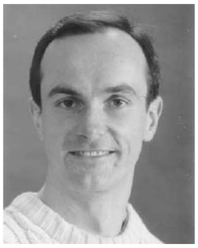
[21] M. Seeger. (2005). Expectation propagation for exponential families, Univ. California Berkeley, Berkeley, CA, Tech. Rep. [Online]. Available: www.kyb.tuebingen.mpg.de/bs/people/seeger.

Oliver Stegle was born in Germany in 1981. He received the Graduate degree in theoretical physics in 2005 from the University of Cambridge, Cambridge, U.K., where he is currently working toward the Ph.D. degree in physics.

His current research interests include Bayesian machine learning applied to bioinformatics and biomedical data modeling.

Sebastian V. Fallert was born in Germany in 1980. He received the B.A. degree in natural sciences (physics) in 2005 and the M.Sci. degree in physics from the University of Cambridge, Cambridge, U.K., where he is currently working toward the Ph.D. degree in physics.

His current research interests include epidemic spreading dynamics from a statistical physics perspective.



David J. C. MacKay was born in U.K. in 1967. He received the B.A. degree in natural sciences (physics) from Trinity College, Cambridge, U.K., in 1988, and the Ph.D. degree in computation and neural systems from California Institute of Technology (Caltech), Pasadena, in 1992.

He was at Darwin College, Cambridge, as a Royal Society Smithson Research Fellow. Since 1995, he has been with the Department of Physics, University of Cambridge, Cambridge, where he was earlier a Lecturer and is currently a Professor. He is the author or coauthor of several papers published in journals and conferences, and a textbook *Information Theory, Inference and Learning Algorithms* (American Statistical Association, 2006).

Søren Brage was born in Denmark in 1973. He received the M.Sc. degree in exercise science and health from the University of Southern Denmark, Odense, Denmark, in 2001, and the M.Phil. degree and the Ph.D. degree in epidemiology from the University of Cambridge, Cambridge, U.K., in 2002 and 2006, respectively.

His current research interests include assessment of physical activity and fitness in populations, with a strong focus on objective biosensing, for studying the relationship between activity and fitness with metabolic diseases in culturally and geographically diverse settings.