Multi-Task Gaussian Processes for Multivariate Physiological Time-Series Analysis

Robert Dürichen, Marco A.F. Pimentel, Lei Clifton, Achim Schweikard, and David A. Clifton

Abstract—Gaussian process (GP) models are a flexible means of performing non-parametric Bayesian regression. However, GP models in healthcare are often only used to model a single univariate output time-series, denoted as single-task GPs (STGP). Due to an increasing prevalence of sensors in healthcare settings, there is an urgent need for robust multivariate time-series tools. Here, we propose a new method using multi-task GPs (MTGPs) which can model multiple correlated multivariate physiological time-series simultaneously. The flexible MTGP framework can learn the correlation within multiple signals even though they might be sampled at different frequencies and have training sets available for different intervals. Furthermore, prior knowledge of any relationship between the time-series such as delays and temporal behaviour can be easily integrated. A novel normalisation is proposed to allow interpretation of the various hyperparameters used in the MTGP. We investigate MTGPs for physiological monitoring with synthetic datasets and two real-world problems from the field of patient monitoring and radiotherapy. The results are compared to standard Gaussian processes and other existing methods in the respective biomedical application areas. In both cases, we show that our proposed framework learned the correlation between physiological time-series efficiently, outperforming the existing state-of-the-art.

Index Terms—Gaussian processes, multivariate data analysis, correlation analysis.

I. INTRODUCTION

Gaussian processes (GPs) are a Bayesian modelling technique that have been widely used for various machine learning tasks, such as dimensionality reduction [1], non-linear classification, and regression [2], [3]. It is a nonparametric method, informally suggesting that the number of parameters in the GP can grow with the number of observed data. Compared to other related techniques, such as, for example, support vector regression (SVR), GP models have the advantage that prior knowledge of the functional behaviour (e.g., periodicity or smoothness) may be easily expressed. The Bayesian nature of its formulation also means that inference is performed in a probabilistic framework, allowing us to reason in the presence of noise, incompleteness, and artefacts, as are typical in realistic healthcare settings.

In the field of biomedical engineering, GPs have been mainly used for modelling physiological time-series. Clifton et al. used GPs to design a vital-sign “early warning system” [4]. Stegle et al. [5] used GPs for a robust regression of noisy heart rate data and Dürichen et al. [6] showed that relevance vector machines can be used efficiently to compensate respiratory motion in robotic radiotherapy. In [7], physiologically-relevant parameters such as the respiratory rate have been estimated directly from the posterior values of the hyperparameters of a GP model that was constructed using mobile sensor data. Recently, GP regression has also been used for the ranking of gene expressions [8].

Most commonly, GPs are used to predict a single output (“task”) based on one or more input time-series. We refer to this model as a single-task GP. Within this work, we explore the potential of an approach using multi-task GP models for physiological time-series analysis. In contrast to most other multi-output models, the aim of a MTGP is to learn the correlation between and within the tasks concurrently. Assuming multiple time-series are given, the intention is to improve the overall modelling accuracy by using a single MTGP model instead of multiple STGPs. MTGPs have been investigated for applications such as the analysis of compiler performance [9], robotic inverse dynamics [10], financial time-series [11], environmental sensor networks [12] and classification of ECG signals [13]. Earlier work has been published within geostatistics in [14], [15], [16]. We note in passing that a multi-task GP is sometimes called a “multi-output” GP in the literature.

One appealing advantage of this framework is that MTGPs can incorporate unevenly-sampled (also discrete) time-series in a single model; no further down-sampling or interpolation is required. Additionally, prior knowledge between the time-series can be easily included such as time shifts or assumptions about similar temporal. This makes the method useful for various applications in healthcare settings, such as compensation of missing data, estimation of time shifts, correlation analysis, and prediction tasks, which will be investigated here on synthetic and real-world datasets. Furthermore, we present a new means of transforming the MTGP correlation hyperparameters between time-series. This increases the interpretability of the coefficients and enables a comparison to other metrics such as Pearson’s correlation coefficient, as is typically required for clinical benchmarking.

After introducing the STGP in section II-A, we derive the MTGP model in section II-B. The following subsections focus on the transformation of the MTGP correlation hyperparameters and on extension of the model (section II-C-E). We subsequently evaluate our approach using several synthetic and real datasets of physiological time-series (section III). In section IV, we discuss further applications of MTGPs within physiological...
data analysis, such as template-matching. To stimulate further research in biomedical applications, we have implemented an open-source MTGP toolbox, which is available on-line.1

II. METHODS

We assume that some training data \( T = \{ \mathbf{x}, \mathbf{y} \} \) are given, where \( \mathbf{x} = \{ x_i \mid i = 1, \ldots, n \} \) is the index set (e.g., the time at which training data are observed) and where \( \mathbf{y} = \{ y_i \mid i = 1, \ldots, n \} \) are the corresponding values of the observed training data with \( x_i, y_i \in \mathbb{R} \), if not stated differently. The aim is to learn a regression model \( y = f(x) + \epsilon \), where \( f(x) \) represents a latent function and where \( \epsilon \sim \mathcal{N}(0, \sigma^2) \) is a noise term. If some test indices are given \( \mathbf{x}^* = \{ x_i^* \mid i = 1, \ldots, p \} \) (e.g., times at which we wish to estimate the value of our time-series model), predictions for unknown test observations \( \mathbf{y}^* = \{ y_i^* \mid i = 1, \ldots, p \} \) can be made. The predicted test observations are denoted \( \hat{\mathbf{y}}^* = \{ \hat{y}_i^* \mid i = 1, \ldots, p \} \). Throughout this paper, we evaluate the accuracy of predictions with respect to a number of metrics. We will use the root mean square error (RMSE), defined as being:

\[
\text{RMSE}(\hat{\mathbf{y}}^*, \mathbf{y}^*) = \left( \frac{1}{p} \sum_{i=1}^{p} (\hat{y}_i^* - y_i^*)^2 \right)^{1/2} = \sqrt{\text{MSE}(\hat{\mathbf{y}}^*, \mathbf{y}^*)}. \tag{1}
\]

As the RMSE is sensitive to the scaling of the observations, we evaluate also the normalised mean square error (NMSE):

\[
\text{NMSE}(\hat{\mathbf{y}}^*, \mathbf{y}^*) = \frac{\sum_{i=1}^{p} (\hat{y}_i^* - y_i^*)^2}{\text{var}(\mathbf{y})}, \tag{2}
\]

where the mean squared error \( \text{MSE}(\hat{\mathbf{y}}^*, \mathbf{y}^*) \) is normalised by the variance of \( \mathbf{y} \). As is common with Bayesian regression models, a predicted observation may be taken as being the mean of the posterior probability distribution. Therefore, we can also evaluate the probability of \( \mathbf{y} \) under the model. Similar to [3], we define the mean standardised log (MSLL) as:

\[
\text{MSLL}(\hat{\mathbf{y}}^*, \mathbf{y}^*) = \frac{1}{p} \sum_{i=1}^{p} \left( - \log p(\hat{y}_i^* | f, x_i^*) + \log p(\hat{y}_i^* | \mathbf{m}(\mathbf{y}), \text{var}(\mathbf{y}), x_i^*) \right), \tag{3}
\]

where the first term represents the log likelihood of \( \hat{y}_i^* \) given our latent function \( f \) and the test index \( x_i^* \). This likelihood is normalised by the second term, which is the log likelihood of \( \hat{y}_i^* \) under a trivial model which predicts using a Gaussian with the mean \( m(\mathbf{y}) \) and variance \( \text{var}(\mathbf{y}) \) of the training labels. Consequently, the MSLL will be negative for complex models and close to zero for simple methods.

A. Single-Task Gaussian Process Models

This section provides a brief introduction to STGP models; a more detailed description can be found in [3]. Gaussian process models assume that the function \( f(x) \) can be interpreted as being a probability distribution over functions

\[
y = f(x) \sim \mathcal{GP}(m(x), k(x, x')), \tag{4}
\]

where \( m(x) \) is the mean function of the process and \( k(x, x') \) is a covariance function which describes the coupling between the \( y \) values, determined according to the distance of the \( x \) values. By modifying the covariance function we can encode our prior knowledge concerning the functional behaviour we wish to model. As shown in [3], there exists a large class of covariance functions which could be used. Two frequently-used examples are the squared-exponential (SE) and periodic (PER) covariance functions:

\[
k_{SE}(r) = \theta_A^2 \exp \left\{ -\frac{r^2}{2\theta_L^2} \right\}, \tag{5}
\]

\[
k_{PER}(r) = \theta_A^2 \exp \left\{ -\frac{\sin^2[(2\pi/\theta_P)r]}{2} \right\}, \tag{6}
\]

where \( \theta_A, \theta_L, \) and \( \theta_P \) are hyperparameters modelling the \( y \)-scaling, \( x \)-scaling (or time-scale if the data are time-series), and period of the covariance functions, respectively, and where \( r = \| x - x' \|_2 \) denotes the Euclidean distance between two indices. The covariance for a vector \( \mathbf{x} \in \mathbb{R}^n \) results in a covariance matrix \( \mathbf{K}(\mathbf{x}, \mathbf{x}) \) of size \( n \times n \), where the covariance function \( k(x_i, x_j) \) gives element \( K_{ij} \). In general, covariance functions have to fulfil Mercer’s theorem, meaning that \( \mathbf{K}(\mathbf{x}, \mathbf{x}) \) has to be symmetric and positive semidefinite and therefore that \( k(\cdot, \cdot) \) is a kernel. Complex covariance functions can be constructed by affine transformations of basic covariance functions. One frequently used example is the quasi-periodic (QP) covariance function, which is a product of (5) and (6)

\[
k_{QP}(r) = \theta_A^2 \exp \left\{ -\frac{r^2}{2\theta_L^2} \right\} \times \exp \left\{ -\frac{\sin^2[(2\pi/\theta_P)r]}{2} \right\}. \tag{7}
\]

Given a training set \( T \), predictions can be made at the test indices \( \mathbf{x}^* \) by computing the conditional distribution \( p(\mathbf{y}^* | \mathbf{x}^*, \mathbf{x}, \mathbf{y}) \) which will be a Gaussian distribution

\[
p(\mathbf{y}^* | \mathbf{x}^*, \mathbf{x}, \mathbf{y}) \sim \mathcal{N}(m(\mathbf{y}^*), \text{var}(\mathbf{y}^*)) \tag{8}
\]

where a mean \( m(\mathbf{y}^*) \) and variance \( \text{var}(\mathbf{y}^*) \). Without loss of generality, the mean function \( m(\mathbf{y}) \) is commonly assumed to be zero. Under this assumption \( m(\mathbf{y}^*) \) and \( \text{var}(\mathbf{y}^*) \) are given by

\[
m(\mathbf{y}^*) = \mathbf{K}(\mathbf{x}, \mathbf{x}^*)^\top \mathbf{K}(\mathbf{x}, \mathbf{x})^{-1} \mathbf{y} \tag{9}
\]

\[
\text{var}(\mathbf{y}^*) = \mathbf{K}(\mathbf{x}, \mathbf{x}^*) - \mathbf{K}(\mathbf{x}, \mathbf{x}^*)^\top \mathbf{K}(\mathbf{x}, \mathbf{x})^{-1} \mathbf{K}(\mathbf{x}, \mathbf{x}^*) \tag{10}
\]

The values of the hyperparameters \( \theta \) may be optimised by, for example, minimising the negative log marginal likelihood (NLML) which is defined as

\[
\text{NLML} = -\log p(\mathbf{y} | \mathbf{x}, \theta) = \frac{1}{2} \log |\mathbf{K}| + \frac{1}{2} \mathbf{y}^\top \mathbf{K}^{-1} \mathbf{y} + \frac{n}{2} \log(2\pi) \tag{11}
\]

Interpreting the NLML as a cost function reveals that the first term penalises model complexity and the second term penalises low data likelihood. Bias-variance trade off is therefore performed by minimising the NLML, which is commonly achieved using gradient descent.

1http://www.robots.ox.ac.uk/~davide
A limitation of the MTGP is that the computational cost for
and

is also known as the
for probability
and
indices
use task-specific numbers of training data.

that
temporal covariance functions within a task, respectively. Note
where
K
is the Kronecker product,

and indices

are vector sets containing hyperparameters for
Kc
and
Kt,
respectively. This leads to a matrix of size
m \times m
for
KMTGP,
as
Kc
has a size of
m \times m
and
Kt
of
n \times n.
We refer to
Kc
as the correlation matrix. Within geostatistics,
this approach is also known as the intrinsic correlation model
[16]. Note the simplifying assumption
n^j = n
for
j = 1, ..., m,
may be relaxed so that the model can be easily extended to
use task-specific numbers of training data.

As with STGPs, the hyperparameters for a MTGP may be
optimised by minimising the NLML, and predictions for test
indices \{x^*, l^*\} can be made by computing the conditional probability
p(y^* | x^*, l, x, l, y).
This method has several useful properties:
- we may have task-specific training indices
n^j
(i.e., training data may be observed at task-specific times);
- automatic learning of the correlation within tasks occurs
by fitting the covariance function in (13); and
- the framework assumes that the tasks have similar
temporal characteristics and hyperparameters
\theta_i.

A limitation of the MTGP is that the computational cost for
evaluating MTGPs is \mathcal{O}(m^3 m^3) compared with
m \times \mathcal{O}(n^3)
for STGPs. Additionally, the number of hyperparameters can
increase rapidly for an increasing number of tasks which
can lead to a multi-modal parameter space with no overall
optimum.

B. Multi-Task Gaussian Process Models

The extension to MTGP models is motivated by the problem
of modelling
m
tasks simultaneously (e.g., multiple physiological
time-series) where each model uses the same index set
x
(e.g., the time). A naive approach is to train an STGP model
independently for each task, as illustrated in Fig.1(a).

We assume that
X = \{x_i^j \mid j = 1, ..., m, i = 1, ..., n^j\}
and
Y = \{y_i^j \mid j = 1, ..., m, i = 1, ..., n^j\}
are the training indices and observations for the
m
tasks, where task
j
has
n^j
number of training data. To specify the affiliation of index
x_i^j
and observation
y_i^j
to task
j,
a label
U
has to be added as an additional input to the model with
l = j,
as shown in
Fig.1(b).

Two independent covariance functions may be assumed,
k_{MTGP}(x, x', l, l') = k_c(l, l') \times k_t(x, x')
(12)

where
k_c
and
k_t
represent the correlation between tasks and
temporal covariance functions within a task, respectively. Note
that
k_c
depends only on the labels
l,
and
k_t
depends only on
the indices
x.
Assuming that
n^j = n
for
j = 1, ..., m,
the covariance matrix
K_{MTGP}
for all
m
tasks can be written as

K_{MTGP}(X, l, \theta_c, \theta_t) = K_c(l, l') \otimes K_t(X, \theta_t)
(13)

where \otimes
is the Kronecker product,
1 = \{j \mid j = 1, ..., m\}
and
\theta_c
and
\theta_t
are vectors containing hyperparameters for
K_c
and
K_t,
respectively. This leads to a matrix of size
m \times m
for
K_{MTGP},
as
K_c
has a size of
m \times m
and
K_t
of
n \times n.
We refer to
K_c
as the correlation matrix. Within geostatistics,
this approach is also known as the intrinsic correlation model
[16]. Note the simplifying assumption
n^j = n
for
j = 1, ..., m,
may be relaxed so that the model can be easily extended to
use task-specific numbers of training data.

As with STGPs, the hyperparameters for a MTGP may be
optimised by minimising the NLML, and predictions for test
indices \{x^*, l^*\} can be made by computing the conditional probability
p(y^* | x^*, l, x, l, y).

This method has several useful properties:
- we may have task-specific training indices
n^j
(i.e., training data may be observed at task-specific times);
- automatic learning of the correlation within tasks occurs
by fitting the covariance function in (13); and
- the framework assumes that the tasks have similar
temporal characteristics and hyperparameters
\theta_i.

A limitation of the MTGP is that the computational cost for
evaluating MTGPs is \mathcal{O}(m^3 m^3) compared with
m \times \mathcal{O}(n^3)
for STGPs. Additionally, the number of hyperparameters can
increase rapidly for an increasing number of tasks which
can lead to a multi-modal parameter space with no overall
optimum.

C. Correlation matrix and normalisation

The remaining challenge is to construct a covariance function
K_c
which fulfils Mercer’s theorem (i.e., one that is a kernel function). The diagonal elements of
K_c
describe the correlation of the tasks with themselves and the non-diagonal
elements correspond to the correlation between tasks. For example, element
K_c[a, b]
represents the correlation between task
a
and
b
to which we refer as \tau_{MT}.
If
K_c
is the identity matrix, the tasks would be modelled independently and share
the same values of the hyperparameters
\theta_i.
Parameterisations
of
K_c
based on a spherical or hyperspherical decomposition were presented in [12], [13]. These parameterisations allow a direct interpretation of the correlation. However, they assume either equally-scaled observations for each task, or require additional scaling hyperparameters.

A more general solution is the so-called “free-form” parameterisation [9]. It is based on the Cholesky decomposition and parameterises the elements of the lower triangular matrix
L.

K_c = LL^T, \quad L = \begin{bmatrix}
\theta_{c,1} & 0 & \ldots & 0 \\
\theta_{c,2} & \theta_{c,3} & \ldots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
\theta_{c,k-m+1} & \theta_{c,k-m+2} & \ldots & \theta_{c,k}
\end{bmatrix}
(14)

where
k = m(m+1)/2
is the number of correlation hyperparameters. One advantage of this approach is that the diagonal elements of
K_c
are not forced to take a value of 1, which leads to an individual
y-scaling hyperparameter
\theta_y
for each task. However, the direct interpretation of
K_c
will become challenging as the elements are no longer take values between
−1
and 1.

To increase the interpretability of
K_c,
a transformation may be performed. We refer to the transformed correlation matrix as
K'_c
which is computed based on the normalised correlation hyperparameters
\theta'_c.
The proposed transformation is based on two constraints. First, the diagonal elements of
K'_c
are assumed to be 1 to eliminate the influence of label-specific scaling. Second, the non-diagonal elements of
K'_c
are computed based on the assumption that the contribution of
\theta'_c
to the diagonal elements of
K_c
is equivalent as of
\theta'_c
to
K'_c.
This leads to the following normalised hyperparameters
\theta'_c:

\theta'_c[l] = \text{sgn}(\theta_c[l]) \sqrt{\frac{\theta_c[l]^2}{\sum_{k=\delta_1+1}^{\delta_2} \theta_c[k]^2}}
(15)

where
\delta_1 = \sum_{i=1}^{j-1} i, \quad \delta_2 = \sum_{i=1}^{j} i, \quad \delta_1 < l \leq \delta_2
and
j \in \{1, ..., m\}.
Here
\theta_c[l]
indicates the
lth
element of the vector
\theta_c.
The normalised MTGP correlation coefficients \tau'_{MT}
can be computed according to (14).

D. Time shift estimation

The predictions provided by MTGPs depend strongly on the correlation between the modelled tasks. In extreme cases, if the tasks are not correlated, the prediction results of MTGPs are comparable to STGPs (however the tasks would still share the same temporal hyperparameter
\theta_t). As a consequence, the MTGP model is influenced if correlated signals are temporally
shifted to each other.

This effect can be compensated by an additional hyperparameter $\theta_s$ for each task which represents a time shift of the time line:

$$x^t = x + \theta^s$$

(16)

Assuming that all tasks will be modelled by an individual shift hyperparameter, $m - 1$ additional hyperparameters are required. If prior knowledge of the tasks is available, these hyperparameters can be easily further constrained such as that multiple tasks share a common hyperparameter $\theta_s$ or that two tasks have a time shift within a specific range. The hyperparameters can be automatically learned by optimising the NLML. This enables the possibility to use the MTGP framework for further applications, such as investigations of time shifts between tasks, concurrent with within- and between-task dynamical modelling.

E. Convolution of kernels

Up to this point, we have assumed that all tasks can be modelled with the same hyperparameter $\theta$. This limitation is mainly motivated by reducing the number of hyperparameters and can be inappropriate for specific applications. If it is required to model task-specific temporal behaviour, individual temporal covariance functions $k^l_t$ for $j = 1, \ldots, m$ can be introduced. Doing so, it has be guaranteed that the resulting matrix $K_t$ still fulfils the Mercer theorem.

In [17] the idea of convoluting two covariance functions has been presented, which has been further discussed in [18] for the case of MTGPs. As the convolution of two covariance functions is again a valid covariance function, a covariance function consisting of, for example, two different SE covariance functions with different hyperparameters $\theta_L$ can be computed as [18]

$$k_{SE \times SE}(r, l', t') = \sqrt{\frac{2\theta_L(l)\theta_L(l')}{\theta_L(l)^2 + \theta_L(l')^2}} \exp \left\{ -\frac{r^2}{\theta_L(l)^2 + \theta_L(l')^2} \right\}$$

(17)

where $\theta_L(l)$ is the x-scaling hyperparameters of task $l$. In the case of $l = l'$, we note that (17) is equivalent to (5). This enables to preserve the temporal characteristics of each tasks, while modelling them simultaneously within the MTGP framework.

III. RESULTS AND DISCUSSION

In the following subsections, we initially illustrate the above-mentioned properties with a synthetic dataset. Later, two real-world biomedical problems are addressed and the performance of MTGPs is compared with that of STGPs and existing domain-specific approaches.

A. Synthetic datasets

1) Prediction of multiple unevenly-sampled tasks: We first investigate if MTGPs can learn the correlation between multiple tasks ($m = 4$) and if this leads to improved prediction results. We use multivariate data presented in [19], which comprise time-series from three optical markers (OM), which are placed along the median line of one subject, at the chest (OM1), at the lower end of the sternum (OM2), and next to the navel (OM3). The 3-D position of each sensor was reduced to its first principle component. Additionally, the fourth task corresponds to a respiration belt (RB) that was placed around the torso next to OM2. Signals are shown in Fig.2(a). Table I lists acquisition parameters for each task. We note that the sampling frequency of OM1-3 is five times higher than that of RB and that task $j$ has $n^j$ number of training data as different training intervals $D^j$ are assumed.

The objective is to predict $y^*_{\hat{y}}$ for OM1 within a range of $x^* \in (20\, s, 70\, s)$. We investigate four evaluation scenarios (S1-S4). In the first scenario (S1), only the time-series from OM1 is considered, which would be equivalent to using a STGP. In scenarios S2-S4, the time-series for OM2, OM3, and RB are integrated successively into the MTGP model. Consequently scenario S4 will consider all 4 tasks. As we know that the physiology being modelled is respiratory motion, we can incorporate this prior knowledge by using a quasi-periodic covariance function (7). However, as the $y$-scaling of each task can be expressed by the correlation covariance matrix, the hyperparameter $\theta_A$ is omitted. The covariance matrix $K_c$ is initialized assuming independent tasks, by setting it to be the identity matrix. Table I also shows that Pearson’s correlation coefficient $\tau$ for all tasks with respect to OM1 indicates a highly positive or negative correlation. The MTGP model has to learn the correlation during training by optimising the NLML.

Fig.2(b-c) show the prediction results for the scenarios and the 95% confidence interval of each prediction. The plot for S1 reveals that the model closely fits the training data. However, soon after the last training observation, the predictions $\hat{y}^*$ tend towards the mean function (of zero). The practical interpretation of this results is that if $\| x^{OM1} - x^* \|_2$ is high, the correlation decreases and therefore the model predicts the prior distribution. Considering more tasks simultaneously, it may be seen that the prediction accuracy of $\hat{y}^*$ improves. As

<table>
<thead>
<tr>
<th>TASK</th>
<th>OM1</th>
<th>OM2</th>
<th>OM3</th>
<th>RB</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D^j$ [s]</td>
<td>(0, 20)</td>
<td>(10, 30)</td>
<td>(25, 40)</td>
<td>(0, 60)</td>
</tr>
<tr>
<td>$f_s$ [Hz]</td>
<td>2.6</td>
<td>2.6</td>
<td>2.6</td>
<td>0.62</td>
</tr>
<tr>
<td>$\tau$</td>
<td>1</td>
<td>-0.96</td>
<td>-0.9</td>
<td>0.89</td>
</tr>
</tbody>
</table>

| TABLE I TIME INTERVAL $D^j$ ENCLOSING THE TRAINING DATA FOR THE $j^{th}$ TASK, SAMPLING FREQUENCY $f_s$, AND PEARSON’S CORRELATION COEFFICIENT $\tau$ WITH RESPECT TO OM1 |

| TABLE II RMSE, NMSE AND MSLL OF $\hat{y}^*$ FOR PREDICTION SCENARIOS S1-S4 |

<table>
<thead>
<tr>
<th>SCENARIO</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>2.244</td>
<td>2.005</td>
<td>1.805</td>
<td>1.474</td>
</tr>
<tr>
<td>NMSE</td>
<td>0.973</td>
<td>0.776</td>
<td>0.629</td>
<td>0.42</td>
</tr>
<tr>
<td>MSLL</td>
<td>0.002</td>
<td>-0.278</td>
<td>-0.527</td>
<td>-0.669</td>
</tr>
</tbody>
</table>
Table II shows, the RMSE decreases from 2.244 for S1 to 1.474 for S4. Interestingly, the RMSE decreases also from S2 to S3, even though OM1 and OM3 do not have training data occurring at a shared interval: OM1, as shown in Table I, has training data in the interval (0, 20) s, while OM3 has training data in the interval (25, 40) s. It can be assumed that the MTGP is able to learn the correlation between OM1 and OM3 via the common intervals of training data for OM1-OM2 and OM2-OM3. Additionally, the correlation between OM1 and RB, which is the signal with the lowest sampling frequency, can be learned accurately, and which leads to a further decrease of the RMSE from 1.805 in S3 to 1.474 in S4. As more tasks are incorporated from S1 to S4, the complexity of the MTGP model increases which is also visible in the decreasing MSLL values.

2) Correlation analysis between tasks: As discussed in section II-C, the correlation coefficients of the MTGP models are influenced by the scaling of the individual tasks due to the selected "free-form" parameterisation. Consequently, it might be difficult to evaluate the resulting correlation coefficients and compare them. Considering the previous experiment, Fig.3(a) shows the resulting matrix $K_s$ for S4. The MTGP correlation coefficients $\tau_{MT}$ varies between $-2.15$ and $3.78$.

The transformed matrix $K_s'$ can be computed according to our proposal in (15). The results are shown in Fig.3(b). The normalised MTGP correlation coefficients $\tau_{MT}'$ show that OM1 is highly negatively correlated with OM2 and OM3, and positively correlated with RB. However, comparing these results with the Pearson’s correlation coefficients in Fig.3(c) reveals differences of up to 0.07. These differences were expected, because Pearson’s correlation coefficient is computed based on the values of the complete signals. In contrast, the MTGP correlation coefficients are based on the available training data and represent the correlation of the output function given by the MTGP model, which is influenced by many factors. (e.g., the temporal covariance functions, hyperparameters $\theta$, and the initialisation of the NLML optimisation). However, nonlinear correlation can be expressed by MTGPs through using different temporal covariance function, which could be useful for certain applications. However, it has to be further investigated how $\tau_{MT}'$ is related to other correlation metrics.

3) Time-shift estimation: To evaluate the influence of time shifts between tasks (section II-D), a new synthetic dataset is considered. Assume two sinusoidal signals are given with a frequency of 1 Hz and additive noise with distribution $\mathcal{N}(0,0.01)$. Let $x_n \in [0,2] s$ for task one and $x_n \in [0,7.5] s$ for task two be the known training data. A squared-exponential covariance function (5) is used as the temporal covariance function. The objective is to predict $y^*$ of task one for the test region $x^* \in (2,7.5) s$ depending on different phase shifts $\phi$ of task two. The mean prediction results of $y^*$ (red dashed lines) are shown in Fig.4(a-d), for $\phi = 0$ to $\phi = \pi/2$. As the phase shift increases from $\phi = 0$ to $\phi = \pi/2$, the normalised MTGP correlation decreases from $\tau_{MT}' = 1$ to $\tau_{MT}' = 0.06$ (Tab.III - the Pearson’s correlation coefficients are shown for comparison). This leads to an increased RMSE from 0.119 to 0.714. If the phase shift is further increased ($\phi > \pi/2$), the correlation becomes negative, the absolute correlation increases and the RMSE decreases. The minimum RMSE is reached for $\phi = \pi/2$ with 0.112 and a correlation of $\tau_{MT}' = -1$ (Fig.4(d)). Additionally, if $|\tau_{MT}'| < 1$, the phase shift of task two is projected into the prediction of $y^*$, as can be observed in Fig.4(b-c). The effect becomes predominant as the distance between a test feature $x^*_n$ to the last training feature $x^*_1$ increases.

The phase shift can be compensated by an additional hyperparameter $\theta_x$. The $\theta_x$ was implemented to shift the time index of task 2. The mean prediction results of a MTGP model with a shift hyperparameter (MTGP$\_\text{shift}$) are shown as green dotted-dashed line in Fig.4. It may be observed that $y^*_\text{pred}$ seems to be independent of the phase shift. This is confirmed by the
RMSE of MTGP, shown in Tab.III, which varies only slightly for different $\phi$. Furthermore, $\tau_{\text{MT}}$ and $\theta_s$ are shown Tab.III. The estimated shifts are 0.123 s and 0.247 s for a phase shift of $\phi = \pi/4$ and $\pi/2$ which correlates to the true shifts of 0.125 s and 0.25 s. However, it might be also observed that in cases where no $\theta_s$ is required ($\phi = \{0, \pi\}$), the RMSE error is slightly increased compared to a MTGP model without $\theta_s$, as an additional hyperparameter has to be trained.

4) Convoluted kernels: Fig.5(a-c) illustrate a synthetic example of three tasks with different temporal characteristics. The three tasks were generated to have the same increasing long-term trend. However, the signals differ in their short-term characteristics and noise components. A sinusoidal component was used in both approaches. Fig.6 shows the results obtained by individual temporal covariance functions $k_j$ for each task $j \in \{1, 2, 3\}$. According to section II-E, a correlation matrix $K_c$ can be constructed by convolution of the individual kernel functions. The results of the convoluted MTGP model with three individual SE covariance functions is shown in Fig.5(d-f). The individual hyperparameters $\theta_L$ are shown in Tab.IV and decrease from 1.18 s for task one down to 0.13 s for task 3. This means that the correlation is lower between two labels $y$ and $y'$ with a time difference $\Delta x$ for task three than for task one. Consequently, the short-term characteristics of task two and three can be better modelled resulting in a decreased RMSE, NMSE and MSLL.

B. Real-world datasets

1) Vital-sign analysis: Conventional procedures for hospital patient monitoring involve frequent measurements of vital signs by nursing staff, such as respiration rate, systolic blood pressure and heart rate. One of the limiting factors of existing methods for patient monitoring is deemed to be how they cope with these typically noisy, sparse and unevenly-sampled vital-sign data [4]. Also, some channels may be missing for long periods of time, which makes it difficult to model this type of dataset. We investigated the correlation between heart rate, respiration rate and systolic blood pressure observations from a cohort of patients recovering from cancer surgery in the Oxford University Hospitals NHS Trust. We applied the standard univariate STGP to each physiological time-series, which assumes independency between the variables, and compared the results with those obtained by using the MTGP method. In the latter, we used the convoluted kernels approach in order to allow for each channel (vital sign) to be modelled with different hyperparameters. The same optimisation algorithm was used in both approaches. Fig.6 shows the results obtained for task one.
for one example patient.

We can see in Fig. 6 that the MTGP provides a significant improvement in the estimated values for the three channels, because the correlation between the vital signs is taken into account. As expected, we found a high correlation between respiration rate and heart rate. Particularly, in periods of incomplete data (days 5 and 6 in Fig. 6), the predictions are more accurate, and can therefore be used to estimate the true value of the data and the distribution over the estimated values. We also note that by using the convoluted-kernel approach, each vital sign is modelled with independent hyperparameters, which may offer useful insights on the recovery of these patients from surgery.

2) Motion compensation in radiotherapy: In robotic radiotherapy, a challenging task is to perform active motion compensation to precisely radiate tumours in the liver or lung, which are constantly moving due to respiration. The internal tumour position is determined by fiducial markers, which can be tracked in X-ray images. As constant X-ray imaging is not possible throughout the treatment, the fiducial positions have to be predicted by a correlation model based on the position of external optical markers [20].

The correlation between internal and external motion is nonlinear, due to different trajectories for inhalation and exhalation, as shown in Fig. 7(a). Clinical practice is to approximate the correlation using a single or dual second-order polynomial model (SP or DP). An alternative approach was presented in [21] using support vector regression (SVR) which is based on the position and velocity of the external marker.

One drawback of both methods is that the majority of external data is discarded as only those training data can be used for which internal and external information are known at the same time. MTGP can overcome this problem, which we investigate using data\(^2\) from a porcine study [21]. Multiple external and internal markers were constantly measured at \(f_s = 2.96\) Hz. The subject was manually ventilated. The investigated internal signal fragment (150 data points) is shown in Fig. 7(b-d). The objective is to predict internal tumour position \(y^*\) depending on the number of internal training points \(n^{int}\) for SP, DP, the SVR approach and the MTGP with different \(k_1\) (eqs. 5-7) which we refer to as MTGP\(_{SE}\), MTGP\(_{PER}\), and MTGP\(_{QP}\), respectively. In general, a lower value of \(n^{int}\) is preferred, as it decreases the number of X-ray images that must be used, and thus the patients are exposed to less radiation. The goal is to map a time-series of spatial marker locations \((n^{int} = 150)\) onto internal fiducial measurements. For \(n^{int} = \{100, 75, 50\}\), the training data were randomly selected. To ensure that a complete breathing cycle is represented, the training data were selected manually for \(n^{int} = \{20, ..., 10\}\). A prior on \(\theta_P\) was specified, as it represents the breathing period. The residual hyperparameters were randomly initialised and the “optimal” hyperparameter set was selected based on the NLML. As discussed in section II-D, a shift hyperparameter \(\theta_S\) was added to the MTGP models to estimate explicitly the delay between internal and external markers that we know exists.

Table V shows the RMSE for all investigated models and the MLSS in parentheses for the MTGP models. Comparing the performance of the MTGP models reveals that MTGP\(_{QP}\) has the lowest RMSE and MLSS across all values of \(n^{int}\).

### Table V

RMSE AND MLSS IN PARENTHESES OF SINGLE AND DUAL POLYNOMIAL MODELS, SVR AND MTGP DEPENDING ON THE NUMBER OF TRAINING POINTS \(n^{int}\) (BEST RESULTS PER \(n^{int}\) HIGHLIGHTED BOLD)

<table>
<thead>
<tr>
<th>(n^{int})</th>
<th>SP</th>
<th>DP</th>
<th>SVR</th>
<th>SE</th>
<th>MTGP (_{PER})</th>
<th>MTGP (_{QP})</th>
<th>QP</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.845</td>
<td>0.226</td>
<td>0.361</td>
<td>0.228</td>
<td>(2.501)</td>
<td>1.906</td>
<td>(0.342)</td>
</tr>
<tr>
<td>75</td>
<td>0.854</td>
<td>0.233</td>
<td>0.370</td>
<td>0.199</td>
<td>(2.602)</td>
<td>1.833</td>
<td>(0.304)</td>
</tr>
<tr>
<td>50</td>
<td>0.851</td>
<td>0.227</td>
<td>0.361</td>
<td>0.321</td>
<td>(2.063)</td>
<td>1.67</td>
<td>(0.434)</td>
</tr>
<tr>
<td>20</td>
<td>0.870</td>
<td>0.988</td>
<td>0.410</td>
<td>0.476</td>
<td>(1.325)</td>
<td>1.91</td>
<td>(0.091)</td>
</tr>
<tr>
<td>15</td>
<td>0.881</td>
<td>0.542</td>
<td>0.421</td>
<td>0.26</td>
<td>(2.343)</td>
<td>1.962</td>
<td>(0.074)</td>
</tr>
<tr>
<td>10</td>
<td>0.949</td>
<td>0.542</td>
<td>0.406</td>
<td>0.396</td>
<td>(1.779)</td>
<td>1.834</td>
<td>(0.107)</td>
</tr>
</tbody>
</table>

\(^2\)http://signals.rob.uni-luebeck.de
This confirms the assumption of respiratory motion as being a quasi-periodic motion. In contrast, the RMSE and MSLL of MTGP\textsubscript{PER} are very high, indicating that a periodic model is not suitable here. This also leads to broader confidence intervals compared to the MTGP\textsubscript{SE} and MTGP\textsubscript{QP}, as is shown in Fig.7(b-d) for $n^{\text{int}} = 15$. In this case, only 15 internal training labels are known. The predicted internal motion is a result of the learned correlation to the external signal. Overall, MTGP\textsubscript{QP} outperforms all investigated approaches. We note also that, in addition to superior regression performance using the MTGP, it offers advantages over the SVR methods by providing explicit quantification of uncertainty in the prediction and robustness to missing or incomplete data. Further improvements of the MTGP approach can be expected by considering multiple external signals.

IV. DISCUSSION

The previous presented synthetic and real-world examples illustrate that MTGP's are a flexible approach that can be used for various biomedical applications. Even though this approach has several advantages compared to other methods, the correct design of the MTGP model and its covariance functions is essential for accurate prediction results. In general, a higher flexibility of the model (e.g., as allowed using convoluted kernels) comes with an increased number of hyperparameters. Consequently, the optimisation of these hyperparameters becomes more challenging as the hyperparameter space may be multimodal. Therefore, if prior knowledge of the function behaviour (e.g. quasi-periodic motion) or of the correlation between task is known it should be used to adapt this framework to the specific application. Furthermore to avoid local optima, priors on the hyperparameters could be included (e.g. the respiration rate in section III-B2). For unknown hyperparameters, the training phase can be repeated multiple times with randomly initialised hyperparameters. The best hyperparameter set can be selected based on the NLML.

Similarly, the “free-form” parameterisation of $\mathbf{K}_c$ allows high flexibility in the model. However this is only practical for a small number of tasks, as the number of hyperparameters increases by $m(m + 1)/2$. If for example a large sensor network has to be modelled, a different parameterisation with additional constraints on the correlation between sensors would be required.

One limitation of this approach is high computational costs $O(m^3 n^2)$, which is a general concern for the use of GP models. An overview of sparse GP methods to overcome this problem is presented in [3], [22], which aim to find a set of pseudo-inputs $\tilde{n}$, with $\tilde{n} \ll n$, to reduce computational complexity. In [9], [23], some of these techniques have been used to investigate sparse MTGPs which reduce the complexity to $O(m n \tilde{n}^2)$. A further decrease of the computational costs is possible by exploiting the Kronecker product [24], limiting the training data to the same time instances for each dimension [25], or by using recursive algorithms for online settings [26]. Applications that require close-to-real-time retraining, such as our application in respiratory motion compensation, would benefit from these techniques, while methods that operate over longer time-scales (such as analysis of vital-sign observations taken every hour) would be less sensitive to computational costs of this kind.

The focus of the examples presented here was the modelling of multivariate physiological datasets with univariate inputs, representing time series. We note that the model can be extended to arbitrary input dimensions, which could be useful for other biomedical applications, such as template matching or registration problems in imaging. A bivariate input space $\mathbf{x}$ representing the spatial coordinates of an image segment with the intensity as observations $y$ could be used with an MTGP, to estimate the optimal offset which maximises the correlation between two image segments. An advantage of this approach would be that each task could have task-specific number of training inputs (e.g., two MRI images with different resolution). Further applications are possible, regarding the analysis of the correlation between noisy signals. If for example the measurement noise is known, this information could be included in a MTGP model to compute a “noise-free” correlation coefficient between signals.

REFERENCES

Robert Dürichen received the Dipl.-Ing. degree (equivalent to M.Sc.) in mechatronics from Technical University Dresden, Germany, in 2009. In 2010, he joined the Institute of Robotics and Cognitive Systems at the University of Lübeck, where he is currently pursuing his PhD in non-invasive motion compensation techniques in radiotherapy. His research interests include signal processing and machine learning applied to multivariate biomedical and robotic applications.

Lei Clifton received a B.Sc. and M.Sc. in electrical engineering from Beijing Institute of Technology, China, and a PhD degree in electrical engineering from Manchester University, UK. After six years of post-doctoral research at the University of Oxford, she was appointed as a Medical Statistician at the Centre for Statistics in Medicine, University of Oxford. Her research interests include statistical signal processing, and machine learning for intelligent health monitoring systems.

Achim Schweikard received his PhD in robotics in 1989 from Technical University of Berlin, Germany. After his PhD he worked as a Research Associate at Stanford University Medical Center and as an Associate Professor for robotics and image processing at Technical University of Munich, Germany. In this position he laid the mathematical foundations for Cyberknife inverse planning. To date more than 200000 cancer patients worldwide have been treated with this technique for respiratory tracking. In 2002 he joined Lübeck University as a full professor to transfer technical and mathematical innovation to the clinic. His research areas of interest include robotics, machine learning, geometric reasoning, and clinical applications of robotics and navigation.

David A. Clifton is a tenure-track member of faculty in the Department of Engineering Science of the University of Oxford, and a Governing Body fellow of Balliol College, Oxford. He is a Research Fellow of the Royal Academy of Engineering. A graduate of the Department of Engineering Science at Oxford, Dr Clifton trained in information engineering and started the Computational Health Informatics (CHI) lab when we was appointed to the faculty in 2013. His research includes “big data” and mobile health projects, with a focus on machine learning.

Marco A.F. Pimentel received the B.Sc. and M.Sc. degrees in biomedical engineering from Universidade Nova de Lisboa, Portugal, in 2009. He was awarded a scholarship from the RCUK in 2010 and joined the Institute of Biomedical Engineering at the University of Oxford, where he is currently pursuing his D.Phil. in engineering science and healthcare innovation. His research interests include signal and image processing, and machine learning applied to the monitoring of complex systems and e-health.