Motivation  Decisions in the intensive care unit (ICU) are
frequently made in settings with a high degree of uncertainty
based on a wide variety of data sources, such as vital signs,
clinical notes, fluids, medications, etc. Clinical data collec-
tion is rapidly expanding, but these data are often sparse and
irregularly-sampled, and contaminated by a variety of noise
interference and human error. The ICU is playing an expand-
ing role in acute hospital care (Vincent 2013), and in such
data-heavy settings, a more concise representation of patient
state and plan care.

Solution  Our proposed technique transforms a variety of
irregularly-sampled clinical data into a new latent space us-
ing the hyperparameters of multi-task GP (MTGP) models.
Patients are compared based on their similarity in the new
hyperparameter space. Our work differs from other work
in that: 1) uses the correlation between and within multi-
time-series to estimate parameters instead of considering
each timeseries separately; 2) infers a compact latent repre-
sentation of the source data, rather than finding patterns that
are common within different timeseries; and 3) leverages the
hyperparameter space. Our work differs from other work
on a single timeseries, but fail to generalize to (or take
advantage of) other related time-series data. In the remain-
der of this paper, we refer to noisy, sparse, heterogeneous,
irregularly-sampled data as "irregularly-sampled" data.

1 Introduction

Goal  High quality clinical care depends on the ability to
combine heterogeneous clinical data to understand the sever-
it of illness (acuity) in patients. Clinical research often
uses risk of mortality as a surrogate for patient acuity, often
evaluated at a single end point, such as after 28-days post-
discharge. Most acuity scores rely on static snapshots of a
patient and do not incorporate evolving clinical information
such as new notes, lab values, etc. Our goal is to provide
a concise representation of these multiple related timeseries
so that they can be compared and assessed.

Challenge  The general issue of comparing signals that are
not aligned and irregularly sampled has been considered be-
fore (see 2.2). Establishing similarity metrics among time-
series data is an important part of many learning tasks and
often is achieved using a variety of summarization methods.
However, many modeling methods fail when applied to ir-
regularly sampled data unless strong assumptions are made
about the functional form present in the underlying data
source. Furthermore, in cases where such methods work,
data imputation is often necessary, which can introduce ad-
ditional sources of error and bias. Finally, many methods
work on a single timeseries, but fail to generalize to (or take
advantage of) other related time-series data. In the remain-
der of this paper, we refer to noisy, sparse, heterogeneous,
irregularly sampled data as "irregularly-sampled" data.

Abstract

The ability to determine patient acuity (or severity of
illness) has immediate practical use for clinicians. We
evaluate the use of multivariate timeseries modeling
with the multi-task Gaussian process (GP) models using
noisy, incomplete, sparse, heterogeneous and unevenly-
sampled clinical data, including both physiological sig-
nals and clinical notes. The learned multi-task GP
(MTGP) hyperparameters are then used to assess and
forecast patient acuity. Experiments were conducted
with two real clinical data sets acquired from ICU pa-
patients: firstly, estimating cerebrovascular pressure reac-
tivity, an important indicator of secondary damage for
traumatic brain injury patients, by learning the inter-
actions between intracranial pressure and mean arterial
blood pressure signals, and secondly, mortality predic-
tion using clinical progress notes. In both cases, MTGPs
provided improved results: an MTGP model provided
better results than single-task GP models for signal in-
terpolation and forecasting (0.91 vs 0.69 RMSE), and
the use of MTGP hyperparameters obtained improved
results when used as additional classification features
(0.812 vs 0.788 AUC).

1 Introduction

Motivation  Decisions in the intensive care unit (ICU) are
frequently made in settings with a high degree of uncertainty
based on a wide variety of data sources, such as vital signs,
clinical notes, fluids, medications, etc. Clinical data collec-
tion is rapidly expanding, but these data are often sparse and
irregularly-sampled, and contaminated by a variety of noise
interference and human error. The ICU is playing an expand-
ing role in acute hospital care (Vincent 2013), and in such
data-heavy settings, a more concise representation of patient
records would help clinical staff to quickly assess patient
state and plan care.

1 The first two authors contributed equally on this work.
2 Corresponding author.
Copyright © 2015, Association for the Advancement of Artificial
Intelligence (www.aaai.org). All rights reserved.

A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment
and Forecasting in ICU with Sparse, Heterogeneous Clinical Data

Marzyeh Ghassemi
Computer Science
MIT
Cambridge, MA 02139 USA
mghassem@mit.edu

Marco A.F. Pimentel *
Engineering Science
University of Oxford
Oxford, UK
marco.pimentel@eng.ox.ac.uk

Tristan Naumann
Computer Science
MIT
Cambridge, MA
tjn@mit.edu

Thomas Brennan
Health Science Technology
MIT
Cambridge, MA
tpb@mit.edu

David A. Clifton
Engineering Science
University of Oxford
Oxford, UK
davidc@robots.ox.ac.uk

Peter Szolovits
Computer Science
MIT
Cambridge, MA
psz@mit.edu

Mengling Feng †
Health Science Technology
MIT
Cambridge, MA
mfeng@mit.edu

Abstract

The ability to determine patient acuity (or severity of
illness) has immediate practical use for clinicians. We
evaluate the use of multivariate timeseries modeling
with the multi-task Gaussian process (GP) models using
noisy, incomplete, sparse, heterogeneous and unevenly-
sampled clinical data, including both physiological sig-
nals and clinical notes. The learned multi-task GP
(MTGP) hyperparameters are then used to assess and
forecast patient acuity. Experiments were conducted
with two real clinical data sets acquired from ICU pa-
patients: firstly, estimating cerebrovascular pressure reac-
tivity, an important indicator of secondary damage for
traumatic brain injury patients, by learning the inter-
actions between intracranial pressure and mean arterial
blood pressure signals, and secondly, mortality predic-
tion using clinical progress notes. In both cases, MTGPs
provided improved results: an MTGP model provided
better results than single-task GP models for signal in-
terpolation and forecasting (0.91 vs 0.69 RMSE), and
the use of MTGP hyperparameters obtained improved
results when used as additional classification features
(0.812 vs 0.788 AUC).

1 Introduction

Motivation  Decisions in the intensive care unit (ICU) are
frequently made in settings with a high degree of uncertainty
based on a wide variety of data sources, such as vital signs,
clinical notes, fluids, medications, etc. Clinical data collec-
tion is rapidly expanding, but these data are often sparse and
irregularly-sampled, and contaminated by a variety of noise
interference and human error. The ICU is playing an expand-
ing role in acute hospital care (Vincent 2013), and in such
data-heavy settings, a more concise representation of patient
records would help clinical staff to quickly assess patient
state and plan care.

1 The first two authors contributed equally on this work.
2 Corresponding author.
Copyright © 2015, Association for the Advancement of Artificial
Intelligence (www.aaai.org). All rights reserved.
Contributions  This paper makes the following contributions:

- We propose a method using MTGP for forecasting patient acuity based on irregularly sampled heterogeneous clinical data.
- We propose a new latent space for representing multi-dimensional timeseries using inferred MTGP hyperparameters.
- We evaluate our approach in two ways: 1) estimating and forecasting a cerebrovascular autoregulation index from noisy physiological time-series data in patients who suffered a traumatic brain injury and 2) transforming irregular ICU patient clinical notes into timeseries, and using MTGP hyperparameters from these timeseries as features to predict mortality probability.

2 Related Work

2.1 Clinical Assessment

In the clinical world, there are practical examples of data being used to infer patient acuity in the form of ICU scoring systems. ICU scoring systems such as SAPS (simplified acute physiology score) use physiologic and other clinical data for acuity assessment. However, in 2012 scoring systems were used in only 10% to 15% of US ICUs (Breslow and Badawi 2012). Recent work has focused on feature engineering for mortality prediction. This is usually accomplished by windowing or aggregating the structured numerical data so that a single feature matrix can be fed into a structured deterministic classifier (Hug and Szolovits 2009; Lehman et al. 2012; Joshi and Szolovits 2012; Ghassemi et al. 2014).

2.2 Timeseries Abstraction

The timeseries abstraction/summarization literature deals more directly with the time-varying nature of data. Dynamic time warping measures similarity between two temporal sequences that may vary in time or speed (Li and Clifford 2012). Another approach is time-series symbolization, which involves discretizing timeseries into sequences of symbols and attaching meaning to the groupings of the symbols (Lin et al. 2007; Saeed and Mark 2006; Syed and Guttag 2011). These approaches rely on some known regularity underlying a signal (e.g. ECG signals), and are often unsuitable for irregularly sampled timeseries. Full latent variable models have been applied to abstracting signals into higher level representations. For example, Fox et al. used beta processes to model multiple related timeseries (Fox et al. 2011), and Marlin et al. used Gaussian mixture models on the first 24 hours of monitor-signals data with hourly-discretization (Marlin et al. 2012). Nevertheless, latent variable approaches are unable to cope with missing and unevenly-sampled data as is, and require either strong assumptions about observations when they change asynchronously, or the computationally expensive approach of modeling time between observations directly as another latent variable.

2.3 Gaussian Processes

Gaussian processes (GP) form the basis for a Bayesian modeling technique that has been used for various machine learning tasks (Rasmussen and Williams 2006). Most commonly, GPs are used to predict a single output (denoted here as “task”) based on one or more input timeseries. We refer to this model as a single-task GP (STGP). Lasko et al. attempted to use Gaussian process regression as a smoothing function of irregularly-sampled signals (Lasko, Denny, and Levy 2013). This is a common usage model for GPs on clinical timeseries: GPs are used to model observed data through the predicted mean function of the timeseries. Clifton et al. used GPs as a framework for coping with data artifacts and incompleteness in mobile sensor data (Clifton et al. 2013b). In a related work (Clifton et al. 2013a), a functional version of extreme value statistics was proposed for physiological data in order to compare different timeseries. Similarly, GPs were used for robust regression of noisy heart rate data (Stegle et al. 2008). The remainder of the related work has used STGP models to predict a single output based on one or more input variables.

3 Methods

In the present study, we explore the potential of a novel approach using MTGP models (Bonilla, Chai, and Williams 2007) to learn the correlation between and within timeseries, and obtain a concise representation of time-varying physiological and clinical data based on the inferred hyperparameters.

Here, we motivate the use of MTGP and describe the method (source code is available on-line1) that we have adapted for hyperparameter construction (Durichen et al. 2014).

---

1http://www.robots.ox.ac.uk/~davidc/publications_MTGP.php
3.1 Multi-Task Gaussian Process Models

The general STGP framework may be extended to the problem of modeling $m$ tasks simultaneously where each model uses the same index set $x$ (e.g., physiological or clinical timeseries). A naive approach is to train a STGP model independently for each task, as illustrated in Figure 1(a). We introduce instead an extension to multi-task GP models proposed in (Bonilla, Chai, and Williams 2007), which makes use of the covariance in related tasks to reduce uncertainty in the inferred signal.

Let $X_n = \{x^j_i \mid j = 1, ..., m, i = 1, ..., n_j\}$ and $Y_n = \{y^j_i \mid j = 1, ..., m, i = 1, ..., n_j\}$ be the training indices and observations for the $m$ tasks, where task $j$ has $n_j$ number of training data. We consider the regression model $\hat{y}_n = g(\bar{x}_n) + \epsilon$, in which $g(x)$ represents the latent function and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is a noise term. GP models assume that the function $g(\bar{x}_n)$ can be interpreted as a probability distribution over functions such that $\hat{y}_n = g(\bar{x}_n) \sim \mathcal{GP}(m(\bar{x}_n), k(\bar{x}_n, \bar{x}_n'))$, where $m(\bar{x}_n)$ is the mean function of the process (assumed = 0) and $k(\bar{x}_n, \bar{x}_n')$ is a covariance function describing the coupling among the independent variables $\bar{x}_n$, as a function of their kernel distance. To specify the affiliation of index $x^j_i$ and observation $y^j_i$ to task $j$, a label $l^j = j$ is added as an additional input to the model, as shown in Figure 1(b). To model the correlation between tasks as well as the temporal behaviour of the tasks within a unified GP model, two independent covariance functions are assumed, and the covariance matrix $K_{MT}$ for all $m$ tasks can be written

$$K_{MT}(X_n, l, \theta_c, \theta_t) = K_c(\theta_c) \otimes K_t(X_n, \theta_t)$$

where $\otimes$ is the Kronecker product, $l = \{j \mid j = 1, ..., m\}$, $K_c$ and $K_t$ represent the correlation and temporal covariance functions, and $\theta_c$ and $\theta_t$ are vectors containing hyperparameters for $K_c$ and $K_t$, respectively. Within geostatistics, this approach is also known as the intrinsic correlation model (Wackernagel 2003).

By modifying the temporal covariance function we can encode our prior knowledge concerning the functional behavior of the tasks that we wish to model. The most frequently-used example is the squared-exponential covariance function (Rasmussen and Williams 2006):

$$K_t = \theta^2_t \exp\left\{-\frac{\|x - x'\|^2}{2\theta^2_t}\right\},$$

where $\theta_t = \{\theta_A, \theta_L\}$, and $\theta_A$ and $\theta_L$ are hyperparameters modeling the $y$-scaling and $x$-scaling (or time-scale if the data are timeseries) of the covariance function, respectively.

To construct a valid positive semidefinite correlation covariance function $K_c$, we used the Cholesky decomposition and the “free-form” parameterization of the elements of the lower triangular matrix $L$ proposed in (Bonilla, Chai, and Williams 2007), such as

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

where $k = \sum_{i=1}^m i$ is the number of correlation hyperparameters.

Identically to STGPs, the hyperparameters $\theta$ for a MTGP may be optimized by minimizing the negative log marginal likelihood via gradient descent (Rasmussen and Williams 2006), and predictions for test indices $\{x^*_n, y^*_n\}$ can be made by computing the conditional probability $p(y^*_n \mid x^*_n, l, x_n, Y_n)$.

Figure 2 shows an example of STGPs and an MTGP applied to a simple synthetic dataset with 4 sample tasks. Tasks 1 and 2 were correlated, task 1 and task 2 were both anti-correlated with task 4, and task 3 was uncorrelated with all other tasks. For this, 4 tasks were sampled from a MTGP model with the following hyperparameters: $\theta_{t,1} = \theta_{A,1} = \theta_{c,1} = \theta_{c,2} = \theta_{c,3} = \theta_{c,6} = \theta_{c,10} = 1$, $\theta_{c,4} = \theta_{c,5} = \theta_{c,0} = 0$, and $\theta_{c,7} = \theta_{c,8} = -1$. Artificial gaps were then randomly created in different tasks at different time points and with different durations. The STGP

Figure 2: (a) A sample function with 4 tasks; (b) Single-task GP (STGP) and (c) multi-task GP (MTGP) predictions on all tasks. The dots represent observations, while dashed lines and colored areas represent the predictive mean and 95% confidence interval, respectively. The line on the bottom represents the mean absolute error (over the 4 tasks) between the predictions and the correspondent reference values. We observe that the overall error obtained in (c) is lower than that in (b), which suggests that the use of MTGP yielded better predictions by taking into account the correlation between the different tasks.
Figure 3: An example of a single-task GP (STGP) and multi-task GP (MTGP) applied to intracranial pressure (ICP) and mean arterial blood pressure (ABP) signals from a traumatic brain injury patient. (a) and (c) show the performance of STGP, whereas (b) and (d) show the improved performance of MTGP, which takes into account the correlation between ICP and ABP. Dots represent observations, crosses represent missing observations (test observations), the dotted line shows the function mean and the shaded area show the 95% confidence interval. We note that the timescale parameter “selected” by the MTGP, which takes into account the correlation between the tasks, is shorter than the one selected by the STGP, which yields to higher likelihood of the test observations (crosses).

The MTGP has several useful properties as compared to the traditional GP:
- We can allow task-specific training indices $n_j$; i.e., training data may be observed at different times for different tasks (Figure 2);
- The correlations within and between tasks are automatically learned from the data by fitting the covariance function in Equation 1; and
- The framework assumes that the tasks have similar temporal characteristics and hyperparameters $\theta_t$.

A limitation of the MTGP is computational cost: $O(m^3 n^3)$ compared with $m \times O(n^3)$ for STGPs. This limitation is not as relevant for our application, given that we are not dealing with densely-sampled time-series data, but data which is sparse and irregular. Another limitation of the MTGP is that the number of hyperparameters can increase rapidly for an increasing number of tasks, which can lead to a multi-modal parameter space.

3.2 Signal Representation via Hyperparameters

We propose using the inferred MTGP hyperparameters $\theta$ that describe the temporal correlation within and between tasks as features that represent our set of observations: $\theta_A$ and $\theta_L$, which respectively govern each output scale of our functions and the input, or time, scale, and $\theta_{c,i}$ that correspond to the correlation between the different tasks (outputs) modelled. In effect, $\theta$ provides a new latent search space to examine and evaluate the similarity of any two given multidimensional functions. Importantly, these parameters are:

1. a means of representing the functional behavior a set of observations $\{y_n, x_n\}$;
2. learned directly from data; and
3. generalizable to any type of longitudinal data, including categorical and numerical types.

4 Experiment 1: From Multiple Noisy Time-Series Data to Acuity Assessment

In this experiment, we use physiological signals from Traumatic Brain Injury (TBI) patients to test the MTGP’s ability to assess and forecast multiple related signals. We examine two noisy timeseries: the intracranial pressure (ICP) and mean arterial blood pressure (ABP). Continuous monitoring of ICP and ABP has become a standard in neurological ICUs. Cerebrovascular autoregulation is an important mechanism to sustain adequate cerebral blood flow (Werner and Engelhard 2007), and impairment of this mechanism indicates an increased risk to secondary brain damage and mortality (Hlatky, Valadka, and Robertson 2005).

Cerebrovascular autoregulation is most commonly assessed based on the Pressure-Reactivity Index (PRx), which is defined as a sliding window Pearson’s correlation between the ICP and ABP (Czosnyka et al. 1997). However, the ICP and ABP timeseries are often contaminated by artifacts and missing data, and PRx can no longer be calculated in these situations. Although methods have been proposed to detect and remove artifacts (Feng et al. 2011), the artifact removal process still creates gaps of missing data in the timeseries.

In this experiment, we demonstrate how the proposed MTGP model can be applied to interpolate the incomplete data in ICP and ABP signals and, more importantly, to accurately estimate PRx.

4.1 Data

The ICP and ABP data were collected from 35 TBI patients who were monitored for more than 24-hours in a Neuro-ICU of a tertiary care hospital between January 2009 and December 2010. The continuously monitored physiological readings were sampled and recorded every 10 seconds. For experimental evaluation, we selected 30 ten-minute windows from each patient recording, where ICP and ABP signals
were free from artifacts and missing values. We then randomly introduced artificial gaps in both signals as shown in Figure 3. We evaluated the PRx estimation accuracy, and we further compared the performance of MTGP to that of STGP, which models each signal independently. For implementation, priors over the hyperparameters were selected after 100 random initializations for each case.

4.2 Results
The quality of predictions are evaluated using the squared error loss, where we compute the squared residual $(y^* - \hat{y}^*)^2$ between the mean prediction ($\hat{y}^*$) and the target ($y^*$) at each test point, and the squared root of the average over the test set to produce the root mean squared error (RMSE). As the RMSE is sensitive to the overall scale of the target values, we additionally evaluate the negative log probability of the target under the model, by defining the mean standardized log loss (MSLL) as

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

where the first term is the log likelihood of $\hat{y}_i^*$ given our latent function $f$ and the test index $x_i^*$. This probability is normalized by the second term, the log likelihood of $\hat{y}_i^*$ under a trivial model that predicts using a Gaussian with mean $m(y_n)$ and variance $\var(y_n)$ of the training labels.

Table 1: Performance of single-task GP (STGP) and multi-task GP (MTGP). PRx-PRx* refers to the difference between the reference PRx (Pearson correlation coefficient of ICP and ABP for a given window) and PRx*, the estimated PRx index (posterior MTGP hyperparameter that measures the interaction between the two tasks).

<table>
<thead>
<tr>
<th>Signal</th>
<th>Measure</th>
<th>STGP</th>
<th>MTGP</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICP</td>
<td>RMSE</td>
<td>0.91</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>MSLL</td>
<td>0.6</td>
<td>0.45</td>
</tr>
<tr>
<td>ABP</td>
<td>RMSE</td>
<td>2.77</td>
<td>1.98</td>
</tr>
<tr>
<td></td>
<td>MSLL</td>
<td>0.65</td>
<td>0.55</td>
</tr>
<tr>
<td>PRx-PRx*</td>
<td>RMSE</td>
<td>-</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Table 1 shows the overall performance of our approach. We note that the MTGP was able to estimate the correlation between the ICP and ABP signals – PRx – accurately even with incomplete data. The average RMSE between the true correlation coefficients and the MTGP estimated ones with the incomplete data was 0.09 (Table 1). This suggests that the posterior hyperparameter of MTGP, which measures the interactions between ICP and ABP, may be used as an index to model the cerebrovascular autoregulation mechanism and thus the risk of secondary brain injury.

We note that the scale of ICP values is normally between 1 to 20 mmHg, and the specific ICP value determines whether the achieved reduction in RMSE is clinically significant. If the ICP has already elevated to somewhere near 20 mmHg, any slight increase in ICP may result in secondary damage to the brain. In this case, even small reductions to RMSE are more desirable to guide the medical interventions.

We also observe that the MTGP provides a significant improvement in interpolating values for both signals, as the correlation between the two physiological variables is taken into account. Particularly, in periods of incomplete data (see Figure 3), the predictions are much more accurate compared to STGP. This shows that the proposed MTGP model can also be used for accurate interpolation and forecasting of ICP and ABP timeseries in the applications of advanced alarming and physiological trajectory analysis.

5 Experiment 2: From Heterogenous Clinical Data to ICU Acuity Forecasting
To demonstrate the effectiveness of the proposed MTGP model on features inferred from sparse, irregularly sampled timeseries, we applied MTGPs to clinical notes from the ICU for mortality prediction as summarized in Figure 4. Gold-standard clinical models typically use population-based acuity scores, such as SAPS-I (Le Gall et al. 1984), based on snapshots of the patient’s status during their stay in the ICU. These scores are inherently limited because patient state (or severity of illness) constantly evolves.

5.1 Data
We used 2001–2006 ICU data from the open-access MIMIC II 2.6 database (Saeed et al. 2011), which includes electronic medical records (EMRs) for 26, 870 ICU patients at the Beth Israel Deaconess Medical Center (BIDMC).

For each patient we extracted the SAPS-I score, calculated from clinical variables over a patient’s first 24-hours in the ICU. We used all notes from nursing, physicians, labs, and radiology recorded prior to the patient’s first discharge from ICU. Discharge summaries were excluded because they typically state the patient’s outcome explicitly. Patients were excluded if their notes had fewer than 100 words, fewer than 6 total notes in their record, or were under the age of 18. Patient mortality outcomes were measured at hospital discharge and 1 year post-discharge.

The final cohort consisted of 10,202 patients with 313,461 notes. A random 30% of the patients (3,040) were held back as a test set. The remaining 70% of patients (7,162) were used to train topic models and mortality predictors. The test set contained 93,411 notes, and the training set had 220,005.

5.2 Clinical Note Decomposition to Timeseries
Beginning from sparse, irregularly sampled clinical notes, we first performed topic modeling as a form of dimensional reduction as described in (Ghassemi et al. 2014). Topics inference was performed on notes using $T = 50$ topics over the words ($W$) in our vocabulary (Blei, Ng, and Jordan 2003; Griffiths and Steyvers 2004). We normalized hyperparameters on the Dirichlet priors for the topic distributions ($\alpha$) and the topic-word distributions ($\beta$) as $\alpha = \frac{K}{T}$ and $\beta = \frac{K}{NW}$.

The topic inference resulted in a 50-dimensional vector of topic proportions for each note in every patient’s record. We
identified topics (matrix $q$), and (3) the inferred MTGP hyperparameters across the nine topic vectors from $q$. Importantly, the admitting SAPS-I score and mean topic members (1 and 2) are both \textit{static measures}. SAPS-I collapses data from the first 24 hours of the record, while the average topic membership collapses the entire per-note timeseries for each patient’s record into an aggregate measure. Our proposed MTGP hyperparameters (3) complement these measures with information about the per-note timeseries.

Table 2: Top five words in chosen topics (enriched for in-hospital mortality/survival).

<table>
<thead>
<tr>
<th>Top Five Words</th>
<th>Possible Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>liver, renal, hepatic, ascites, dialysis</td>
<td>Renal Failure</td>
</tr>
<tr>
<td>thick, secretions, vent, trach, resp</td>
<td>Respiratory infection</td>
</tr>
<tr>
<td>remains, family, gtt, line, map</td>
<td>Systematic organ failure</td>
</tr>
<tr>
<td>increased, temp, hr, pt, cc</td>
<td>Multiple physiological changes</td>
</tr>
<tr>
<td>intubated, vent, ett, secretions, propofol</td>
<td>Respiratory failure</td>
</tr>
<tr>
<td>name, family, neuro, care, noted</td>
<td>Discussion of end-of-life care</td>
</tr>
<tr>
<td>cabg, pain, ct, artery, coronary</td>
<td>Cardio-vascular surgery</td>
</tr>
<tr>
<td>chest, pneumothorax, tube, reason, clip</td>
<td></td>
</tr>
<tr>
<td>pain, co, denies, oriented, oriented, neuro</td>
<td>Responsive patient</td>
</tr>
</tbody>
</table>

5.4 Outcome Classification

We considered five feature prediction regimes that combined subsets of the feature matrices 1, 2, and 3 as an aggregate feature matrix. We trained two supervised classifiers that were identical in the five feature sets used, but provided different objective functions for optimization: Lasso logistic regression and L2 linear kernel SVM.

Classifiers were trained to create classification boundaries for two clinical outcomes: in-hospital mortality and 1-year post-discharge mortality. All outcomes had large class imbalance (e.g., in-hospital mortality rates of 10.9%). To address this issue, we randomly sub-sampled the negative class in the training set to produce a minimum 70%/30% ratio between the negative and positive classes. Test set distributions were not modified, and reported performance reflects those distributions. Due to space constraints, we only reported results on a completely held out test set. We performed 5-fold cross-validation on the remaining data, and cross-validation results were similar to those obtained on the completely held-out test set.

We evaluated the performance of all classifiers using the area under the Receiver Operating Characteristic curve (AUC) on the held-out test set. Table 3 reports results from the Lasso model. Results obtained using the L2 linear kernel SVM were not statistically different.
The main limitation in using this approach to characterize timeseries is computational cost. We conducted an exhaustive grid search over the constrained hyperparameter space. Computational costs may be addressed using a recently proposed Bayesian optimization for automatically tuning the MTGP hyperparameters (Swersky, Snoek, and Adams 2013) in large datasets. In a "real-time" setting, the computational cost for \( m \) tasks is \( O(n^3, n^3) \). An overview of sparse GP methods is presented in (Quinonero-Candela and Rasmussen 2005), which aims to find a smaller set of pseudo-inputs to reduce computational complexity. Further improvement is possible by 1) exploiting the Kronecker product (Stegle et al. 2011), 2) limiting the training data to the same time instances of each dimension of the data (Evgeniou, Michelli, and Pontil 2005), or 3) by using recursive algorithms (Pillonetto, Dinuzzo, and De Nicolao 2010). Applications that require close-to-real-time retraining (e.g. Experiment 2), would benefit from these techniques, while operating over longer time-scales would be less sensitive.

7 Acknowledgements
This research was funded in part by the Intel Science and Technology Center for Big Data, the National Library of Medicine Biomedical Informatics Research Training grant (NIH/NLM 2T15 LM007092-22), and the R01 grant EB001659 from the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health (NIH). MAFP was supported by the RCUK Digital Economy Programme and FCT, Portugal. DAC was supported by a Royal Academy of Engineering Research Fellowship and by the Wellcome Trust and EPSRC. MF was supported by the A*STAR Graduate Scholarship.

Appendix
References


Vincent, J.-L. 2013. Critical care—where have we been and where are we going? *Critical Care* 17(Suppl 1):S2.
