BMC Medical Informatics and Decision Making

TECHNICAL ADVANCE Open Access

Sparse multi-output Gaussian processes for online medical time series prediction

Li-Fang Cheng[1], Bianca Dumitrascu[2], Gregory Darnell[2], Corey Chivers[3], Michael Draugelis[3], Kai Li[4] and Barbara E Engelhardt[4,5*]

Abstract

Background: For real-time monitoring of hospital patients, high-quality inference of patients' health status using all information available from clinical covariates and lab test results is essential to enable successful medical interventions and improve patient outcomes. Developing a computational framework that can learn from observational large-scale electronic health records (EHRs) and make accurate real-time predictions is a critical step. In this work, we develop and explore a Bayesian nonparametric model based on multi-output Gaussian process (GP) regression for hospital patient monitoring.

Methods: We propose MedGP, a statistical framework that incorporates 24 clinical covariates and supports a rich reference data set from which relationships between observed covariates may be inferred and exploited for high-quality inference of patient state over time. To do this, we develop a highly structured sparse GP kernel to enable tractable computation over tens of thousands of time points while estimating correlations among clinical covariates, patients, and periodicity in patient observations. MedGP has a number of benefits over current methods, including (i) not requiring an alignment of the time series data, (ii) quantifying confidence regions in the predictions, (iii) exploiting a vast and rich database of patients, and (iv) inferring interpretable relationships among clinical covariates.

Results: We evaluate and compare results from MedGP on the task of online prediction for three patient subgroups from two medical data sets across 8,043 patients. We find MedGP improves online prediction over baseline and state-of-the-art methods for nearly all covariates across different disease subgroups and hospitals.

Conclusions: The MedGP framework is robust and efficient in estimating the temporal dependencies from sparse and irregularly sampled medical time series data for online prediction. The publicly available code is at https://github.com/bee-hive/MedGP.

Keywords: Gaussian processes, Electronic health records, Sparse time series, Spectral mixture kernel

Background

Large-scale collections of electronic health records (EHRs) are becoming useful for understanding disease progress, early diagnosis, and personalized treatments for many clinical diseases [1–3]. EHRs contain rich patient information—disease history, demographics, vital signs, and lab results—that clinicians use to diagnose and treat patients. In this work, we are interested in developing a statistical framework that leverages medical data from a set of reference patients to enable personalized, real-time monitoring of new hospital patients. In particular, we consider data from the Hospitals at the University of Pennsylvania (HUP) containing information for over 260,000 patients, and the public Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-III) data set with more than 53,000 admissions from 38,000 patients in intensive care units (ICUs) [4].

*Correspondence: bee@princeton.edu
[4]Department of Computer Science, Princeton University, Princeton, NJ USA
[5]Center for Statistics and Machine Learning, Princeton University, Princeton, NJ, USA
Full list of author information is available at the end of the article

Robust models of patient state are essential as the basis for important downstream analyses of patient data. In particular, these models allow smoothing of noisy data across time, estimates of patient clinical covariates values and uncertainty in those estimates at any time point, and forecasting of patient state based on trends of specific covariates across time. For example, we might wish to predict the time-to-event for septic shock based on patient state. Early diagnosis of sepsis is extremely effective at reducing the mortality rate of sepsis. Sepsis is one of the leading causes of death in critically ill patients in the United States [5]. Each year an estimated 750,000 cases of sepsis or septic shock occur in the US. The mortality rate of septic patients ranges from 20% to 30%, and accounts for roughly 9.3% of all US deaths [6, 7]. Sepsis is often developed during a patient's stay in the hospital. However, accurate diagnosis of sepsis is difficult due to heterogeneous symptoms across patients [8].

A time-to-event prediction for septic shock would greatly improve if it were built upon an underlying model of patient state. Predicting septic shock without a model of patient state is challenging: Many of the covariates, lab results in particular, are sparsely sampled across patients. For example, vital signs (respiratory rate, heart rate, systolic blood pressure, and body temperature) are generally taken once every three to four hours for inpatient data, and once every hour for patients in the intensive care unit (ICU). Blood tests requiring a blood draw are generally performed at most once a day (Fig. 1; Table 1). Data missingness is systematic and not at random [9]: a doctor will generally order a test to inform patient state relevant to a specific diagnosis. Time-to-event models thus benefit greatly from the use of a patient state model to avoid these challenging properties of medical data in the downstream analysis.

However, these inpatient data also pose challenges to developing patient state models. In particular, these time series data are not aligned across patients to a reference time point or disease onset; instead, patient intake is at time 0 and release is hours or days later. The time intervals between observations are non-uniform, and no two observations are generally taken at the same time. The sparsity over patients and uncalibrated time series make the physiological progression of patient state within patients or joint analysis of time series across patients difficult to model using many existing time series analyses.

In this work, we build a statistical framework that uses sparse, heterogeneous EHR time series data to monitor and predict vital signs and lab results for each patient in an online way. To do this, we first designed a nonparametric model based on Gaussian process (GP) multivariate regression to explore the correlations both within each clinical covariate across time and across clinical covariates given rich EHR reference data. Our model includes a highly structured GP kernel regularized using sparsity-inducing priors to avoid overfitting, allow interpretability, and ensure computational tractability. Second, we propose a framework based on nonparametric density estimation to tailor the empirical model to a patient-specific model for each new patient. For real-time monitoring, we update the empirical distribution from reference patients with patient-specific observations as measurements are observed. We evaluate our method, MedGP, on over 6,000 patients from three disease groups with more than four million measurements from the HUP data, and on one disease group from the MIMIC-III data set. We compare results to state-of-the-art approaches for patient online monitoring and investigate similarities and differences in correlations among covariates across disease groups.

Related work

Related work falls into three areas of medical time series analysis: (i) incorporating noisy, heterogeneous, irregular, and sparsely sampled time series data; (ii) combining information across multiple time series; and (iii) exploiting reference data in addition to observations about the current patient to enable patient-specific predictions for a new hospital patient.

Most prior work has focused on modeling each clinical covariate separately. Due to the irregularity and temporal sparsity of medical data, conventional time series models, such as hidden Markov models (HMMs), autoregressive (AR) models, state-space models, and linear dynamical systems (LDS), are challenging to apply because of the assumption of regular measurement sampling in time. Recent work has focused on developing methods to compensate for the missing data in order to work with models that assume complete data. Methods such as kernel support vector machine (SVM), matrix factorization, and k-nearest neighbors (KNNs) were applied for missing data imputation to improve sepsis or septic shock prediction [10, 11]. In other work, a hierarchical switching LDS model was used to monitor the physiological signals during neonatal sepsis; the model allows the latent state of a patient to change during periods with fewer observations [12]. In an alternative approach, noisy and sparse time series data were smoothed temporally by putting Gaussian priors on the mean parameters of the Gaussian mixture model, which is related to a Gaussian process prior, although the distribution is over a finite-dimensional vector [13].

Gaussian processes (GPs) are useful approaches for time series analysis because they can naturally capture irregular time series observations and estimate prediction uncertainties in a probabilistic framework [14]. For these reasons, GPs have been applied to the analysis of medical time series data. Previous work used a single-output GP regression model to smooth and impute each

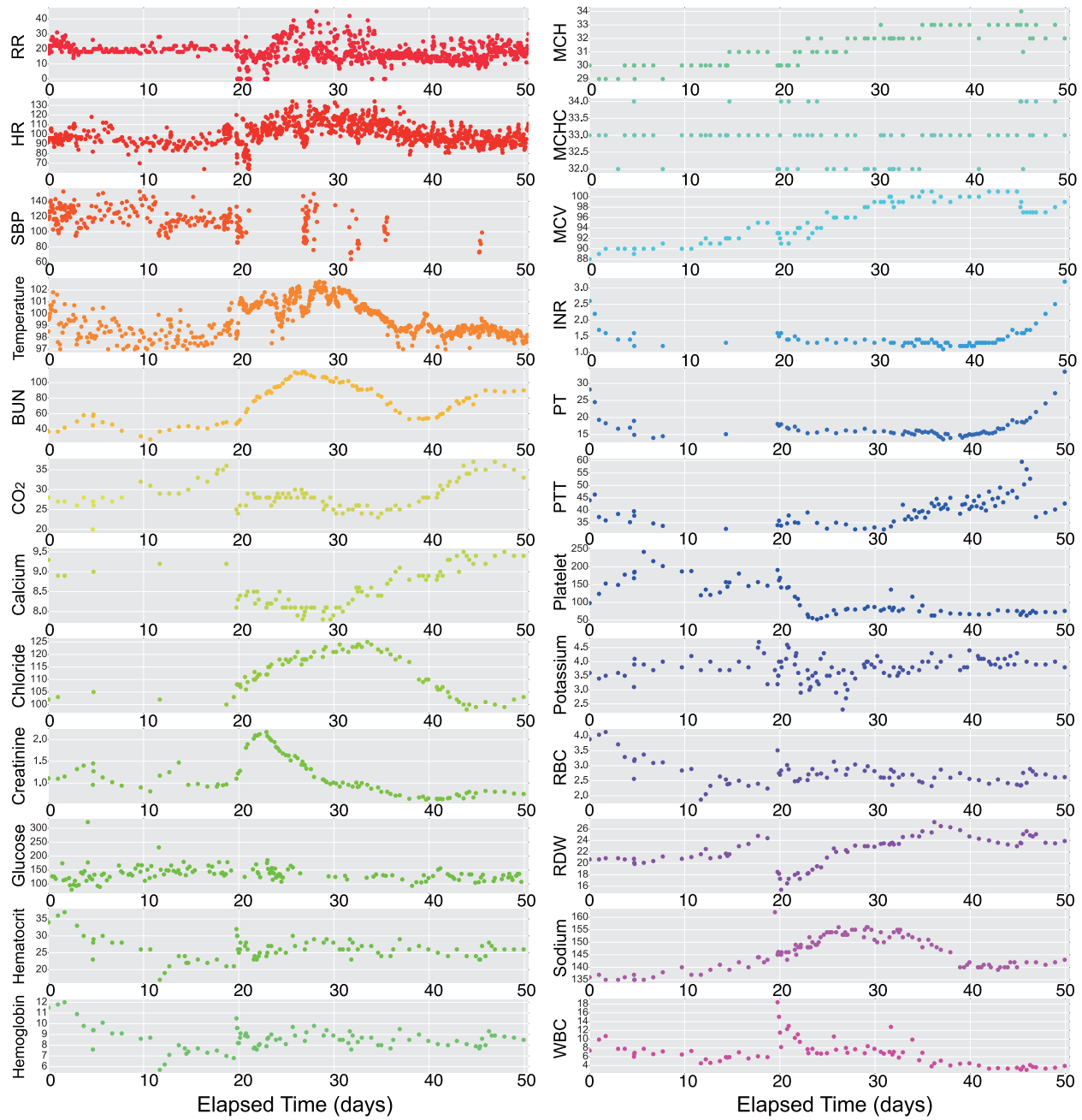

Fig. 1 An example of time series data of 24 clinical covariates for a septic patient in the HUP data. The 24 covariates include four vital signs—respiratory rate (RR), heart rate (HR), systolic blood pressure (SBP), body temperature—and 20 lab results. The time series are aligned by the patient's admission time. The density of sampling varies widely over the 24 covariates. A full description of these covariates can be found in Table 1

covariate independently [15, 16]. The Probabilistic Subtyping Model (PSM) added patient-specific information for smoothing temporal trajectories of clinical covariates and clustering disease subtypes [17]. PSM learns a mixture model based on a B-spline and GPs to impute the clinical measurements for patients with scleroderma. Demographic covariates, including gender, ethnicity, and clinical history, were also incorporated in the model. In an extension of PSM, the authors adapted patient-specific information to forecast specific clinical covariates [18]; the time series for each covariate was still modeled independently.

Table 1 The 24 clinical covariates modeled in MedGP

Type	Covariate	Sepsis	Neoplasms	Heart Failure	MIMIC-III
Vital	Respiratory rate (RR)	87,076	493,964	147,445	291,466
Vital	Heart rate (HR)	96,317	527,989	227,951	294,746
Vital	Systolic blood pressure (SBP)	84,909	447,666	104,129	124,587
Vital	Body temperature (Temp)	80,597	364,286	94,468	56,533
Lab	Blood urea nitrogen (BUN)	12,528	71,825	21,751	25,102
Lab	Carbon dioxide (CO_2)	12,672	72,784	21,844	20,979
Lab	Calcium level	10,388	66,051	18,867	20,568
Lab	Chloride	10,100	68,534	21,421	26,248
Lab	Creatinine	12,689	72,928	21,889	25,237
Lab	Glucose point-of-care (Glucose POC)	20,444	170,872	54,239	24,196
Lab	Hematocrit (Hct)	12,752	74,060	22,035	24,810
Lab	Hemoglobin (Hgb)	13,005	75,646	27,891	21,226
Lab	Mean cell hemoglobin (MCH)	12,587	69,736	18,379	20,877
Lab	Mean cell hemoglobin concentration (MCHC)	12,577	69,682	18,359	20,885
Lab	Mean cell volume (MCV)	12,587	69,751	18,380	20,875
Lab	International normalization ratio (INR)	5,733	38,810	17,005	15,735
Lab	Prothrombin time (PT)	5,722	38,844	17,007	15,734
Lab	Partial thromboplastin time (PTT)	5,872	41,894	19,596	17,185
Lab	Platelet	12,586	69,945	18,367	21,395
Lab	Potassium level	12,830	77,395	28,470	27,200
Lab	Red blood cell (RBC)	12,600	69,776	18,387	20,876
Lab	Red cell distribution width (RDW)	12,580	69,757	18,381	20,877
Lab	Sodium level	12,848	78,617	28,597	26,383
Lab	White blood cell (WBC)	12,581	69,950	18,384	20,960

This table includes the total number of observations for each covariate across patients in three disease groups—sepsis, neoplasms, and heart failure—in the HUP data, and the heart failure patients in the MIMIC-III data

The idea of capturing the joint dynamics between vital signs and lab tests has also been explored. Using high-frequency regularly sampled time series, the dynamics between heart rate (HR) and blood pressure (BP) were modeled using a mixture of an LDS model [19] and a switching vector autoregressive model (SVAR) [20]. The joint dynamics estimated across covariates were reported to be associated with hospital mortality. In other work [21], a multivariate spline-based approach with linear mixed effects was used to predict multiple longitudinal outcomes and time-to-death of patients. Time series graphical models (TGMs) [22, 23] have also been studied and applied for analyzing multivariate medical time series of ICU patients [24]. TGMs model the partial correlations between each dimension of the multivariate time series as an undirected graph. However, both TGMs and SVAR models follow the assumptions of vector autoregressive (VAR) models, and thus assume the sampling interval of the time series is fixed across dimensions. In practice, this means missing data imputation needs to be done in advance [23]. Coupled Latent Trajectory Model (C-LTM) [25], an extension of PSM, adapted conditional random fields (CRFs) to update the distribution of the target covariate from five other auxiliary covariates. While tackling the challenge of irregular sampling and jointly modeling multiple covariates, C-LTM is limited by requiring temporal alignment across patients, as in PSM.

Several multi-output GP frameworks have been proposed for other application areas. In the geostatistics literature, the linear model of coregionalization (LMC) characterizes correlations between outputs through a set of kernels and coregionalization matrices that estimate weights for pairwise outputs [26, 27]. In the machine learning literature, related models include multi-task GPs [28], semiparametric latent factor models [29], and multi-task kernel learning [30]. These can be viewed as variations of the LMC with different parameterizations and constraints. Convolution processes (CPs) have also been adapted to model multiple correlated outputs through

the convolution of smooth kernels and latent processes [31]. This approach usually has fewer hyperparameters and more efficient computation as compared to LMC, but only squared exponential (SE) kernels have been shown to be computationally tractable. Applying a multi-task GP (MTGP) framework [28] to clinical time series analysis has also been considered in two studies [32, 33]; both studies considered one patient as one task and used the remaining patients as reference training data. Other work adapted the LMC framework with one SE kernel to model three sparsely sampled clinical covariates (intracranial pressure, mean arterial blood pressure, and pressure-reactivity index) jointly [32]. The MTGP was shown to outperform a single-task GP (STGP) in prediction error. Both MTGP and CP have also been used with an SE kernel to model three densely sampled vital signs (respiratory rate, systolic blood pressure, and heart rate); both methods showed improvements as compared to a single-task GP [33].

Our work is distinct from previous research in several ways. First, we use the GP regression framework to model multiple irregularly sampled medical time series using a sparse structured multi-output kernel. In contrast to related work [32, 33], our kernel uses a mixture of flexible spectral kernels [34], allowing periodic behavior and both short-term and long-term dependencies within and across the clinical covariates over time. Second, we use the LMC framework to enable an interpretable quantification of cross-correlation and sparsity between covariates. Third, we model many more clinical covariates (24) compared with previous studies (at most six); in the online medical setting, efficient and scalable computation in this multi-view model is essential. To do this we use a sparse and low-rank formulation of the shared covariance matrix across clinical covariates to estimate and regularize the relationships between covariates in order to learn about covariate relationships specific to patient subgroups and to prevent overfitting.

In our methodology, MedGP, we trained a GP model on each reference patient separately, and used these models to estimate the empirical population-level model using nonparametric density estimation. This approach avoids training procedures that iterate through all reference patients, which is computationally intractable for an online system [32, 33]. To speed up training, we optimized the implementation in C++ using multithreading. Finally, in order to personalize the model for a new patient, we update the empirical population-level model on-the-fly to estimate patient specific parameters as measurements from the new patient are observed.

Methods

In this section, we describe our method, MedGP, for estimating the underlying dynamic processes jointly across a large number of sparsely sampled clinical covariates. We first describe the design of the Gaussian process kernel for capturing the temporal correlations within and between covariates. Next, we introduce the sparsity-inducing prior to regularize the LMC weight matrix. We then describe estimation of the parameters in the empirical prior and in the kernel. Next, we describe how to learn a patient-specific kernel by first building a population-level model from reference patients and then performing online updating of the parameters when observations about a new patient accumulate. Finally, we describe methods to perform computationally tractable online inference in these models, concluding with a discussion of computational complexity.

Gaussian processes

Gaussian processes (GPs) are distributions over arbitrary functions. By definition, a GP is a collection of random variables, any finite collection of which have a joint Gaussian distribution. Alternatively, a GP can be described as a distribution on an arbitrary function, defined as

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}')), \tag{1}$$

where $m(\mathbf{x})$ is the *mean function*:

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})], \tag{2}$$

and $\kappa(\mathbf{x}, \mathbf{x}')$ is the *covariance function* or *kernel*:

$$\kappa(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]. \tag{3}$$

Any finite number of function values jointly have a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{K}$ between any pair of observations, defined by the kernel function,

$$\begin{aligned} [f(x_1), f(x_2), \cdots, f(x_T)]^\top &\sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{K}), \\ \boldsymbol{\mu} &= [m(x_1), m(x_2), \cdots, m(x_T)]^\top, \\ \mathbf{K}_{i,j} &= \kappa(x_i, x_j). \end{aligned} \tag{4}$$

Properties of the function $f(\mathbf{x})$ such as smoothness or periodicity are determined by the kernel function $\kappa(\mathbf{x}, \mathbf{x}')$. One of the most commonly used kernels is the squared exponential (SE) kernel

$$\kappa(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\ell^2}\right), \tag{5}$$

which is parameterized by a length scale ℓ and a scale factor σ. The functions generated by a GP with an SE kernel are smooth because the kernel function is infinitely differentiable [35]. The value of the length scale ℓ determines the distribution of changes over the function value with respect to changes in the input $\mathbf{x}$, encouraging a specific smoothness. Due to its simplicity, SE is used in many applications; however, the properties of the functions that

it captures are fairly limited. Periodic functions, for example, are not well modeled by an SE kernel, but instead captured by a periodic kernel

$$\kappa(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left[-\frac{4\sin^2\left(\frac{\pi||\mathbf{x}-\mathbf{x}'||}{p}\right)}{\ell^2}\right], \tag{6}$$

where p is the period of the function. When modeling medical time series, the SE kernel or the periodic kernel are often used in combination to capture the unknown source-specific smoothness and periodicity of the trajectories of clinical covariates [15, 33].

Gaussian process regression with a structured multi-output kernel

Our first goal is to jointly model multiple clinical covariates—vital signs and lab tests—over time for each patient using GP regression. For the ith patient, we denote the time series of the dth covariate as a vector $\mathbf{x}_{i,d}$, representing the time points that the dth covariate was observed, and the corresponding observation vector $\mathbf{y}_{i,d}$:

$$\mathbf{x}_{i,d}^\top = \left[x_{i,d,1}, x_{i,d,2}, \ldots x_{i,d,t} \ldots, x_{i,d,T_{i,d}}\right], \tag{7}$$

$$\mathbf{y}_{i,d}^\top = \left[y_{i,d,1}, y_{i,d,2}, \ldots y_{i,d,t} \ldots, y_{i,d,T_{i,d}}\right], \tag{8}$$

where t indexes time, and $T_{i,d}$ is the total number of observations for the dth covariate of the ith patient.

To represent the time series data over all D covariates, we define the flattened data,

$$\mathbf{x}_i^\top = \left[\mathbf{x}_{i,1}^\top, \mathbf{x}_{i,2}^\top, \ldots, \mathbf{x}_{i,D}^\top\right], \tag{9}$$

$$\mathbf{y}_i^\top = \left[\mathbf{y}_{i,1}^\top, \mathbf{y}_{i,2}^\top, \ldots, \mathbf{y}_{i,D}^\top\right], \tag{10}$$

where $\mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^{T_i \times 1}$, $T_i = \left(\sum_{d=1}^{D} T_{i,d}\right)$. Let $\mathcal{F}_i$ be a multi-output function over time for the ith patient. We capture the relationship between time and clinical observations as a GP regression model:

$$\mathbf{y}_i = \mathcal{F}_i(\mathbf{x}_i) + \boldsymbol{\epsilon}_i, \tag{11}$$

where $\boldsymbol{\epsilon}_i$ is the residual noise vector. Marginally at the tth observation of the dth covariate, the residual noise is modeled as

$$\epsilon_{i,d,t} \sim \mathcal{N}\left(0, \sigma_{i,d}^2\right), \tag{12}$$

where $\sigma_{i,d}^2$ is the covariate-specific residual variance for each individual.

We assume that the function $\mathcal{F}_i$ is drawn from a patient-specific Gaussian process $\mathcal{GP}_i$ with mean function $\mu_i(\mathbf{x})$ and kernel $\kappa_i(\mathbf{x}, \mathbf{x}')$:

$$\mathcal{F}_i \sim \mathcal{GP}_i(\mu_i(\mathbf{x}), \kappa_i(\mathbf{x}, \mathbf{x}')). \tag{13}$$

We set $\mu_i(\mathbf{x}) = \mathbf{0}$ [35].

We designed the kernel $\kappa_i(\mathbf{x}, \mathbf{x}')$ to capture predictive and generalizable covariance structure across medical time series data. Assuming the covariates are correlated across time, we adapted the linear model of coregionalization (LMC) framework [26, 27]. We used a set of Q *basis kernels* $\{\kappa_q(\mathbf{x}, \mathbf{x}')\}_{q=1}^Q$ to model D covariates jointly. The kernel for the cross-covariance of any pair of covariate types is modeled by a weighted structured linear mixture of the Q basis kernels. The full joint kernel is written as a block structured function

$$\kappa_i(\mathbf{x}_i, \mathbf{x}_i') = \sum_{q=1}^{Q} \begin{bmatrix} b_{q,(1,1)}\kappa_q(\mathbf{x}_{i,1}, \mathbf{x}_{i,1}') & \cdots & b_{q,(1,D)}\kappa_q(\mathbf{x}_{i,1}, \mathbf{x}_{i,D}') \\ b_{q,(2,1)}\kappa_q(\mathbf{x}_{i,2}, \mathbf{x}_{i,1}') & \cdots & \vdots \\ \vdots & \ddots & \vdots \\ b_{q,(D,1)}\kappa_q(\mathbf{x}_{i,D}, \mathbf{x}_{i,1}') & \cdots & b_{q,(D,D)}\kappa_q(\mathbf{x}_{i,D}, \mathbf{x}_{i,D}') \end{bmatrix}, \tag{14}$$

where $b_{q,(d,d')}$ scales the covariance (defined by the qth basis kernel) between covariates d and d', and $\kappa_i(\mathbf{x}_i, \mathbf{x}_i) \in \mathbb{R}^{T_i \times T_i}$. We collapsed $b_{q,(d,d')}$ into a set of weight matrices $\{\mathbf{B}_q\}_{q=1}^Q$, where each $\mathbf{B}_q$ is a symmetric positive definite matrix

$$\mathbf{B}_q = \begin{bmatrix} b_{q,(1,1)} & b_{q,(1,2)} & \cdots & b_{q,(1,D)} \\ b_{q,(2,1)} & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ b_{q,(D,1)} & b_{q,(D,2)} & \cdots & b_{q,(D,D)} \end{bmatrix} \in \mathbb{R}^{D \times D}. \tag{15}$$

If the inputs, observation times are the same for all covariates, we can further simplify Eq. (14) with the Kronecker product $\otimes$. That is, if $\mathbf{x}_{i,1} = \mathbf{x}_{i,2} = \cdots = \mathbf{x}_{i,D} \triangleq \mathbf{x}_{i,*}$ and $\mathbf{x}_{i,1}' = \mathbf{x}_{i,2}' = \cdots = \mathbf{x}_{i,D}' \triangleq \mathbf{x}_{i,*}'$:

$$\kappa_i(\mathbf{x}_i, \mathbf{x}_i') = \sum_{q=1}^{Q} \mathbf{B}_q \otimes \kappa_q(\mathbf{x}_{i,*}, \mathbf{x}_{i,*}'), \tag{16}$$

although in practice we do not often see this situation in medical time series data. For simplicity, we only use the index when date come from different individual.

Properties of the time series observations, such as periodicity and short term dependencies, are captured in the Q basis kernels. For medical covariates, the properties and patterns of each patent's time series observations may vary. As a trivial example, when a patient is under age 18, their pulse will be well correlated with their age, height, and weight; above age 18, the correlation among pulse, age, height, and weight is more variable within age than across ages. Furthermore, only a few vital signs, such as heart rate, blood pressure, and body temperature, are known to be periodic with a 24-h period (i.e., a circadian rhythm), but whether there is a similar period for specific lab results, such as white blood cell counts or pressure of carbon dioxide in the blood, is unclear [36].

To handle the heterogeneity of patterns within covariates and across patients, we selected the spectral mixture (SM) kernel as the basis kernel [34]. The SM kernel is a general form of a variety of stationary kernels, including the squared exponential (SE) kernel and the periodic

kernel, and has also shown good performance in modeling processes generated from more complex kernels through a mixture of kernels approach [34]. The basis kernel $\kappa_q(x_t, x_{t'})$ is written as

$$\kappa_q(x_t, x_{t'}) = \exp\left(-2\pi^2\rho^2 v_q\right)\cos\left(2\pi\rho\mu_q\right), \tag{17}$$

where $\rho = |x_t - x_{t'}|$ is the absolute distance in time. In our work, the mixture weights for each basis kernel are encoded in $\mathbf{B}_q$.

To be used for GP regression, $\kappa_i(\mathbf{x}, \mathbf{x}')$ must be a valid Mercer kernel, i.e., the Gram matrix must be positive definite for all $\mathbf{x}$ and $\mathbf{x}'$. Since the matrix produced by each basis kernel $\kappa_q(\mathbf{x}, \mathbf{x}')$ is symmetric positive definite, we only need to ensure that every $\mathbf{B}_q$ is positive definite to produce a Mercer kernel. To do this, we parameterized $\mathbf{B}_q$ as

$$\begin{aligned}\mathbf{B}_q &= \mathbf{A}_q\mathbf{A}_q^\top + \begin{bmatrix}\lambda_{q,1} & 0 & \cdots & 0\\ 0 & \lambda_{q,2} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \lambda_{q,D}\end{bmatrix} \\ &= \mathbf{A}_q\mathbf{A}_q^\top + \mathrm{diag}(\lambda_q),\end{aligned} \tag{18}$$

$$\mathbf{A}_q = \begin{bmatrix}a_{q,(1,1)} & \cdots & a_{q,(1,R_q)}\\ \vdots & \ddots & \vdots\\ a_{q,(D,1)} & \cdots & a_{q,(D,R_q)}.\end{bmatrix} \tag{19}$$

Here $\mathbf{A}_q \in \mathbb{R}^{D\times R_q}$, $\lambda_q \in \mathbb{R}^{D\times 1}$. We let R_q denote the number of non-zero columns in $\mathbf{A}_q$, or the rank for $\mathbf{B}_q$ when $\lambda_q = \mathbf{0}$.

For any two observations from the same patient of different covariates at different times, denoted as $x_{d,t}$ and $x_{d',t'}$, the prior covariance from the GP kernel is

$$\kappa(x_{d,t}, x_{d',t'}) = \sum_{q=1}^{Q} b_{q,(d,d')}\kappa_q(x_t, x_{t'}). \tag{20}$$

We summarize the parameters and hyperparameters of our SM-LMC kernel in Table 2.

Sparsity-inducing priors on weight matrix B_q

As the number of medical covariates included in the model increases, we need to increase the number of basis kernels Q and corresponding R_q to allow greater representational flexibility. However, too many basis kernels may lead to overfitting and will become computationally intractable. To avoid this, we regularized the elements of each weight matrix $\mathbf{B}_q$ by introducing structured sparsity-inducing priors on each $\mathbf{A}_q$ matrix as follows.

We included two layers of sparsity-inducing priors for flexible, data-adaptive shrinkage behavior, modified from previous work [37, 38]. First, we put column-wise sparsity-inducing priors to regularize each column in $\mathbf{A}_q$. This corresponds to regularizing the degrees of freedom of the functions, or number of latent processes generated from each basis kernel in the LMC model [39]. Second, we put sparsity-inducing priors on each matrix element $a_{q,(d,r)}$ in $\mathbf{A}_q$ to produce element-wise sparsity. The effect of element-wise sparsity is to perform model selection on the number of basis kernels that each pair of covariates uses for covariance representation. Finally, we put sparsity-inducing priors on the elements of λ_q to shrink the covariance for observations from the same covariate.

In practice, we implemented each layer of the prior as a two-layer hierarchical gamma distribution. The generative model is written as

$$\begin{aligned}\tau_{q,(r)} &\sim \mathrm{Gamma}(\xi, \eta),\\ \phi_{q,(r)} &\sim \mathrm{Gamma}(\gamma, \tau_{q,(r)}),\\ \delta_{q,(d,r)} &\sim \mathrm{Gamma}(\beta, \phi_{q,(r)}),\\ \psi_{q,(d,r)} &\sim \mathrm{Gamma}(\alpha, \delta_{q,(d,r)}),\\ a_{q,(d,r)} &\sim \mathcal{N}(0, \psi_{q,(d,r)}),\end{aligned} \tag{21}$$

where each element $a_{q,(d,r)}$ has a Gaussian distribution. Parameters $\phi_{q,(r)}$ and $\tau_{q,(r)}$ control the column-specific shrinkage, while parameters $\psi_{q,(d,r)}$ and $\delta_{q,(d,r)}$ control the local shrinkage of each element in the $\mathbf{A}_q$ matrix. For vector λ_q, we regularized each element with a local Laplace prior:

$$\lambda_{q,(d)} \sim \mathrm{Laplace}(0, \beta_\lambda). \tag{22}$$

For our results, we set $\alpha = \beta = \gamma = \xi = 0.5$ to recapitulate two layers of the horseshoe prior, using a statistically equivalent prior represented by a hierarchical gamma with four layers [38, 40–42]. Parameters $\psi_{q,(d,r)}$, $\delta_{q,(d,r)}$, $\phi_{q,(r)}$, and $\tau_{q,(r)}$ were estimated during optimization. We set $\beta_\lambda = 0.01$ to regularize the diagonal terms $\lambda_{q,(d)}$. The hyperparameter η controls the overall shrinkage profile of the hierarchical gamma prior (see Additional file 1: Appendix A for more details). We chose η over $\{0.01, 0.1, 1.0\}$ using cross-validation prediction error.

Table 2 The list of hyperparameters for modeling the $d = 1 : D$ clinical variables and $q = 1 : Q$ mixture kernels

Notation	Size	Description
v_q	Q	Squared exponential part of qth basis kernel
μ_q	Q	Periodicity of qth basis kernel
$a_{q,(d,r)}$	$\sum_{q=1}^{Q} D \times R_q$	Weights of (d, d') for qth basis kernel
$\lambda_{q,}$	D	Intra-covariate weights of the dth covariate for qth basis kernel

Parameter learning

To estimate the parameters for the regularized kernel, we optimized the posterior probability. We denote all

parameters that were estimated directly as $\boldsymbol{\theta}$ and hyperparameters in the sparsity-inducing prior as $\boldsymbol{\theta}_f$:

$$\begin{aligned}\boldsymbol{\theta} = \big\{&\mu_q, \nu_q, a_{q,(d,r)}, \lambda_{q,(d)},\\ &\psi_{q,(d,r)}, \delta_{q,(d,r)}, \phi_{q,(r)}, \tau_{q,(r)}\big\},\end{aligned} \tag{23}$$

$$\text{for } q = 1, \cdots, Q; d = 1, \cdots, D; r = 1, \cdots, R_q$$

$$\begin{aligned}&\boldsymbol{\theta}_f = \{\alpha, \beta, \gamma, \xi, \eta, \beta_\lambda\},\\ &\alpha = \beta = \gamma = \xi = 0.5.\end{aligned} \tag{24}$$

The posterior density of our model is then

$$\begin{aligned}&p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}_f) \propto p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\theta}_f)\\ &\propto p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) \Big[\prod_{q=1}^{Q} \prod_{d=1}^{D} \prod_{r=1}^{R_q} p(a_{q,(d,r)}|\psi_{q,(d,r)})\\ &p(\psi_{q,(d,r)}|\alpha, \delta_{q,(d,r)}) p(\delta_{q,(d,r)}|\beta, \phi_{q,(r)})\Big]\\ &\times \Big[\prod_{q=1}^{Q} \prod_{r=1}^{R_q} p(\phi_{q,(r)}|\gamma, \tau_{q,(r)}) p(\tau_{q,(r)}|\xi, \eta)\Big]\\ &\times \Big[\prod_{q=1}^{Q} \prod_{d=1}^{D} p(\lambda_{q,(d)}|\beta_\lambda)\Big] \Big[\prod_{q=1}^{Q} p(\nu_q) p(\mu_q)\Big].\end{aligned} \tag{25}$$

The term $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ is found by calculating the GP marginal likelihood given the values of $\boldsymbol{\theta}$ [35], which is

$$\begin{aligned}\log p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = &-\frac{1}{2}\mathbf{y}^\top (K_{|\boldsymbol{\theta}} + \boldsymbol{\epsilon} I)^{-1}\mathbf{y}\\ &-\frac{1}{2}\log |K_{|\boldsymbol{\theta}} + \epsilon I|\\ &-\left(\frac{\sum_{d=1}^{D} T_{i,d}}{2}\right)\log(2\pi).\end{aligned} \tag{26}$$

We use $K_{|\boldsymbol{\theta}}$ to denote the covariance matrix given $\boldsymbol{\theta}$.

We thus estimated $\boldsymbol{\theta}$ by solving the posterior optimization problem, for $\mathcal{Q}(\boldsymbol{\theta}) = \log p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}_f) = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta})$.

See equations in Additional file 1: Appendix B for the derivation of $\mathcal{Q}(\boldsymbol{\theta})$.

Due to the conjugacy of the hierarchical gamma priors, we optimized parameters $\psi_{q,(d,r)}$, $\delta_{q,(d,r)}$, $\phi_{q,(r)}$, $\tau_{q,(r)}$ directly using maximum a posteriori (MAP) estimates of their posterior distribution (or mean when the mode does not exist). Our optimization procedure then consists of two parts. In the first part, we used the update equations to estimate $\psi_{q,(d,r)}$, $\delta_{q,(d,r)}$, $\phi_{q,(r)}$, and $\tau_{q,(r)}$ conditional on current estimates of $\hat{\mu}_q$, $\hat{\nu}_q$, $\hat{a}_{q,(d,r)}$, and $\hat{\lambda}_{q,(d)}$ directly (details can be found in Additional file 1: Appendix B). In the second part, we estimated parameters μ_q, ν_q, $a_{q,(d,r)}$, and $\lambda_{q,(d)}$ using a scaled conjugate gradient method to find the local maximum, conditioned on current estimates of $\hat{\psi}_{q,(d,r)}$, $\hat{\delta}_{q,(d,r)}$, $\hat{\phi}_{q,(r)}$, and $\hat{\tau}_{q,(r)}$. We iterated over the two steps until the change in $\mathcal{Q}(\boldsymbol{\theta})$ reached the convergence criterion (< 0.005) or until the maximum number of iterations (≥ 30).

Estimating the population-level model and online updating

The GP with the structured kernel described above lets us model the patient-specific joint dynamics between covariates within the same patient. We now describe how we built a populatio n-level empirical prior from a set of mixture kernels estimated from all training patients, and how we apply this empirical prior to a new patient.

To estimate the empirical priors across training patients, we trained one GP kernel for each patient separately, and then we clustered and extracted the estimates of the basis kernels (defined by hyperparameters μ_q and ν_q). The idea here is that, when we estimate a set of patient-specific mixture kernels, we would like to understand the high-level properties of these mixture kernels shared across patients in the same patient group. Then, we can estimate the group-specific distributions of the hyperparameters through in the estimates of basis kernels belonging to each cluster. For instance, a circadian rhythm (24-h periodicity) may be observed in some covariates for some patient, groups but the period across patients could vary within a range. Across the space of μ and ν, the spectral kernels vary substantially (Fig. 2a). For each basis kernel that was estimated, the characteristic period is $1/\mu_q$ and the length scale is $1/2\pi\sqrt{\nu_q}$ [34]. There are different ways to define the features of a kernel. Here, we used the temporal features of the learned kernels directly (Fig. 2b). The temporal spacing of two adjacent points is one hour, and we use kernel values within a 72 h window. We then used a Gaussian mixture model (GMM) to perform clustering on the kernels, estimated across patients and we chose the best number of kernel clusters Q' ($1 \leq Q' \leq Q$) based on Bayesian information criterion (BIC). For the MedGP implementation, we adapted the open source scikit-learn package [43]. We used version 0.18.1, with ten random restarts, a maximum of 2,000 iterations, and allowing each mixture component to have its own covariance matrix.

For each identified kernel cluster, we estimated one set of parameters μ_q and ν_q for the basis kernel, and the weight coefficients—elements in $\mathbf{B}_q$ matrices, computed using the $\mathbf{A}_q$ matrices and $\boldsymbol{\lambda}_q$ vectors. We do this by building an empirical distribution using kernel density estimation (KDE) with a Gaussian kernel over the GP kernel hyperparameters assigned to that cluster. The bandwidth of the kernel density estimator was chosen based on Silverman's "rule-of-thumb" [44]. We estimated each new parameter using density-weighted means with the density from the univariate KDE as the weights. When there were multiple kernels in a patient cluster, the estimated $\mathbf{B}_q$ matrices were added based on the additive assumption of our kernel before aggregating to estimate the population-level kernel for that cluster. To allow online updating, we estimated the elements of the new empirical $\mathbf{A}_q$ matrix and $\boldsymbol{\lambda}_q$ vector corresponding to each new $\mathbf{B}_q$

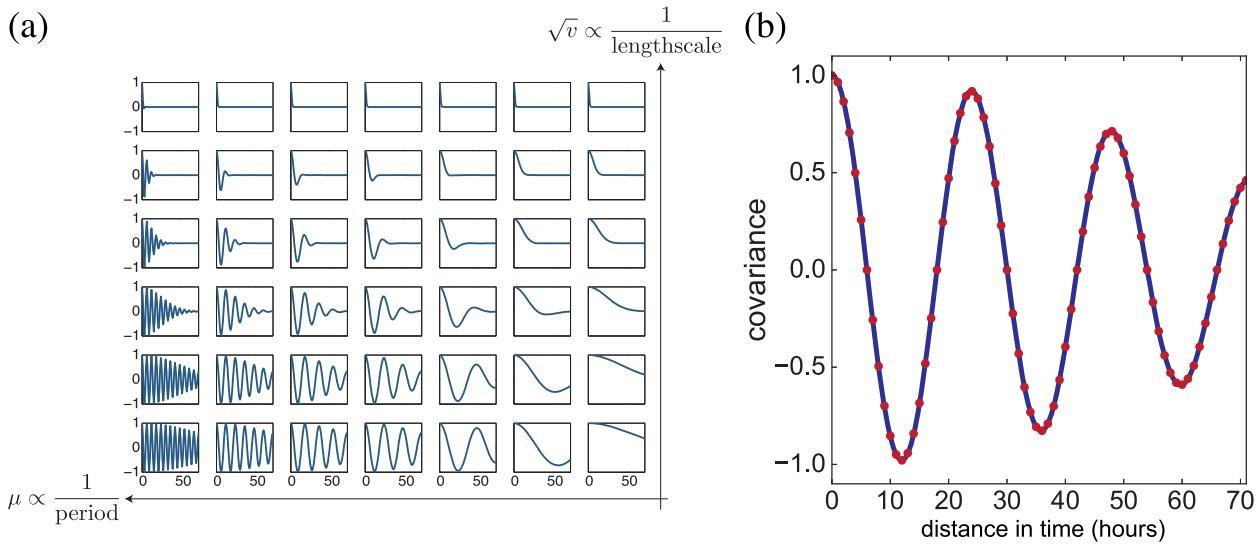

Fig. 2 Illustrations of the basis kernels and the features for kernel clustering. (**a**) An example of a discrete set of basis kernels with different μ and v within a 72-h window. (**b**) An example of the 72-dim temporal features (shown with red dots) taken from a kernel for GMM clustering

matrix using singular value decomposition (SVD). For the univariate GP regression, we did not use density weighted means because we found them to be unstable; instead we used a grid-based search to identify the hyperparameters with the highest posterior probality with respect to the kernel density estimates.

As the number of vital signs and lab measurements for a new patient accumulated, we update the hyperparameters to estimate a patient-specific kernel. Indeed, we update the kernel sequentially every time a new observation arrives. To do this in a computationally tractable way, we used the momentum method [45] with almost a 72-h window of previous observations to update the kernel hyperparameters when predicting the value of next observation. For all experiments, we chose the momentum as 0.9 and the learning rate as 10^{-5}. For elements in the $\mathbf{A}_q$ matrices, we do not update the values if the elements were set to near zero in the empirical prior so as to maintain the empirical sparsity structure.

Efficient inference in MedGP

The main bottleneck of our method is in learning patient-specific kernel hyperparameters. Let $T_i = \sum_{d=1}^{D} T_{i,d}$ denote the total number of samples of the ith patient; the computational cost to compute the Gram matrix is $\mathcal{O}\left(QT_i^2\right)$, which increases linearly with the chosen number of basis kernels. To find the MAP estimates of the parameters, we need to invert and compute the determinant of the Gram matrix $(K_{|\boldsymbol{\theta}} + \boldsymbol{\epsilon} I)$ in Eq. (26). The computational complexity for the full matrix inversion is $\mathcal{O}(T_i^3)$ using Cholesky decomposition. When calculating the gradients for optimizing the hyperparameters, the cost is dominated by $\mathcal{O}(QDRT_i^2)$ after the inverse Gram matrix is pre-computed, which is linear with the total number of the kernel hyperparameters. In practice, the complexity of each iteration is either $\mathcal{O}\left(T_i^3\right)$ or $\mathcal{O}(QDRT_i^2)$. That is, the patient with the most measurements is the main bottleneck for training. In our implementation, we mitigate the bottleneck using optimized linear algebra functions in Intel MKL library with multithreading and computing the gradients of the hyperparameters in parallel.

Results

We analyzed the performance of the method, multi-output GP with a sparse SM-LMC kernel and online updating, MedGP by applying it to time series data from the Hospital of the University of Pennsylvania (HUP) and the public MIMIC-III data set [4]. We first introduce the HUP and MIMIC-III data and preprocessing procedures, and then we show experimental results and comparisons with baseline and state-of-the-art methods for online monitoring of time-series data with correlated clinical covariates.

Medical data preprocessing

The HUP medical time-series data consist of electronic health records (EHRs) from more than 260,000 patients admitted to a University of Pennsylvania Hospital. For each patient, the data include many heterogeneous clinical covariates, including ICD-9 codes, patient demography, length-of-stay, vital signs, and lab results. We jointly modeled the 24 covariates with the greatest number of observations across patients (Table 1). We selected three

groups of discharged patients from these data: 1365 septic patients, 952 patients with heart failure, and 4723 patients with neoplasms. Each patient has at least one observation for each of the 24 covariates, and in total over four million observations were evaluated.

For each clinical covariate, we first removed obvious artifacts (e.g., values outside of the possible range in living humans). For the patients with neoplasms or heart failure, we used the full patient length-of-stay in training and testing. For septic patients, the disease progression varies substantially across patients, and the distribution of the covariates changes dramatically depending on the disease phase. To address this issue, we segmented the time series data into four disjoint partitions based on clinical status: *no sepsis*, *pre-sepsis*, *sepsis*, and *recovery*. To label each stage, we incorporated prior clinical domain knowledge. For instance, we identified sepsis stages using ICD-9 codes and positive blood culture results. Since our model assumes stationarity, to better estimate the temporal correlation across covariates, we chose the *recovery* stage before the patients' discharge to test our method, since this is a relatively stable stage. We used the bed unit information to identify if the patient is in a stable state. That is, when a patient is transferred to step-down bed, we labeled the time series after the transfer as *recovery*. The median length-of-stay after preprocessing is 140 h for the sepsis group, 285 h for the heart failure group, and 197 h for the neoplasms group.

We applied similar preprocessing procedures to the MIMIC-III data. We selected patients with a heart failure diagnosis that eventually had a routine discharge. We removed artifacts such as out-of-bounds values for each covariate, and applied the criteria to each patient that at least five measurements were taken for all 24 selected covariates. We extracted 1004 heart failure under these criteria and used 1003 of them, excluding one patient with more than 50K measurements due to memory constraints.

Experimental setup

We applied MedGP to the three selected groups of patients separately, and evaluated characteristics and performance of MedGP under two different experimental settings. In the first analysis, we evaluated the model's ability to learn the covariance between a pair of highly correlated clinical covariates, and we measured the imputation performance in an online setting. In the second analysis, we follow the same online setting, but instead jointly model all 24 clinical covariates, including four vital signs and 20 lab covariates. In both settings, we evaluated our method using 10-fold cross-validation at the patient level. That is, for each fold we ran the kernel clustering step on the kernels from the training patients to estimate a set of population-level basis kernels and $\mathbf{B}_q$ matrices. This set of kernels was then applied to the held-out patients to predict the value of each covariate using observations from all other covariates measured at the same time as, or earlier than, the test observation (i.e., no future information included). After each prediction, we updated the patient-specific kernel parameters using the new observations from the test patient.

We compared our method to several univariate methods that modeled each covariate separately: (i) a naive one-lag prediction procedure, which predicts an observation equal to the last observation available from the same patient; (ii) an independent GP with squared exponential (SE) or spectral mixture (SM) kernels fitting each covariate separately (we tested with $Q = 1$ for SM); (iii) the multi-resolution Probability Subtyping Model (PSM) combining linear regression, B-splines, and independent GPs [17]. To estimate the spectral kernel parameters, for each patient we initialized 1000 random kernels by drawing uniformly from a length scale range (between 6 and 72 h) and period range (between 24 and 72 h). We computed the marginal likelihood of all random kernels for each patient, and then initialize optimization using the kernels with the highest marginal likelihood. The elements in the $\mathbf{A}_q$ matrices are initialized randomly between -1.5 and 1.5.

We compared results from MedGP to these various methods using two metrics: (i) mean absolute error (MAE) of the predicted observations with the true observations, and (ii) 95% coverage, the percentage of true observations that fell within the predictive 95% confidence region. We quantified and reported the improvements with respect to both metrics compared to all three baselines (naive prediction, univariate GP, and PSM). To test if the differences in prediction results from different approaches were statistically significant, we performed paired t-tests for the results of each covariate and compared the p-values with a Bonferroni corrected threshold (dependent on the number of jointly modeled covariates in each experiment).

We note that the original PSM was designed to model scleroderma disease [17]. Thus, to make it applicable to our different patient groups, several adjustments were made. First, we omitted the population and environmental factors selected for their relevance to scleroderma. Second, we chose the knots of the B-spline basis by sampling every hour for vital signs and every 24 h for lab results between zero and the longest length-of-stay for patients in each disease group. Third, to make PSM training feasible on the scale of our data set, we limited the maximum number of subtypes to ten for the sepsis and heart failure groups, and 20 for the neoplasms group.

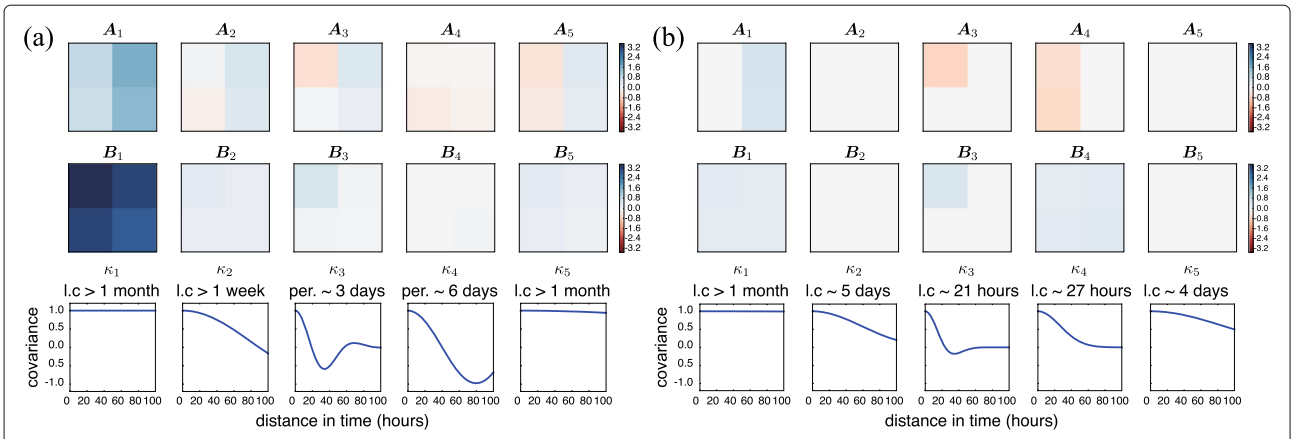

Fig. 3 The trained kernel for one patient jointly modeling PT and INR. For both the (**a**) SM-LMC kernel and (**b**) sparse SM-LMC kernel, the $\mathbf{A}_q$ matrices (upper row), $\mathbf{B}_q$ matrices (middle row), and the basis kernel κ_q (bottom row) are illustrated. The zero elements are colored in light grey. Here *l.c* denotes length scale for each basis kernel, and *per.* denotes period. The length-of-stay for this patient was over 90 days

Results of two lab covariates

As a proof of principle, we jointly modeled two well correlated lab covariates, prothrombin time (PT) and international normalization ratio (INR), on three HUP subgroups. PT measures the time it takes for the plasma in the blood to clot, and is often ordered to check bleeding problems. INR is an international standard for PT to account for possible variations across different labs. For the same patient, the two covariates usually have similar trajectories over time (Fig. 1).

We trained the kernels for one patient's INR and PT time series data both with and without the structured sparse prior (Fig. 3). Both $\mathbf{A}_q$ and $\mathbf{B}_q$ matrices estimated using the sparse prior have higher levels of sparsity versus those estimated without using the sparse prior. We observed that, for both methods, one of the estimated basis kernels κ_1 captures long-term (around one month) dependencies. However, with the sparse prior, the estimated weights associated with this long term kernel $\mathbf{A}_1$ are rank one instead of rank two. This means the trajectories of the two covariates are similar enough to be explained by one instead of two functions, and thus fewer hyperparameters. Moreover, two basis kernels were found with zeros weights $\mathbf{A}_2$ and $\mathbf{A}_5$ (Fig. 3b), suggesting that the prespecified number of basis kernels may be reduced. We also found that the off-diagonal elements in the $\mathbf{B}_q$ matrices in both cases have nonzero values, suggesting a nonzero covariance between PT and INR observations. In particular, two basis kernels captured the covariance between PT and INR: one with a greater than one-month trend (Fig. 3b, $\mathbf{B}_1$ and κ_1), and one with a 27-h trend (Fig. 3b, $\mathbf{B}_4$ and κ_4). Here, the sparse kernel has 18 nonzero hyperparameters, whereas there are 40 for the non-sparse kernel. We can compare the two fitted kernels using both log marginal likelihoods and model selection scores. The log marginal likelihoods of the two kernels are -118.16 (SM-LMC) and -128.50 (sparse SM-LMC), indicating a better fit for the SM-LMC model without sparsity. However, the Bayesian information criterion (BIC) values, which take into account the number of parameters in a model, were 353.63 (SM-LMC) and 309.79 (sparse SM-LMC), where values closer to zero reflect better models. Thus, using a sparse prior has the advantage of a expressive but more compact kernel representation.

We then ran our model on all three disease groups separately, and compared our method with the univariate baselines under the scenario of online imputation of the same two well-correlated clinical covariates. For independent GPs, we used gradient descent to optimize the hyperparameters. For PSM, we performed grid search for the parameters of the B-spline and the independent GP kernel. For our method, we set $Q = 5$ and $R_q = 2$ for the $\mathbf{A}_q$ matrices for training. In the sepsis and heart failure groups, three nonzero basis kernel functions ($Q' = 3$) were found for the model using the SM-LMC kernel, while only two non-zero basis kernel functions ($Q' = 2$) were found using the sparse SM-LMC kernel; the number of non-zero hyperparameters were 18 and 12 respectively. In the neoplasms group, the number of nonzero basis kernels were the same as the pre-specified number ($Q' = Q = 5$). With 10-fold cross-validation, we found that results using the SM-LMC kernel showed smaller imputation error than those using the baselines for both PT and INR (Fig. 4). The mean absolute errors (MAEs) showed that the non-sparse SM-LMC kernels perform imputation the best among the related approaches. On the other hand, looking at the 95% coverage, results using non-sparse or sparse SM-LMC kernel were well calibrated with respect

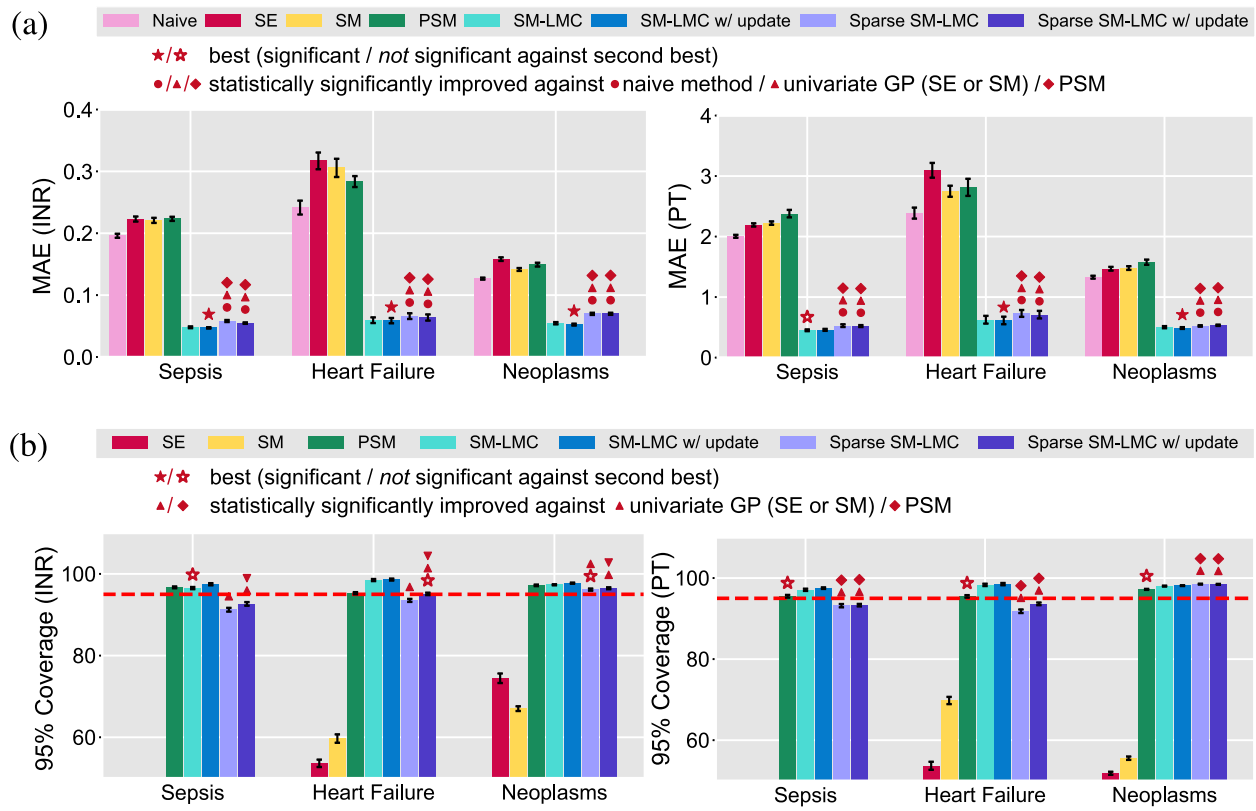

Fig. 4 The results of prediction when jointly modeling INR and PT. The figure illustrates (**a**) mean absolute error (MAE), and (**b**) 95% coverage (the dashed red line indicates 95%). The error bars denote ±1 standard error

to the confidence region compared with independent GPs, although sometimes slightly worse than PSM. Note that in this experiment we used a p-value threshold $p < 0.005$ to detect statistical significance, which reflects the Bonferroni correction. The results indicate that the sparse prior finds models with sparse structure while maintaining predictive performance in this two covariate case.

Results of a joint model including 24 vital signs and lab covariates

In the second experimental setting, we jointly modeled 24 vital signs and clinical covariates ($D = 24$) for all three disease groups (Table 1). We set the number of basis kernels $Q = 5$ and the number of nonzero columns in $\mathbf{A}_q$ as $R_q = 8$ in this experiment for the three HUP subsets. For the MIMIC-III heart failure subset, we set $Q = 4$. Detailed results of the best setup as well as the results for different Q may be found in Additional file 1: Appendix C and Appendix D.

Estimating population-level kernels

We first visualized the population-level kernels estimated from the three patient groups of the HUP data (Figs. 5, 6 and 7) and the MIMIC-III patient subgroup (Fig. 8). We observed shared patterns in the basis kernels κ_q and the weight matrices $\mathbf{B}_q$ across all patient groups. Comparing the estimated population-level kernels, we found at least one long-term smoothing basis kernel with length scale longer than three days, and one 24- to 25-h periodic basis kernel, which indicates the existence of circadian rhythms in specific covariates as expected. Furthermore, in the neoplasms group, which consists of more patients than the other two groups, we found additional short-term smoothing basis kernels and one 12- to 13-h periodic basis kernel, which may correspond to known circasemidian rhythm of clinical covariates, such as body temperature. We also observed an 11-h periodic kernel in the MIMIC-III subset.

In addition to the characteristics of the basis kernels, our model with the sparse prior also showed interpretable cross-covariate patterns (Figs. 5b to 8b). Based on the $\mathbf{B}_q$ matrices, we identified groups of well correlated covariates. For instance, lab covariates hematocrit (Hct), hemoglobin (Hgb), and red blood cell (RBC) count showed the highest levels of correlation. Since both Hct and Hgb are known to be proportional to the number

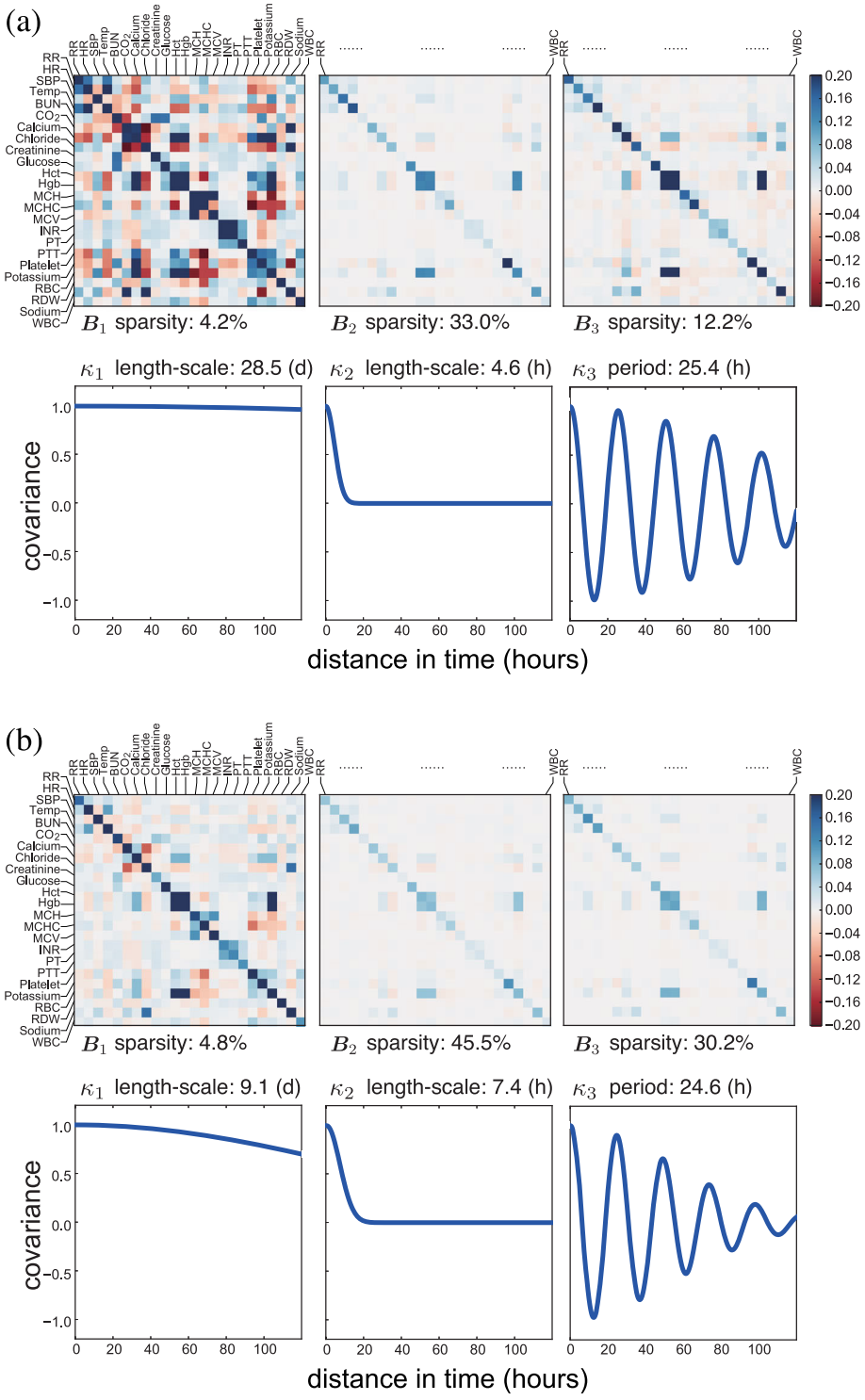

Fig. 5 The estimated population-level basis kernels and corresponding $\mathbf{B}_q$ matrices for septic patients. We show the kernels estimated (**a**) without a sparse prior and (**b**) with a sparse prior ($Q' = 3$). The sparsity of the $\mathbf{B}_q$ matrices is calculated as the percentage of nearly zero entries (i.e., values $\leq 10^{-3}$). The units for length-scale or period are (d) for days and (h) for hours

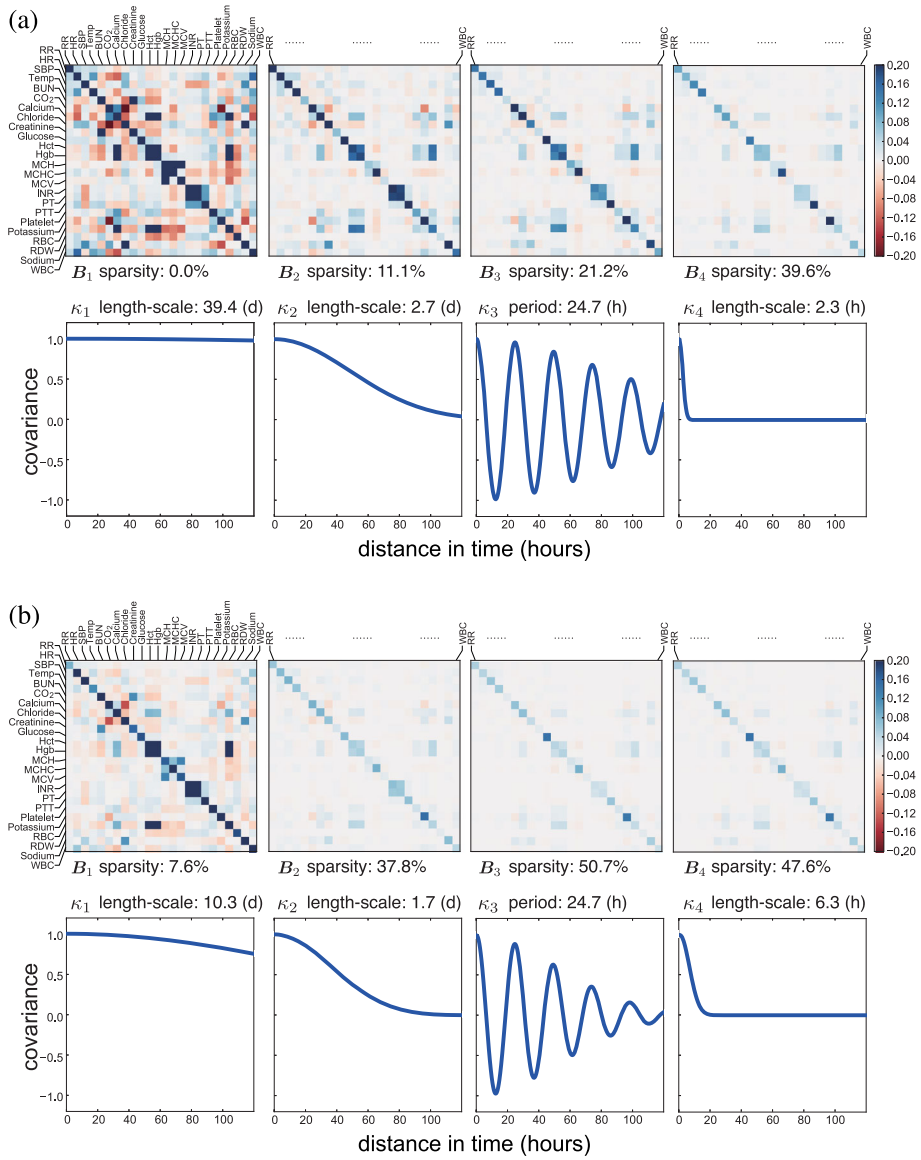

Fig. 6 The estimated population-level basis kernels and corresponding $\mathbf{B}_q$ matrices for patients with heart failure. We show the kernels estimated (**a**) without a sparse prior and (**b**) with a sparse prior ($Q' = 4$). The sparsity of the $\mathbf{B}_q$ matrices are calculated as the percentage of nearly zero entries (i.e., values $\leq 10^{-3}$). The units for length-scale or period are (d) for days and (h) for hours

of red blood cells, this positive correlation was encouraging [36]. The pair of lab covariates studied in the previous section, INR and PT, also showed substantial positive correlation. We found that the four vital signs—respiratory rate (RR), heart rate (HR), systolic blood pressure (SBP), and body temperature (Temp)—had substantial correlations with each other as well as weak correlations with some lab covariates. Another identifiable set of well-correlated covariates includes lab measurements of carbon dioxide (CO_2), calcium, chloride, potassium, and sodium. The three lab covariates related to the concentration of hemoglobin—mean cell hemoglobin (MCH), mean cell volume (MCV), and mean cell hemoglobin concentration (MCHC)—appeared to have substantial correlation (Fig. 5). The correlations modeled in these covariance matrices are exploited for accurate prediction and imputation in the MedGP framework.

To learn more about the importance of each kernel type across all subsets, we visualized the percent coverage of each type of kernel clusters found in the patients subsets (Fig. 9). The coverage of each kernel type is computed as the ratio of patients that have non-zero $\mathbf{B}_q$ matrix

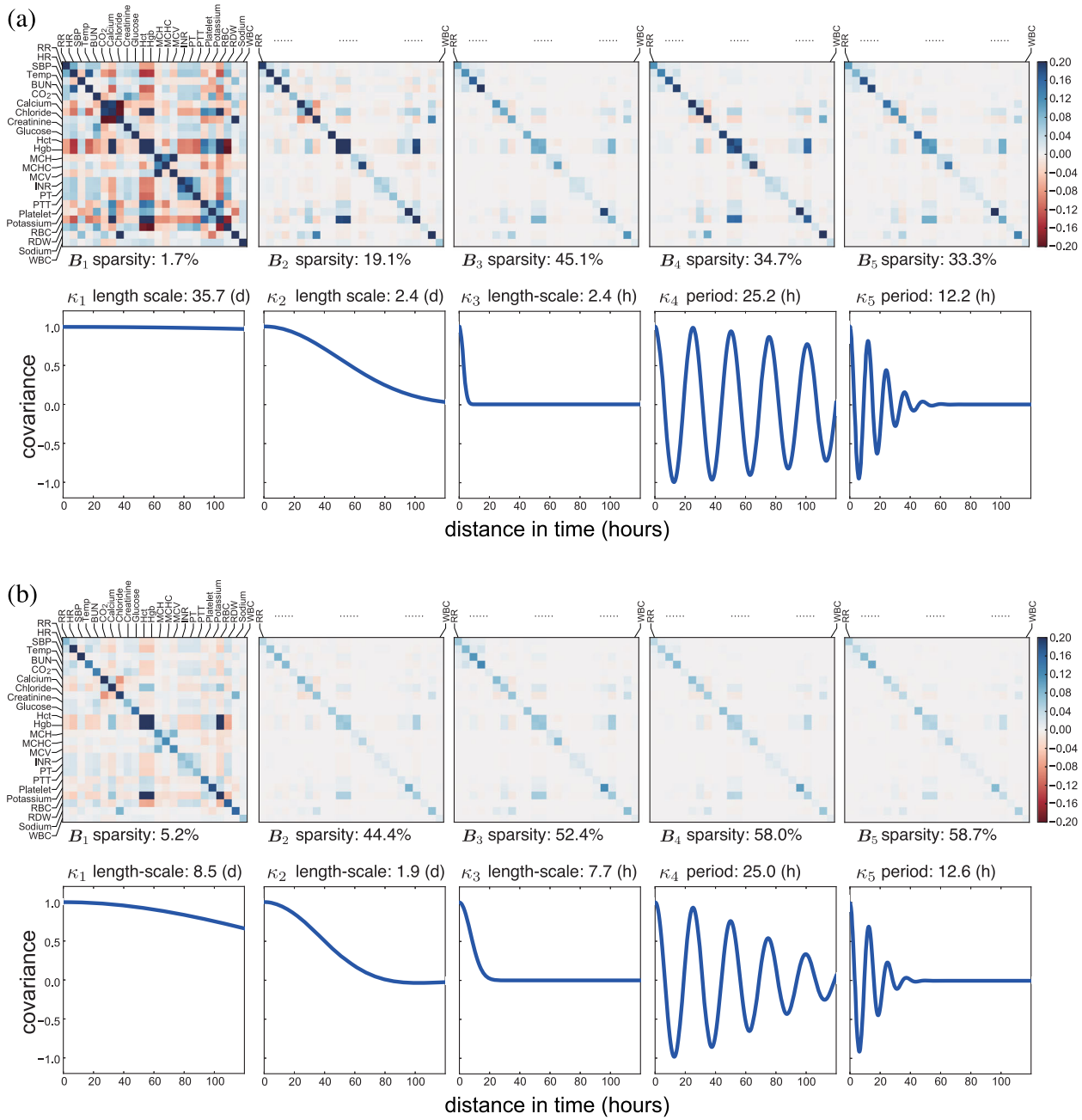

Fig. 7 The estimated population-level basis kernels and corresponding $\mathbf{B}_q$ matrices for patients with neoplasms. We show the kernels estimated (**a**) without a sparse prior ($Q' = 5$) and (**b**) with a sparse prior ($Q' = 5$). The sparsity of the $\mathbf{B}_q$ matrices are calculated as the percentage of nearly zero entries (i.e., values $\leq 10^{-3}$). The units for length-scale or period are (d) for days and (h) for hours

corresponding to it. We found that the kernel clusters with long-term (length scale $>$ 3 days) and short-term (length scale $<$ 12 h) dependencies have the highest coverage across the four subsets. In the MIMIC-III patients subset, the coverages of the short-term kernel, and the 12-h and 24-h periodic kernels are higher than that of in the HUP subsets. We think this is because the higher sampling frequency in the MIMIC-III patient subset enables more accurate estimation of the short-term and periodic dependencies.

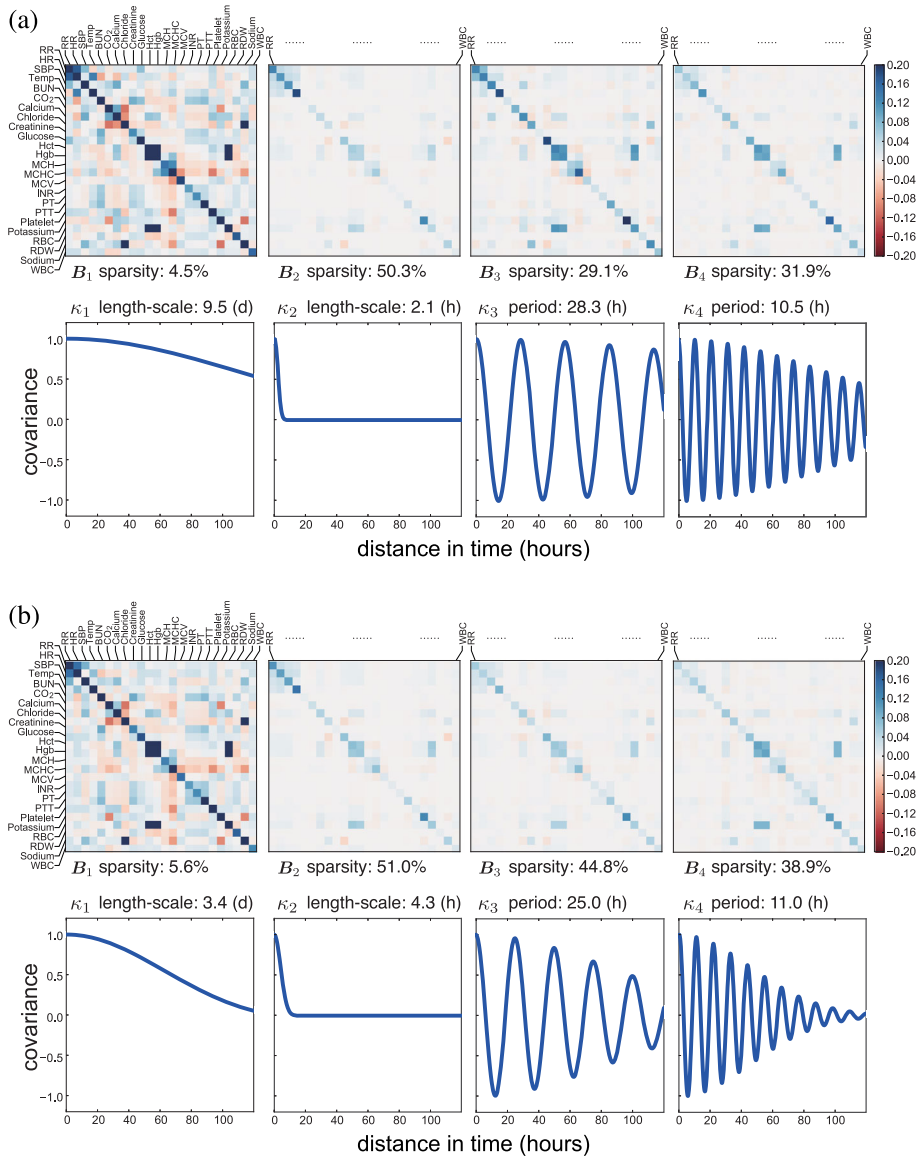

Fig. 8 The estimated population-level basis kernels and corresponding $\mathbf{B}_q$ matrices for 1003 patients with heart failure in MIMIC-III data set. We show the kernels estimated (**a**) without a sparse prior and (**b**) with a sparse prior ($Q' = 4$). The sparsity of the $\mathbf{B}_q$ matrices is calculated as the percentage of nearly zero entries (i.e., values $\leq 10^{-3}$). The units for length scale or period are (d) for days and (h) for hours

Results for online imputation

Next, we used the trained kernels to perform online imputation for each patient subgroup, where the goal is to predict the next observation for each covariate given the observations at previous time points. Across these methods, we used the percentage of improvement in MAE over three types of baselines—naive prediction, univariate GP (with SE or SM kernel), and PSM—to compare results for each of the 24 clinical covariates; we visualized the results separately (Figs. 10, 11 and 12; Figure B–E in Additional file 1: Appendix C; Figure F–U in Additional file 1: Appendix D). We also show the results of variations of our method for comparison (with or without the proposed sparse prior; with or without online updating). We performed paired t-tests on predictions from MedGP and each baseline to quantify the improvements, and statistical significance was evaluated using Bonferroni-corrected $p < 4.17 \times 10^{-4}$.

Comparing results with the independent GP model—specifically, selecting the best results from the SE or SM kernel, we found that MedGP, and in particular sparse SM-LMC with online updating, outperformed the

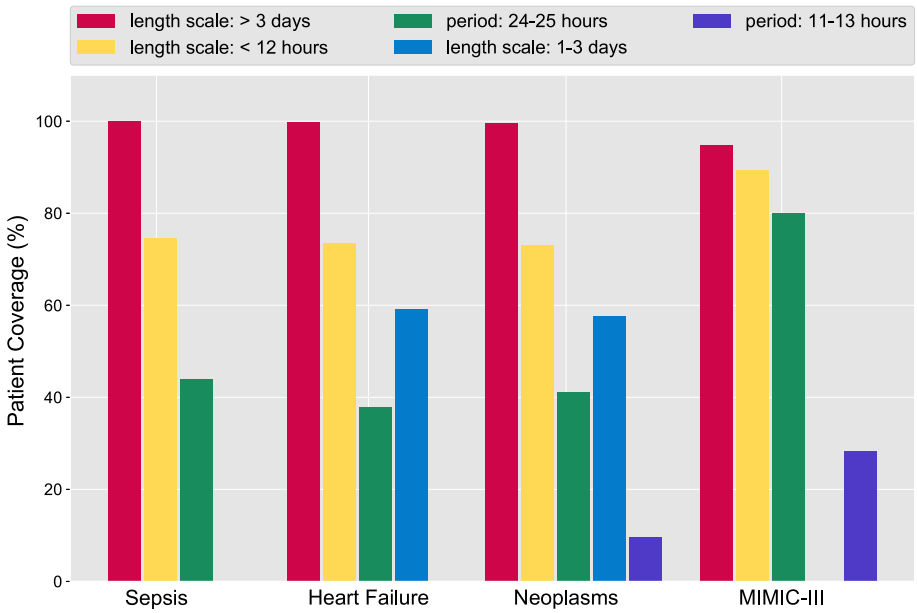

Fig. 9 The coverage over patients for each discovered kernel. We show the proportion of subjects that a kernel include, of each fuelypes of the population-level kernel clusters

independent GP model on the online imputation task for most covariates across the four patient groups (Fig. 10). In the HUP data, we found 18, 21 and 22 covariates significantly improved by MedGP in the sepsis, heart failure, and neoplasms groups, respectively. In the MIMIC-III patients subset, we found 19 covariates were improved. For all four groups, the number of covariates that were improved significantly by MedGP is greater than using SM-LMC kernels without the sparse prior. We found that the covariates that were well correlated in $\mathbf{B}_q$ usually showed significant positive improvements over independent GPs; Hct, Hgb, and RBC are notable examples. Similar observations could be made for INR and PT, the pair of lab covariates studied previously (Fig. 4). Across 24 covariates, the MAEs for INR and PT were slightly worse compared with only modeling these two covariates. However, we also observed that using the sparse prior with the SM-LMC kernel led to better performance as compared to not using the sparse prior, indicating that sparse regularization is helpful when jointly modeling heterogeneous covariates. Finally, there were some covariates for which MedGP did not improve over univariate GPs in two or more disease groups, including red cell distribution width (RDW), white blood cell count (WBC) and platelets.

When the baseline method is the naive one-lag method, for all four patient groups, we found fewer covariates with significant improvements compared with improvements over univariate GPs (Fig. 11). In particular, the covariates for which the naive method had an advantage were lab covariates that have piece-wise linear behavior, such as mean cell hemoglobin (MCH) and mean cell hemoglobin concentration (MCHC Fig. 1). In the case of piece-wise linear behavior, our kernel does not improve the performance compared with the naive approach since the time series are neither smooth nor periodic. Moreover, we also found that the naive method performed better in respiratory rate, PTT, platelet, RDW, and white blood cell (WBC) count. Overall, however, our method improved online prediction results for 18, 20 and 20 of the 24 covariates in sepsis, heart failure, and neoplasms groups, respectively. In the MIMIC-III subset, we found 14 covariates were improved significantly over the naive method.

When the baseline method is PSM [17], we found that our method outperformed PSM for most of the lab covariates, but PSM outperformed MedGP in imputation of vital signs and two lab covariates: glucose point-of-care (Glucose POC) and potassium (Fig. 12). For vital signs and glucose level, PSM has an advantage because of a higher sampling rate in those covariates and the highly structured mean function in the HUP subsets. The sampling rates are usually every 4 h for vital signs and every 8 h for glucose, which is more frequent than other lab covariates. Since PSM uses a B-spline basis function to capture the empirical mean, it may tolerate non-stationarity better. However, in the MIMIC-III subset, our method improved in imputing glucose and three vital signs (RR, SBP, Temp) over PSM. We think this reflects the higher sampling rate of the covariates that allows better estimation of the short-term temporal dependencies. Overall,

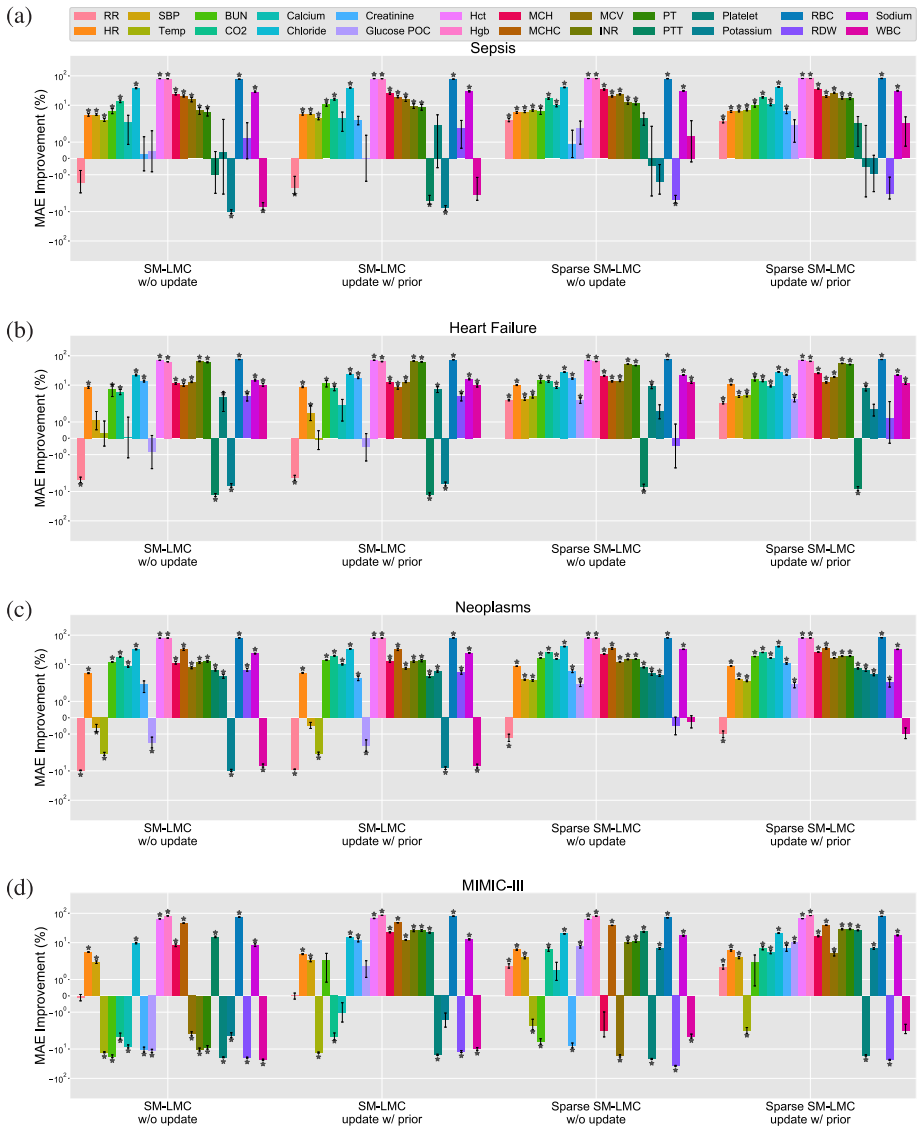

Fig. 10 The percent improvement using MedGP for online imputation compared to independent (unvariate) GPs. The figures depicts the results of 24 covariates for the (**a**) sepsis, (**b**) heart failure, and (**c**) neoplasms and (**d**) MIMIC-III heart failure subgroups. The *y*-axis is on log scale. The error bars denote ±1 standard error. The ⋆ indicates statistical significance p < 4.17 x 1e-4

MedGP significantly improved the imputation of 17, 20 and18 covariates in sepsis, heart failure, neoplasms subsets, respectively in the HUP data set, and 16 covariates in the MIMIC-III subset when compared with PSM. We contrast the PSM approach of structuring the mean function with our approach of structuring the kernel function, which leads to different types of gains in this problem.

Next, we looked at the calibration of the 95% coverage estimates (Figure D–E in Additional file 1: Appendix C; Fig N–U in Additional file 1: Appendix D). We found that MedGP outperformed independent GPs in terms of calibration of the 95% confidence region for all covariates. For this evaluation, values closer to 95% are better. We observed that the coverage using the non-sparse SM-LMC kernel was usually higher than the coverage using the sparse SM-LMC kernel in the three HUP subgroups, indicating that MedGP may slightly underestimate covariate-specific noise. In contrast, in the MIMIC-III subset, we observed that MedGP gave consistently more accurate 95% coverage than without regularization in most covariates. We also found that, in all patient subsets, online updating significantly improves

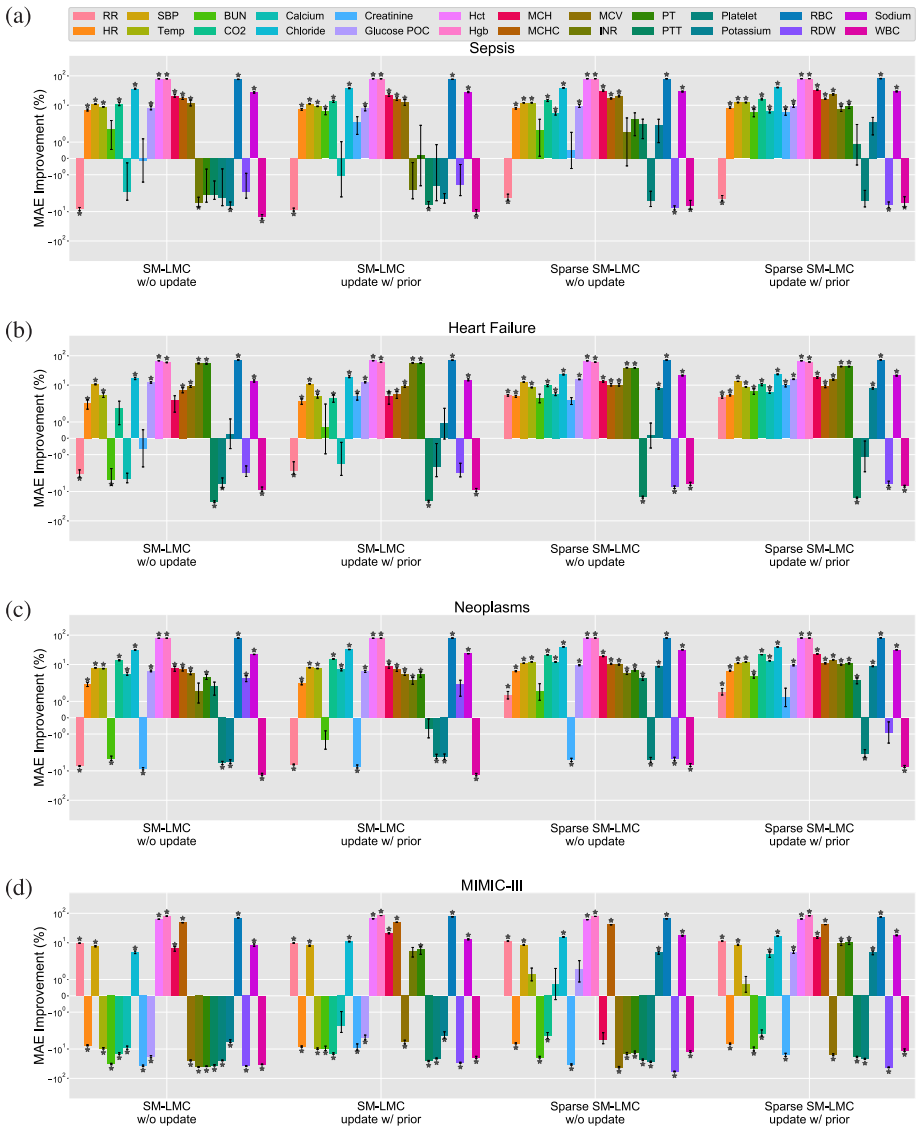

Fig. 11 The percent improvement using MedGP for online imputation compared to the naive method. The figures depicts the results of 24 covariates for the (**a**) sepsis, (**b**) heart failure, and (**c**) neoplasms and (**d**) MIMIC-III heart failure subgroups. The *y*-axis is on log scale. The error bars denote ± 1 standard error. The $\star$ indicates statistical significance $p < 4.17$ x 1e-4

the accuracy of the 95% coverage. Among all tested methods, PSM tended to overestimate the 95% confidence region. We think this is because PSM assumes that the input time series are aligned by patient status, and this alignment is not the case in our data. With unaligned data, PSM learned large marginal variance parameters due to high empirical variance of the observations across patients at the same elapsed time. In contrast, the estimation of marginal covariance parameters in MedGP is not affected by alignment because estimates are patient-specific. We also observed that, for either MedGP or PSM, the coverage was lower for some covariates in the MIMIC-III subset than in HUP subsets, such as temperature, CO_2, and PTT. This potentially reflects greater non-stationarity in the MIMIC-III subset, whose records were from intensive care units (ICUs) instead of regular hospital beds.

Finally, we compared the prediction performance of MedGP compared with the version without patient-specific online updating. We observed that online updating significantly improves the imputation errors of at least 12 out of 24 covariates in sepsis, heart failure, neoplasms, and the MIMIC-III subset. Similarly, evaluating the 95% coverage, all 24 covariates were improved by

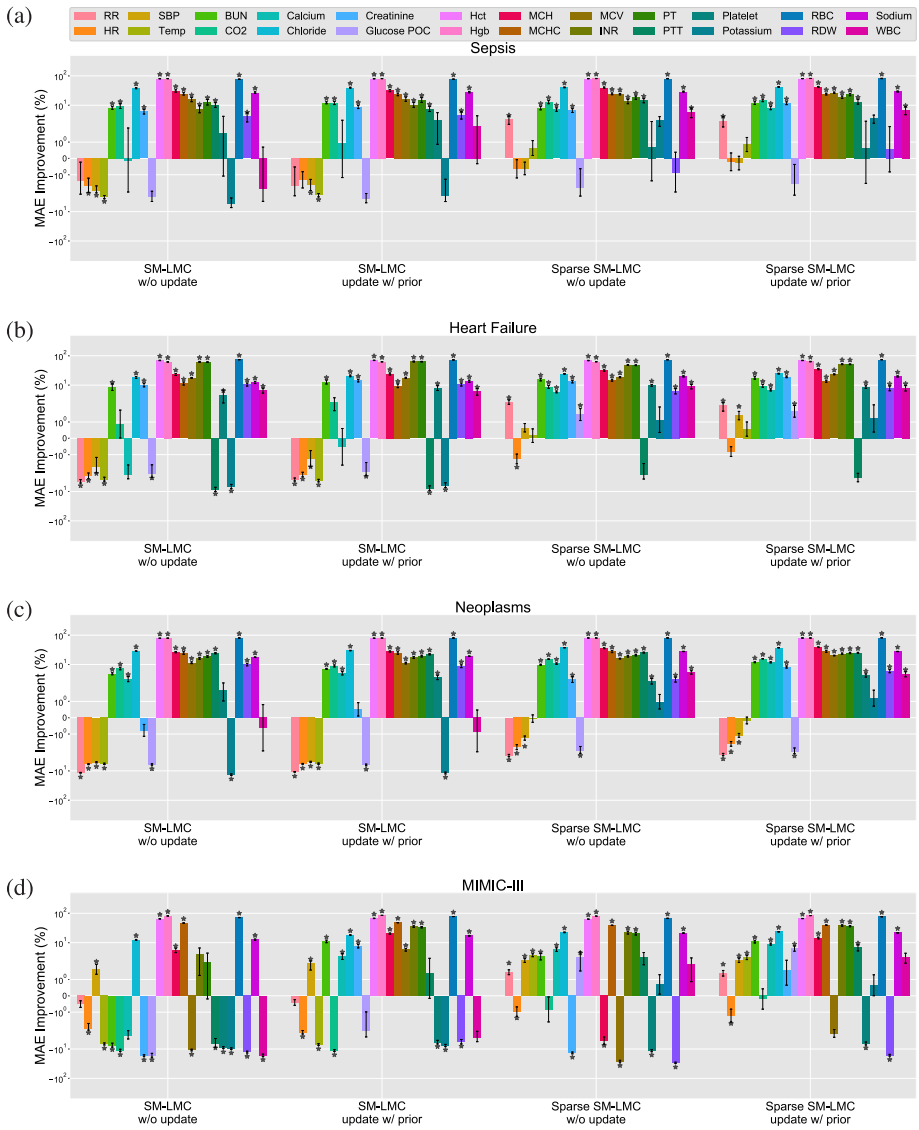

Fig. 12 The percent improvement using MedGP for online imputation compared to PSM. The figure depicts the results of 24 covariates for the (**a**) sepsis, (**b**) heart failure, and (**c**) neoplasms and (**d**) MIMIC-III heart failure subgroups. The *y*-axis is on log scale. The error bars denote ±1 standard error. The ⋆ indicates statistical significance p < 4.17 x 1e-4

online updating across the three diseases groups in HUP, and 18 covariates were improved in the MIMIC-III subset (Figure D–E in Additional file 1: Appendix C; Figure N–U in Additional file 1: Appendix D). This improvement highlights the importance of updating the empirical priors with patient-specific observations for this problem.

Computational efficiency and scalability

In this section, we compare computational speed between different implementations of our method. For patients with only a few observations, an existing implementation using conventional GP inference is sufficient for computationally tractable online inference. However, since our data include a large number of patients with potentially thousands of observations each, we implemented an exact inference algorithm in C++ and optimized it through Intel MKL libraries and customized multithreading blocks. In the experimental setting of $Q = 5$, $D = 24$, and $R_q = 8$, there are 1114 hyperparameters to estimate. We summarized the runtime under different implementations for one patient with 2028 unique time points and 6679 observations (Table 3); the tests were performed using a machine with Intel® Xeon® CPUs running at 2.40GHz. Using our optimized

Table 3 Training time (in seconds) for a single iteration under different implementations of MedGP

Implementation	Sequential	Multithreading
Computing Gram matrix	11	2
Inverting Gram matrix	13	3
Computing gradients	2497	97
Total per iteration	2521	102

The total number of observations across time for this patient is 6679. The sequential test used a single CPU, while the multithreading test used 35 CPUs—one thread per CPU

implementation, for patients with a large number of observations ($T_i \geq 5000$), we accelerated training by a factor of 10 to 25 on average as compared with the sequential approach. We also compared our implementation with the standard GPy [46] implementation under different sample sizes and Q, and reached empirically at least three times speed up. We provide these results in Additional file 1: Appendix E.

The proposed framework can be parallelized at the patient level and is suitable for analysis when patient data are observed in a streaming form. For each reference patient, we distributed the optimized training process on a computing cluster to estimate the patient-specific hyperparameters in parallel. In addition, the population-level kernels could be updated sequentially; the computationally expensive GP training procedure does not need to be applied to patient data in bulk. That is, when we receive more data from new patients, we only need to update the kernel density estimators. Our framework provides better computational efficiency compared to models designed for smaller collections of observations (e.g., approximately two hundred observations for each patient) as in most previous work. Those approaches are computationally intractable when working on a set of rich patient observations of the magnitude of the HUP data due to large matrix inversions and summing marginal likelihoods across patients at each iteration.

Discussion

We showed that our method, MedGP, improves performance for online prediction of 24 clinical covariates as compared with independent univariate GPs, a naive method of propagating the previous observation, and an earlier state-of-the-art approach, PSM [17]. We found that, for well-correlated covariates, our method improves online imputation performance substantially over the related methods in most tested covariates. The improvements over the naive one-lag prediction and univariate GPs were significant in both vital signs and lab covariates. We found that PSM was, in general, better at predicting vital signs with more densely sampled observations. However, our approach does not require patient time series alignment and shows better calibration of the 95% confidence region as compared to PSM.

There are several directions that will be explored using the MedGP framework motivated by the present results. The first direction is to allow time-varying covariances by specifically modeling non-stationarity. Some possible approaches to explore include incorporating state-space models or change point detection [47, 48], and extending those methods to work on multivariate scenarios. Another direction of interest is to consider latent subpopulation-level structured kernels through multivariate medical time series. We expect that our results could be further improved through incorporating hierarchical methods with proper features or metrics to represent the differences between patients within the same disease group and across disease groups more carefully. For instance, the original PSM used three levels of hierarchy based on the subpopulations of patients with scleroderma, including population level, subpopulation level, and individual level. Our model may benefit from such an approach, but more efficient inference procedures are needed to train on our large data set [49]. We should point out that this is possible through, for instance, deriving corresponding stochastic variation inference (SVI) algorithm. For example, previous work develops an SVI algorithm for a semiparametric latent factor model (SLFM) with $R_q = 1$ [50], which could be generalized to apply to MedGP.

For future applications, we will use the framework to monitor the health status of patients in a hospital setting and identify those patients at high risk for acute diseases in order to assist with decision making in treatment plans. Specifically, MedGP can impute latent state in patients at any time point, including confidence region around those estimates; this latent state can be used for a number of downstream analyses that require complete knowledge of patient state at specific time points. For instance, the changes of dynamics and temporal correlations between two vital signs have been found to be useful for disease detection given high-frequency regularly sampled time series [19, 20]. We demonstrated that MedGP accurately estimates the temporal correlations in the presence of sparse, unaligned time-series data for up to 24 covariates, and we would expect to further associate the cross-covariate dynamics to more complicated diseases, such as septic shock [51], where the interactions of multiple covariates are jointly taken into consideration for diagnosis.

Conclusions

In this paper, we propose a flexible and efficient framework for estimating the temporal dependencies across multiple sparse and irregularly sampled medical time series data. We developed a model with multi-output Gaussian process regression with a highly structured

kernel. We fit this model using an optimized implementation of exact GP inference to three different disease groups in the HUP medical data set and the MIMIC-III ICU data set. We demonstrate in the results that our model is a robust and reliable estimate of patient state upon which downstream medical analyses can be built.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12911-020-1069-4.

Additional file 1: Appendix.

Abbreviations

EHR: electronic health record; ICD: International Classification of Diseases; HUP: Hospitals at the University of Pennsylvania; MIMIC-III: Multiparameter Intelligent Monitoring in Intensive Care; ICU: intensive care unit; GP: Gaussian process; STGP: single-task GP; MTGP: multi-task GP; HMM: hidden Markov model; AR: autoregressive; LDS: linear dynamical systems; SVM: support vector machine; KNN: *k*-nearest neighbor; PSM: Probabilistic Subtyping Model; SVAR: switching vector autoregressive model; TGM: time series graphical model; VAR: vector autoregressive; C-LTM: Coupled Latent Trajectory Model; CRF: conditional random field; LMC: linear model of coregionalization; SE: squared exponential; SM: spectral mixture; HR: heart rate; BP: blood pressure; SBP: systolic blood pressure; RR: respiratory rate; PT: prothrombin time; INR: international normalization ratio; PTT: partial thromboplastin time; Hct: hematocrit; Hgb: hemoglobin; RBC: red blood cell; Temp: body temperature; BUN: blood urea nitrogen; CO_2: carbon dioxide; MCH: mean cell hemoglobin; MCV: mean cell volume; MCHC: mean cell hemoglobin concentration; RDW: red cell distribution width; WBC: white blood cell; POC: point-of-care; MAE: mean absolute error; SVI: stochastic variational inference; SLFM: semiparametric latent factor model

Acknowledgements

The authors would like to thank Derek Aguiar and Niranjani Prasad for providing insightful comments and discussion on the manuscript.

Authors' contributions

L-FC and BEE conceived the idea and designed the experiments. L-FC, CC, and MD provided and processed the patient data. L-FC, BD, and GD implemented and ran the experiments. L-FC, CC, MD, and BEE analyzed the results. L-FC and KL developed the computational efficiency results. L-FC and BEE wrote the paper; all authors edited the paper. All authors read and approved the final manuscript.

Funding

BEE was supported by NIH R01 MH101822, NIH R01 HL133218, NIH U01 HG007900, a Sloan Faculty Fellowship, and NSF CAREER AWD1005627. L-FC was supported by the Helen Shipley Hunt Fund for Innovation. The funders had no role in the design of the study nor the collection, analysis, and interpretation of data or writing the manuscript.

Availability of data and materials

The MIMIC-III dataset is publicly available upon request through https://mimic.physionet.org/gettingstarted/access/. The HUP dataset is not publicly shareable due to constraints of the IRB.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

BEE is on the SAB for Celsius Therapeutics and Freenome, and is currently employed by Genomics plc and Freenome and on a year leave-of-absence from Princeton University (2019–2020). L-FC is now at Verily Life Sciences, LLC. All work was done at Princeton University.

Author details

[1]Department of Electrical Engineering, Princeton University, Princeton USA. [2]Lewis-Sigler Institute, Princeton University, Princeton, NJ, USA. [3]University of Pennsylvania Health System, Philadelphia, PA, USA. [4]Department of Computer Science, Princeton University, Princeton, NJ USA. [5]Center for Statistics and Machine Learning, Princeton University, Princeton, NJ, USA.

Received: 9 November 2018 Accepted: 5 March 2020
Published online: 08 July 2020

References

1. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA. 2013;309(13):1351–2.
2. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 2013;20(1):117–21.
3. Ghassemi M, Celi LA, Stone DJ. State of the art review: the data revolution in critical care. Crit Care. 2015;19(1):118.
4. Johnson AE, Pollard TJ, Shen L, Lehman L-wH, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MiMIC-III, a freely accessible critical care database. Sci Data. 2016;3(1).
5. Hotchkiss RS, Karl IE. The pathophysiology and treatment of sepsis. N Engl J Med. 2003;348(2):138–50.
6. Angus DC, Linde-Zwirble WT, Lidicker J, Clermont G, Carcillo J, Pinsky MR. Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. Crit Care Med. 2001;29(7):1303–10.
7. Kumar G, Kumar N, Taneja A, Kaleekal T, Tarima S, McGinley E, Jimenez E, Mohan A, Khan RA, Whittle J, Jacobs E, Nanchal R. Nationwide trends of severe sepsis in the 21st century (2000-2007). Chest. 2011;140(5):1223–31.
8. Pierrakos C, Vincent J-L. Sepsis biomarkers: review. Crit Care. 2010;14(1):15.
9. Newgard CD, Lewis RJ. Missing data: How to best account for what is not known. JAMA. 2015;314(9):940–1.
10. Kim J, Blum JM, Scott CD. Temporal features and kernel methods for predicting sepsis in postoperative patients. Technical Report, , University of Michigan, USA. 2010.
11. Ho JC, Lee CH, Ghosh J. Septic shock prediction for patients with missing data. ACM Trans Manag Inf Syst. 2014;5(1):1–1115.
12. Stanculescu I, Williams CKI, Freer Y. A hierarchical switching linear dynamical system applied to the detection of sepsis in neonatal condition monitoring. In: Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence. AUAI Press; 2014.
13. Marlin BM, Kale DC, Khemani RG, Wetzel RC. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. In: Proceedings of the 2nd ACM SIGHIT symposium on International Health Informatics - IHI '12. ACM Press; 2012.
14. Roberts S, Osborne M, Ebden M, Reece S, Gibson N, Aigrain S. Gaussian processes for time-series modelling. Philos Trans R Soc Lond A Math Phys Eng Sci. 2012;371(1984).
15. Stegle O, Fallert SV, MacKay DJC, Brage S. Gaussian process robust regression for noisy heart rate data. IEEE Trans Biomed Eng. 2008;55(9):2143–51.
16. Lasko TA, Denny JC, Levy MA. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. PloS One. 2013;8(6):1–13.
17. Schulam P, Wigley F, Saria S. Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. AAAI Press; 2015.
18. Schulam P, Saria S. A framework for individualizing predictions of disease trajectories by exploiting multi-resolution structure. In: Advances in Neural Information Processing Systems 28. Curran Associates, Inc.; 2015. p. 748–56.
19. Nemati S, Lehman L-WH, Adams RP, Malhotra A. Discovering shared cardiovascular dynamics within a patient cohort. In: 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE; 2012. p. 6526–9.
20. Lehman L-WH, Adams RP, Mayaud L, Moody GB, Malhotra A, Mark RG, Nemati S. A physiological time series dynamics-based approach to patient monitoring and outcome prediction. IEEE J Biomed Health Inform. 2015;19(3):1068–76.

21. Rizopoulos D, Ghosh P. A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. Stat Med. 2011;30(12):1366–80.
22. Dahlhaus R. Graphical interaction models for multivariate time series. Metrika. 2000;51(2):157–72.
23. Tank A, Foti N, Fox E. Bayesian structure learning for stationary time series. In: Proceedings of the Thirty-first Conference on Uncertainty in Artificial Intelligence. AUAI Press; 2015.
24. Gather U, Imhoff M, Fried R. Graphical models for multivariate time series from intensive care monitoring. Stat Med. 2002;21(18):2685–701.
25. Schulam P, Saria S. Integrative analysis using coupled latent variable models for individualizing prognoses. J Mach Learn Res. 2016;17(234):1–35.
26. Journel AG, Huijbregts CJ. Mining Geostatistics: Academic Press; 1978.
27. Goovaerts P. Geostatistics for Natural Resources Evaluation: Oxford university press; 1997.
28. Bonilla EV, Chai KM, Williams CKI. Multi-task Gaussian process prediction. In: Advances in Neural Information Processing Systems 20; 2008. p. 153–60.
29. Teh YW, Seeger M, Jordan MI. Semiparametric latent factor models. In: Proceedings of the International Workshop on Artificial Intelligence and Statistics, vol. 10; 2005.
30. Titsias MK, Lázaro-Gredilla M. Spike and slab variational inference for multi-task and multiple kernel learning. In: Advances in Neural Information Processing Systems 24. Curran Associates, Inc.; 2011. p. 2339–47.
31. Álvarez MA, Lawrence ND. Computationally efficient convolved multiple output Gaussian processes. J Mach Learn Res. 2011;12:1459–500.
32. Ghassemi M, Pimentel MAF, Naumann T, Brennan T, Clifton DA, Szolovits P, Feng M. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence; 2015. p. 446–53.
33. Dürichen R, Pimentel MAF, Clifton L, Schweikard A, Clifton DA. Multitask Gaussian processes for multivariate physiological time-series analysis. IEEE Trans Biomed Eng. 2015;62(1):314–22.
34. Wilson AG, Adams RP. Gaussian process kernels for pattern discovery and extrapolation. In: Proceedings of the 30th International Conference on Machine Learning. JMLR.org; 2013. p. 1067–75.
35. Rasmussen CE, Williams CKI. Gaussian Processes for Machine Learning: The MIT Press; 2006.
36. Widmaier EP, Raff H, Strang KT. Vander, Sherman, Luciano's Human Physiology: the Mechanisms of Body Function. 9th Edition. Boston: McGraw-Hill Higher Education; 2004.
37. Polson NG, Scott JG. Shrink globally, act locally: Sparse Bayesian regularization and prediction. Bayesian Stat. 2010;9:501–38.
38. Gao C, Brown CD, Engelhardt BE. A latent factor model with a mixture of sparse and dense factors to model gene expression data with confounding effects. arXiv preprint arXiv:1310.4792. 2013.
39. Álvarez MA, Rosasco L, Lawrence ND. Kernels for vector-valued functions: a review. Found Trends Mach Learn. 2012;4(3):195–266.
40. Carvalho CM, Polson NG, Scott JG. The horseshoe estimator for sparse signals. Biometrika. 2010;97(2):465–80.
41. Armagan A, Clyde M, Dunson DB. Generalized beta mixtures of Gaussians. In: Advances in Neural Information Processing Systems 24. Curran Associates, Inc.; 2011. p. 523–31.
42. Zhao S, Gao C, Mukherjee S, Engelhardt BE. Bayesian group factor analysis with structured sparsity. J Mach Learn Res. 2016;17(196):1–47.
43. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
44. Silverman BW. Density Estimation for Statistics and Data Analysis: CRC press; 1986.
45. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. In Nature. 1986;533–536.
46. GPy. GPy: A Gaussian process framework in Python. 2012. Version 1.8.5.
47. Adams RP, MacKay DJC. Bayesian online changepoint detection. arXiv preprint arXiv:0710.3742. 2007.
48. Saatçi Y, Turner R, Rasmussen CE. Gaussian process change point models. In: Proceedings of the 27th International Conference on Machine Learning; 2010. p. 927–34.
49. Feinberg V, Cheng L-F, Li K, Engelhardt BE. Large linear multi-output gaussian process learning for time series. arXiv preprint arXiv:1705.10813. 2017.
50. Nguyen TV, Bonilla EV. Collaborative multi-output Gaussian processes. In: Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence. AUAI Press; 2014.
51. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. Sci Transl Med. 2015;7(299):ra122.

Publisher's Note