Medical data wrangling with sequential variational autoencoders

Daniel Barrejón, Pablo M. Olmos, Antonio Artés-Rodríguez

***Abstract*—Medical data sets are usually corrupted by noise and missing data. These missing patterns are commonly assumed to be completely random, but in medical scenarios, the reality is that these patterns occur in bursts due to sensors that are off for some time or data collected in a misaligned uneven fashion, among other causes. This paper proposes to model medical data records with heterogeneous data types and bursty missing data using sequential variational autoencoders (VAEs). In particular, we propose a new methodology, the Shi-VAE, which extends the capabilities of VAEs to sequential streams of data with missing observations. We compare our model against state-of-the-art solutions in an intensive care unit database (ICU) and a dataset of passive human monitoring. Furthermore, we find that standard error metrics such as RMSE are not conclusive enough to assess temporal models and include in our analysis the cross-correlation between the ground truth and the imputed signal. We show that Shi-VAE achieves the best performance in terms of using both metrics, with lower computational complexity than the GP-VAE model, which is the state-of-the-art method for medical records.**

***Index Terms*—Deep learning, VAE, missing data, heterogeneous, sequential data**

I. Introduction

Since machine learning emerged, all the primary attention focused on working with homogeneous data sets, where too few artifacts such as outliers or missing data barely appear. But real-world data sets are quite different. Data is usually organized in databases containing incomplete, noisy, and more critical, heterogeneous information sources. These scenarios are quite common in medical applications. For instance, Electronic Health Records (EHR) may contain information from monitoring sensors, different physicians' diagnoses, or visits to the hospital. A heterogeneous medical footprint hence defines each patient. This kind of information will exhibit missing data due to sensors' failures or due to temporal gaps between each visit to the hospital, to name a few.

In the literature, the common assumption is that the lost information from a data set is Missing Completely at Random (MCAR). However, the most usual scenario is that missing data follows some kind of pattern. For example, in human monitoring applications the sensors tracking different sources might disconnect for some amount of time, not intermittently, generating *bursts of missing data*. For medical data sets missing patterns can appear simultaneously across different attributes as it is shown in Figure 1.

The recent literature on machine learning (ML) approaches to handle noise and missing data in medical records is dominated by deep learning methods. In this regard, recurrent neural networks (RNN) stand as one of the most popular approaches. In [1] the authors propose Long-Short-Term Memory (LSTM) networks [2], to recognize patterns in multivariate time series of clinical measurements. This work was extended in [3] with binary indicators of missingness as features. A different approach is proposed in [4], where Gated Recurrent Units (GRU) are modified to incorporate missing masks, hence modeling the time intervals between clinical appointments. Other works like BRITS [5] also look into the bidirectional capabilities of RNNs and exploit this property to impute missing values in time series with underlying nonlinear dynamics.

Although the above RNN-based methods show impressive results dealing with time series forecasting, they do not benefit from the flexibility and the underlying data correlations inferred by probabilistic deep generative models (DGMs). DGMs capture inner correlations that can be present in high-dimensional data employing a low-dimensional latent space. In the framework of VAEs, the heterogeneous incomplete variational autoencoder (HI-VAE) [6], the mixed VAE (VAEM) [7], the MIWAE [8], the Partial VAE presented in [9] or similar works [10] [11] propose efficient methods to jointly model different data types and missing data in a single DGM. Among DGMs able to deal with sequential data, GP-VAE [12] stands out. GP-VAE implements a latent probabilistic model in which a Gaussian process captures the correlation of the low-dimensional latent variable along time, and this GP relies on a VAE to implement the observation model. However, GP-VAE cannot deal with heterogeneous observations. Finally, DGM-like solutions to deal with tabular or sequential based on generative adversarial networks (GANs), such as GAIN in [13], the gated recurrent GAN in [14], MisGAN in [15] and VIGAN [16] do not show to outperform the imputation ability of other VAE-based methods and are harder to train due to the min-max underlying optimization problem.

In this paper, we consider modeling sequential heterogeneous data when missing data comes in bursts, a scenario in which none of the previous DGMs have been tested to date. On the one hand, we show that when errors come in bursts, standard error metrics such as normalized mean-squared error (NRMSE) do not reflect well the imputation accuracy, and we study the correlation between the ground-truth signal and the imputed one. In this setup, we demonstrate that GP-VAE struggles to deal with long-missing data bursts since the underlying GP correlation quickly decays, driving the GP posterior to a non-informative mean and large variance.

To better deal with bursty missing patterns, we propose the sequential heterogeneous incomplete VAE (Shi-VAE). This

 DOI: 10.1109/JBHI.2021.3123839

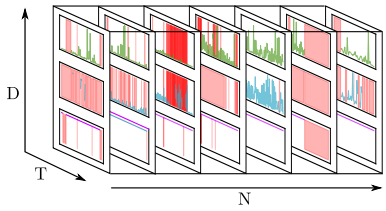

Fig. 1: Example of heterogeneous streams of data with missing values from the medical data set. Red vertical lines correspond to missing values. Each row corresponds to a different type of data: the first two correspond to positive real-valued data and the third to binary data. D refers to dimensionality of the dataset, T to the temporal dimension and N to the number of samples.

model generalizes the HI-VAE model in [6], including a latent temporal structure driven by LSTMs following a similar idea as in [17]. The extended memory properties of these networks provide a more robust ability to cope with missing bursts, efficiently capturing into the low-dimensional latent projection the correlation to past observations. Besides, Shi-VAE comes with efficient training methods based on amortized variational inference that can handle massive data sets. As a representative example of a medical database, we demonstrate the superior ability of Shi-VAE to deal with complex time-series using two real data sets. First, we consider the data set from the 2012 Physionet Challenge [18] which contains measurements of 35 electrophysiological signals for 12,000 patients monitored during 48 on the intensive care unit (ICU). Second, we consider a data set of human passive monitoring coming from mobile devices. It contains heterogeneous attributes (distance travelled, mobile phone usage, quality of sleep, etc.) and a challenging presence of bursty missing data. The Shi-VAE code to reproduce our experiments can be found in https://github.com/dbarrejon/Shi-VAE. Overall, we claim the following contributions:

- We propose Shi-VAE as a robust generative model to handle heterogeneous time series corrupted with missing data.
- We demonstrate that NRMSE is a partial metric when it comes to compare imputation models in the presence of missing data in bursts.
- We propose to use a temporal correlation metric to compare the different models. This metric is more sensitive to detect over-smooth solutions.

We organize the paper as follows. Firstly, Section II introduces the problem statement we want to tackle. Section III presents Shi-VAE. In Section IV we present the two data sets we have used to validate our model and the results we have found. Section V presents our final remarks.

II. A Human Monitoring Database

Through patients' mobile phones and other wearable devices, continuous sensor data can be collected in a non-invasive manner, providing valuable information about everyday activity patterns. The possibility of inferring emotional states by analyzing smartphone usage data [19], [20], GPS traces of movement [21], social media data [22], and even sound recordings [23] has become a growing research focus over the past decade.

One of the databases that we use in this paper was collected using the mobile application eB2 MindCare [24] in a collaboration we carried out with two public mental health hospitals in Madrid (Hospital Universitario Fundación Jiménez Díaz and Hospital Universitario Rey Juan Carlos). This study was approved by the Fundación Jiménez Díaz Research Ethics Committee (Study code: LSRG-1-005 16). We periodically capture passive monitoring information from $N = 170$ psychiatric patients using eB2 MindCare, thus registering different signals for every user. In particular, we are working with daily summary representations of every variable. The seven attributes we work with are listed in Table I along with the fraction of missing values across all patients.

Regarding the positive variables, distance, steps total, and vehicle are related to the patient's mobility. App usage is a positive variable that measures the total amount of active time the user has been using the phone, with social applications, phone calls, etc. Sleep is a positive variable that counts the total time a person has slept during a day. Regarding binary variables, sport explains whether the person has done any sport $x_t = 1$ or not $x_t = 0$ during the day and steps home states whether the person was at home $x_t = 1$ or not $x_t = 0$ at that particular day.

Variable	Type	Missing Percentage [%]
Distance	Positive	42
Steps Home	Binary	66
Steps Total	Positive	22
App Usage	Positive	38
Sport	Binary	62
Sleep	Positive	31
Vehicle	Positive	44

TABLE I: Human Monitoring data set.

Finally, we remark that, although the number D of attributes is the same for every patient ($D = 7$), the signal length T per patient is very diverse. The average sequence length is 233. Figure 2 illustrates the whole population and the missing pattern. From the Figure 2 we can observe that almost any day comes with missing values, and hence we can expect long bursts of missing attributes.

In this paper, we demonstrate the superior ability of the proposed Shi-VAE to capture the non-trivial correlations among the database attributes and accurately impute missing values.

III. Proposed Model

We first introduce a general notation of the problem and then present the Shi-VAE model.

A. Notation

We define our data set as $\mathcal{D} = \{\mathbf{X}^1, \ldots, \mathbf{X}^N\}$, where N corresponds to the total number of samples in the data

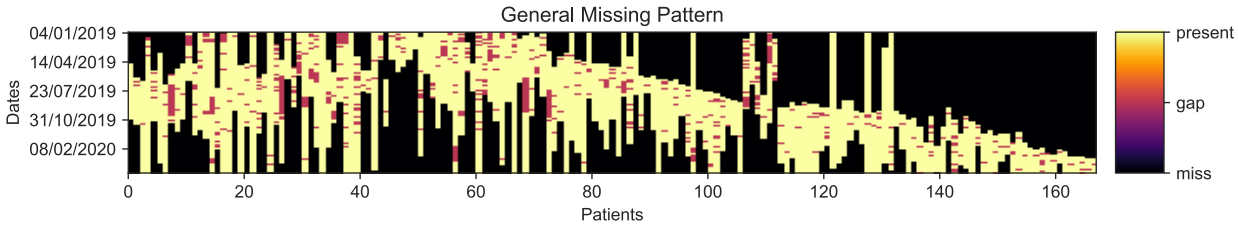

Fig. 2: Overall view of the human monitoring database. Each patient has a given sequence length. Black means no record of that patient, magenta means a complete missing day and yellow that at least there is one variable present at that day.

set. Each sample $\mathbf{X}^n \in \mathbb{R}^{T^n \times d}$ has T^n observations $\mathbf{x}_t = [x_{t1}, \dots, x_{td}]^\top \in \mathbb{R}^d$, where d refers to the dimension or attribute. From now on, we use $\mathbf{X}^n = \mathbf{X}$ in order to relax notation. We consider heterogeneous attributes:

- **Continuous Variables**:
  1) **Real-valued data**: Data taking real values, *i.e.*, $x_{td} \in \mathbb{R}$.
  2) **Positive-valued data**: Data taking only positive values, *i.e.*, $x_{td} \in \mathbb{R}^+$.
- **Discrete Variables**:
  1) **Binary Data**: Data can only be either 1 or 0, *i.e.*, $x_{td} \in [0, 1]$.
  2) **Categorical data**: Data taking values in a finite unordered set, *i.e.*, $x_{td} \in \{-1, 0, 1\}$, or $x_{td} \in \{$'negative, 'neutral, 'positive'$\}$.

Furthermore, we assume that any $\mathbf{x}_t$ can have both observed values and missing values. Let us define $\mathcal{O}_t$ as the index set for the observed attributes at time t and $\mathcal{M}_t$ as the missing index at the same time. Hence $\mathcal{O}_t \cap \mathcal{M}_t = \emptyset$. With this notation, we can split this sentence into a vector containing observed attributes $\mathbf{x}_t^o$, and a complementing vector containing missing attributes $\mathbf{x}_t^m$.

B. The sequential heterogeneous incomplete VAE (Shi-VAE)

This section presents the Shi-VAE probabilistic generative model, which extends the capabilities of a standard VAE to sequential heterogeneous data streams and handles missing data. In Shi-VAE, the temporal dependencies and shared correlations among attributes are captured by a latent hierarchy of low-dimensional latent variables: a continuous latent variable $\mathbf{z}_t \in \mathbb{R}^K$, which follows a Mixture of Gaussian's (MoG) Prior distribution [25], and a discrete latent variable $\mathbf{s}_t$ that represents the component of the MoG[1]. We model the dependence between these two latent variables and the temporal data as follows:

$$p(\mathbf{X}, \mathbf{Z}, \mathbf{S}) = \prod_{t=1}^{T} p_{\theta_x}(\mathbf{x}_t|\mathbf{z}_{\leq t}, \mathbf{s}_t) p_{\theta_z}(\mathbf{z}_t|\mathbf{z}_{<t}, \mathbf{s}_t) p_{\theta_s}(\mathbf{s}_t), \quad (1)$$

[1]Another option for the prior would be to use a mixture of posteriors as prior also known as VampPrior [26]. However, to us it is more reasonable to use a prior that is not dependent on the posteriors distributions, due to the implicit dependencies present in the model.

where $\mathbf{Z} = \mathbf{z}_{\leq T}$ and $\mathbf{S} = \mathbf{s}_{\leq T}$. The joint probability density function is parameterized by $\theta = \{\theta_x, \theta_z, \theta_s\}$. From now on, we omit this dependency to further relax notation. Following [6], we assume that given the latent variable $\mathbf{z}_t$ encodes all the correlation among attributes and hence they are all conditionally independent

$$p(\mathbf{x}_t|\mathbf{z}_{\leq t}, \mathbf{s}_t) = \prod_{d \in \mathcal{O}_t} p(x_{td}|\mathbf{z}_{\leq t}, \mathbf{s}_t) \prod_{d \in \mathcal{M}_t} p(x_{td}|\mathbf{z}_{\leq t}, \mathbf{s}_t). \quad (2)$$

The actual expression for each of the likelihood factors $p(x_{td}|\mathbf{z}_{\leq t}, \mathbf{s}_t)$ depends on the data-type of every attribute, as we develop in the next sub-section.

The temporal dependency is encoded into the term $p_{\theta_z}(\mathbf{z}_t|\mathbf{z}_{<t}, \mathbf{s}_t)$, which implements a RNN-based model to capture the temporal data correlation along time:

$$p(\mathbf{z}_t|\mathbf{z}_{<t}, \mathbf{s}_t) = \mathcal{N}(\mathbf{z}_t|\boldsymbol{\mu}_{0,t}, \boldsymbol{\Sigma}_{0,t}), \quad (3)$$

where $\boldsymbol{\mu}_{0,t}$ and $\boldsymbol{\Sigma}_{0,t}$ define the parameters of the conditional prior distribution, and they are obtained as the output of a deep neural network (DNN) $\varphi_\omega^{\text{prior}}(\cdot)$ that extracts features from the past hidden state $\mathbf{h}_{t-1}$ and the current discrete state $\mathbf{s}_t$:

$$[\boldsymbol{\mu}_{0,t}, \boldsymbol{\Sigma}_{0,t}] = \varphi_\omega^{\text{prior}}(\mathbf{h}_{t-1}, \mathbf{s}_t), \quad (4)$$

where $\boldsymbol{\Sigma}_{0,t}$ is considered a diagonal matrix. The hidden state $\mathbf{h}_{t-1}$ encodes the information of the process $\mathbf{z}$ up to time $t-1$, and it is updated along time using an LSTM with the following state update recurrence

$$\mathbf{h}_{t-1} = f_{\boldsymbol{\tau}}(\mathbf{y}_{t-1}, \mathbf{h}_{t-2}), \quad (5)$$

where $\mathbf{y}_{t-1} = \varphi_\omega^{\mathbf{z}}(\mathbf{z}_{t-1})$ is the output of a DNN with input $\mathbf{z}_{t-1}$. We choose to work with LSTM [2] due to the ability to better cope with long sequences, but any other RNN architectures such as GRU [27] could be used. Besides, in order to prevent the exploding gradient problem that can arise in RNNs, we clip the gradients to 0.5.

Finally, for the discrete latent variable $\mathbf{s}_t$ we assume an informative time-independent prior:

$$p(\mathbf{s}_t) = \text{Categorical}(\mathbf{s}_t|\boldsymbol{\pi}), \quad (6)$$

where $\pi_k = 1/L$, where L is the number of components in the mixture.

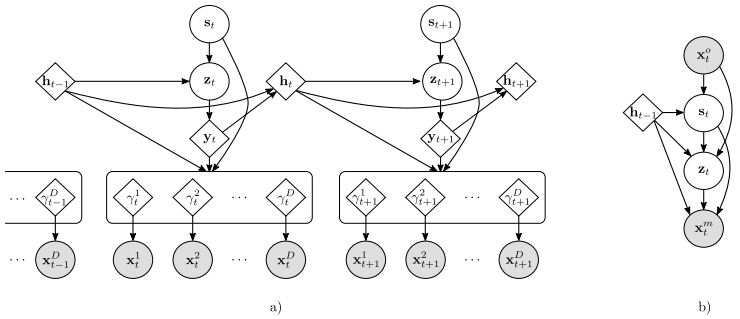

Fig. 3: On a), Shi-VAE generative model. On b), Shi-VAE inference model.

C. Heterogeneous Decoder

We propose to use a factorized decoder that can handle different data-types for each attribute. A DNN is used to provide the likelihood parameters, e.g. mean and variance of a Gaussian distribution, given $\mathbf{h}_{t-1}, \mathbf{s}_t$, and $\mathbf{y}_t$. We denote the likelihood parameters for the d-th attribute at time t as $\gamma_t^d = \varphi_{\omega,d}^{\text{dec}}(\mathbf{h}_{t-1}, \mathbf{s}_t, \mathbf{y}_t)$, where $\varphi_{\omega,d}^{\text{dec}}$ is the *decoder* DNN, as it translates latent information into the observed variable space. Hence, the general likelihood expression is:

$$p(x_{td}|\mathbf{z}_{\leq t}, \mathbf{s}_t) = p(x_{td}|\gamma_t^d) \tag{7}$$

We consider the following data-types and associated likelihood forms:

1) **Real-valued data**: We assume a Gaussian likelihood distribution, *i.e.*,

$$p(x_{td}|\gamma_t^d) = \mathcal{N}(\mu_{x,t}^d, \sigma_{x,t}^{2,d}),$$
$$\text{where } [\mu_{x,t}^d, \sigma_{x,t}^{2,d}] = \varphi_{\omega,d}^{\text{dec}}(\mathbf{y}_t, \mathbf{s}_t, \mathbf{h}_{t-1}) \tag{8}$$

2) **Positive real-valued data**: We assume a log-Gaussian likelihood distribution, *i.e.*,

$$p(x_{td}|\gamma_t^d) = \log\mathcal{N}(\mu_{x,t}^d, \sigma_{x,t}^{2,d}), \tag{9}$$
$$\text{where } [\mu_{x,t}^d, \sigma_{x,t}^{2,d}] = \varphi_{\omega,d}^{\text{dec}}(\mathbf{y}_t, \mathbf{s}_t, \mathbf{h}_{t-1})$$

3) **Binomial data**: We assume a Bernoulli likelihood distribution, *i.e.*,

$$p(x_{td}|\gamma_t^d) = Be(p_{x,t}^d), \tag{10}$$
$$\text{where } p_{x,t}^d = \sigma\big(\varphi_{\omega,d}^{\text{dec}}(\mathbf{y}_t, \mathbf{s}_t, \mathbf{h}_{t-1})\big),$$

and $p_{x,t}^d$ is the probability parameter of the Bernoulli distribution and σ is the sigmoid function.

4) **Categorical data**: We assume a multinomial likelihood distribution where the parameters of the likelihood are the C-dimensional output of a DNN with a log-softmax output

$$\log p(x_{td} = c|\gamma_t^d) = \varphi_{\omega,d}^{\text{dec}}(\mathbf{y}_t, \mathbf{s}_t, \mathbf{h}_{t-1})|_c \tag{11}$$

for $c = 1, \ldots, C$.

The left part of Figure 3 illustrates the generative model defined by Equations (1)-(7). From this figure we can see the motivation of having a shared latent space on $\mathbf{z}$ and $\mathbf{s}$ but an independenent heterogeneous decoder where each likelihood for x_t^d is parameterized by γ_t^d.

D. Model training with Variational Inference

Variational training [28] involves optimizing a parameterized family of distributions $q_\eta(\cdot)$ that approximate the latent posterior distribution given the observed data. This optimization is carried out by maximizing the well-known evidence lower bound (ELBO).

The variational distribution for our model is defined as $q_\phi(\mathbf{x}_{\leq T}^m, \mathbf{z}_{\leq T}, \mathbf{s}_{\leq T}|\mathbf{x}_{\leq T}^o)$ and it only depends on the observed attributes. Firstly, we need to define the variational distribution over the latent variable $\mathbf{z}_t$

$$q_{\phi_z}(\mathbf{z}_t|\mathbf{z}_{<t}, \mathbf{s}_t, \mathbf{x}_t^o) = \mathcal{N}(\boldsymbol{\mu}_{z,t}, \boldsymbol{\Sigma}_{z,t}), \tag{12}$$
$$\text{where } [\boldsymbol{\mu}_{z,t}, \boldsymbol{\Sigma}_{z,t}] = \varphi_\omega^{\text{enc}}(\varphi_\omega^{\mathbf{x}}(\tilde{\mathbf{x}}_t), \mathbf{h}_{t-1}, \mathbf{s}_t),$$

where $\tilde{\mathbf{x}}_t$ denotes a D-dimensional vector where the missing dimensions have been replaced by zeros following the zero filling approach as described in [6], $\boldsymbol{\mu}_{z,t}$ and $\boldsymbol{\Sigma}_{z,t}$ are the parameters of the variational distribution and $\varphi_\omega^{\mathbf{x}}$ and $\varphi_\omega^{\text{enc}}$ are neural networks. $\boldsymbol{\Sigma}_{z,t}$ is a diagonal matrix. The variational distribution for the discrete latent space $\mathbf{s}_t$ is defined as

$$q_{\phi_s}(\mathbf{s}_t|\mathbf{x}_t^o, \mathbf{z}_{<t}) = \text{Categorical}(\boldsymbol{\pi}(\varphi_\omega^{\mathbf{s}}(\tilde{\mathbf{x}}_t, \mathbf{h}_{t-1}))), \tag{13}$$

where the probability for each category is given by the output of the DNN $\varphi_\omega^{\mathbf{s}}(\cdot)$ followed by a log soft-max function. The variational distribution will then be composed of the variational distribution from Equation (12), the variational distribution from Equation (13) and $p(\mathbf{x}_t^m|\mathbf{z}_{\leq t}, \mathbf{s}_t)$, *i.e.*

$$q_\phi(\mathbf{x}_{\leq T}^m, \mathbf{z}_{\leq T}, \mathbf{s}_{\leq T}|\mathbf{x}_{\leq T}^o) = \prod_{t=1}^T q_{\phi_z}(\mathbf{z}_t|\mathbf{z}_{<t}, \mathbf{s}_t, \mathbf{x}_t^o) \tag{14}$$
$$q_{\phi_s}(\mathbf{s}_t|\mathbf{x}_t^o, \mathbf{z}_{<t})$$
$$p(\mathbf{x}_t^m|\mathbf{z}_{\leq t}, \mathbf{s}_t).$$

The inference model is shown at the right part of Figure 3. By expanding the following expression

$$\log p(\mathbf{X}^o) \geq \int q(\mathbf{X}^o, \mathbf{X}^m, \mathbf{Z}, \mathbf{S}) \log \frac{p(\mathbf{X}, \mathbf{Z}, \mathbf{S})}{q(\mathbf{X}^o, \mathbf{X}^m, \mathbf{Z}, \mathbf{S})} d\mathbf{Z} d\mathbf{S} d\mathbf{X}^m, \tag{15}$$

we obtain the ELBO objective training function defined in Equation (16). The first term inside the sum in Equation (16)

$$\log p(\mathbf{X}^o) \geq \sum_{t=1}^{T} \Bigg[\underbrace{\mathop{\mathbb{E}}_{\substack{q(\mathbf{s}_t|\mathbf{x}_t^o,\mathbf{z}_{<t},)\\ q(\mathbf{z}_t|\mathbf{x}_t^o,\mathbf{z}_{<t},\mathbf{s}_t)}} \left[\log p(\mathbf{x}_t^o|\mathbf{z}_{\leq t},\mathbf{s}_t)\right]}_{\text{Reconstruction}} - \underbrace{\mathop{\mathbb{E}}_{q(\mathbf{s}_t|\mathbf{x}_t^o,\mathbf{z}_{<t})} \left[\beta \mathrm{KL}\big(q(\mathbf{z}_t|\mathbf{z}_{<t},\mathbf{x}_t^o,\mathbf{s}_t)||p(\mathbf{z}_t|\mathbf{z}_{<t},\mathbf{s}_t)\big)\right] - \beta \mathrm{KL}(q(\mathbf{s}_t|\mathbf{x}_t^o,\mathbf{z}_{<t},)||p(\mathbf{s}_t))}_{\text{Regularization}} \Bigg] \quad (16)$$

is the average reconstruction log-likelihood (e.g. how well we explain the observed data given the latent space induced by the approximated posterior), while the other two Kullback-Leibler (KL) divergence terms act like regularizers that penalize for posteriors far from the prior latent distributions. Although the expectation over $q(\mathbf{s}_t|\mathbf{x}_t^o)$ can be computed analytically, since $\mathbf{s}_t$ is a discrete variable, due to the temporal dependencies encoded on the hidden state of the RNN $\mathbf{h}_t$ we approximate such expectations at low complexity by sampling from $q(\mathbf{s}_t|\mathbf{x}_t^o)$ using the Gumbel-softmax trick [29]. Finally, in Equation (16) β is a regularization parameter that we gradually increase during training, in a way the KL terms do not dominate over the reconstruction term during the earlier stages of training. Upon training, data is normalized. Standard-scaling is used for real attributes, and also to the logarithm of positive attributes. Categorical data is one-hot encoded.

E. The GP-VAE probabilistic model

As discussed in the introduction, GP-VAE [12] stands out as the state-of-the-art VAE to handle temporal series. Before addressing the experimental section, it is relevant to compare at this point the GP-VAE probabilistic model with respect to Shi-VAE. In GP-VAE, the latent temporal variable $\mathbf{z}_t$ is modeled with a Gaussian Process (GP) [30], i.e., $\mathbf{z}_t \sim \mathcal{GP}\left(m_z(\cdot), k_z(\cdot,\cdot)\right)$. The GP prior on the latent space is flexible and robust but it comes at the cost of inverting the kernel matrix, which has a time complexity of $\mathcal{O}(T^3)$. In contrast, the RNN-based correlation model in (3) comes with a computational cost that grows linearly in T. Moreover, designing a kernel function for GP-VAE that accurately captures correlations in feature space and also in the temporal dimension is challenging.

As in Shi-VAE, in GP-VAE given $\mathbf{z}_t$ all the attributes are conditionally independent. Indeed, the GP-VAE and its inference machinery [12] does not consider heterogeneous observations, and all observations are modelled with real-valued Gaussian distributions.

IV. Experimental Results

In this section we test the ability of Shi-VAE to exploit hidden correlations between attributes and infer trustworthy reconstructions in the presence of missing bursts. The following models are tested against Shi-VAE in the different experiments:

- **Mean**: We replace the missing values with the mean corresponding to the subsampled signal.
- **Last Obs Carried Forward (LOCF)**: We impute using the last observed value for a given attribute.
- **KNN**: We use k-nearest neighbor with normalized Euclidean distance to find similar samples, and then impute with a weighted average of the neighbors.
- **Matrix Factorization (MF)**: We subsample and factorize the data into two low-rank matrices and impute the missing entries with matrix completion [31].
- **MICE**: We use Multiple Imputation by Chained Equations (MICE), a very common method for missing value imputation which imputes those missing values from multiple imputations with chained equations [32].
- **GP-VAE**: The GP-VAE described in Section III-E.

We remark that both MF and MICE are "genie-aided" in the sense that they observe future values of the signal with-in a window to impute the results. The rest of the algorithms perform missing data imputation in an on-line fashion. Both GP-VAE and Shi-VAE reconstruct missing values by projecting the observed sequence to the latent space and then reconstruct the missing values using the generative model. The following python packages were used in order to implement the following methods: *fancyimpute* for Mean, KNN and MF; *autoimpute* for LOCF and *scikit-learn* for MICE [2].

We show results for three data sets. First, a synthetic data set generated by a heterogeneous HMM (Hidden Markov Model) with large hidden space, the human monitoring database described in Section II, and the well known medical data set Physionet [18]. While in the first database, the generated data set does not contain any missing data, note that both Physionet and the human monitoring database have quite a lot of missing observations. We evaluate performance over artificial missing data that we further incorporate into the data streams in all cases. We introduce missing sequences of random length for every variable to emulate missing bursts. A visual example can be seen in Figure 4. Each burst is generated sampling a random length from a uniform distribution $\mathcal{U}(3,10)$ and placing the burst in a random position given by an observed value. For every case (database and % introduced missing data), we create 10 random masks with a different missing pattern each, that we use to compute average errors and standard deviations around them. All masks implemented in the experiments are accessible in the code repository https://github.com/dbarrejon/Shi-VAE

We used the default setups for all the baselines model except for the GP-VAE, where we set the latent dimension to 2 in the synthetic data set, to 35 for Physionet and to 5 for the other two databases, since these values provided optimized results after cross-validation. The cross-validated parameter configuration for the Shi-VAE is described in Table II.

[2] autoimpute(0.12.1) [33], fancyimpute(0.5.5) [34], mice(0.23.2) [35]

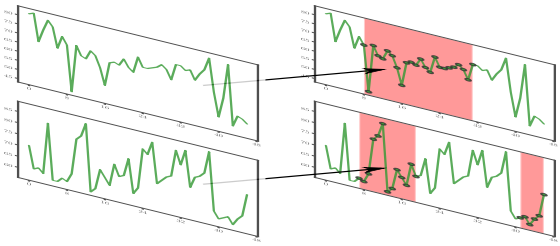

Fig. 4: Generation of missing artificial bursts for different sequences. The red masks and the corresponding missing entries (black markers) indicate the bursts of missing data.

Parameter	Synthetic	Physionet	Human Monitoring
Epochs	100	100	100
Annealing Epochs	20	20	50
Dimension $\mathbf{z}$	2	35	5
Dimension $\mathbf{h}$	10	10	10
L	3	10	3
T	100	48	-
Optimizer	Adam	Adam	Adam
Learning Rate	$5e-3$	$5e-3$	$5e-3$
Activation Layers	ReLU	ReLU	ReLU
Split Train/Val/Test	800/100/100	4K/4K/4K	135/15/17
Batch Size	64	64	64

TABLE II: Parameter configuration for the different experiments.

A. Evaluation Metrics

We use two different types of metrics to compare our models: standard error metrics and cross-correlation metrics between the ground-truth sequence and the reconstructed one. Before presenting the evaluation metrics, we will introduce some basic notation. Let us define $\mathbf{X}_d$ as a $N \times T$ matrix where we compact the d-th attribute across all data points and time. This is the matrix before introducing the artificial missing bursts. The imputed matrix for such attribute is defined as $\hat{\mathbf{X}}_d$ (equal to $\mathbf{X}_d$ for non-missing entries). Therefore, x_{td}^n is the entry at time t and data point n of $\mathbf{X}_d$. N_d is the number of missing entries in $\mathbf{X}_d$.

1) Error metrics: : We use a different type of error depending on the type of data:

- **Continuous data**, *i.e.* real and positive: we consider the normalized root mean squared error (NRMSE) evaluated only at missing entries

$$err(d) = \frac{\sqrt{1/N_d \sum_n \sum_t (x_{td}^n - \hat{x}_{td}^n)^2}}{\max(\mathbf{X}_d) - \min(\mathbf{X}_d)}. \quad (17)$$

- **Binary data and categorical data**: we consider the classification accuracy error evaluated at the missing entries.

$$err(d) = \frac{1}{N_d} \sum_n \sum_t I(x_{td}^n \neq \hat{x}_{td}^n), \quad (18)$$

where $I(\cdot)$ is the indicator function.

The average imputation error for all the attributes is given by Error $= 1/D \sum_d err(d)$, where D is the number of attributes.

2) Cross Correlation: : On temporal data sets, evaluating the performance of a given model based on standard error metrics might not be conclusive enough, as our experiments demonstrate. We augment our experiments by analyzing $\phi(d)$, which is defined as the sum of the cross correlation between any missing burst in $\mathbf{X}_d$ (a portion of a given row) and its corresponding imputation in $\hat{\mathbf{X}}_d$, normalized by the total number of missing entries N_d. To simplify notation, assume $\mathbf{w}$ and $\hat{\mathbf{w}}$ are the true and imputed values of a missing burst respectively in $\mathbf{X}_d$, then we accumulate in $c(\mathbf{w}, \hat{\mathbf{w}})$ the maximum value of the normalized cross correlation, i.e.

$$c(\mathbf{w}, \hat{\mathbf{w}}) = \max[(\mathbf{w} - \mu_{\mathbf{w}}) \star (\hat{\mathbf{w}} - \mu_{\hat{\mathbf{w}}})], \quad (19)$$

$\star$ is the cross correlation operator, and $\mu_{\mathbf{w}}$ is the average signal value during the burst.

Hence

$$\phi(d) = \frac{\sum_{\mathbf{w}, \hat{\mathbf{w}} \in \mathbf{X}_d} c(\mathbf{w}, \hat{\mathbf{w}})}{N_d} \quad (20)$$

We also report the average correlation across all attributes, *i.e.* Cross. Corr $= 1/D \sum_d \phi(d)$.

B. Synthetic Data set

This data set is composed of $N = 1000$ samples of length $T = 100$ from a three-state HMM model. At each time instant the HMM produces four outputs of different nature: real, positive, binary and categorical. Each state is characterized by different emission distribution for each data type. The transition probabilities have been forced to be smooth, so that really abrupt changes are not likely to happen. Over the clean database, we generate missing masks with overall missing rates of 10%, 30% and 50%. For all the baselines, including the GP-VAE, we work with subsampled slots of length $T = 50$ of every individual signal. For the Shi-VAE, we consider the whole signal.

In Figure 5 we display both reconstruction errors per attribute at different missing rates (a) and cross correlation for the real and positive attributes (b). In terms of reconstruction error, GP-VAE obtains the best results for the continuous variables by a small margin compared to Shi-VAE. This is due to the GP-VAE assuming a fully Gaussian distribution. However, for the binary and categorical view their performance is the same. Their distance with respect to the other baselines is remarkable. On the other hand, in terms of cross correlation, observe in Figure 5 (b) that Shi-VAE is able to reconstruct signals that are more correlated to the true distribution of the data. This raises an important question on how temporal models that are explicitly designed to impute missing values should be analyzed, whether it is more important to just focus on standard error metrics, or metrics considering temporal dependencies should be used when assessing the validity of temporal models.

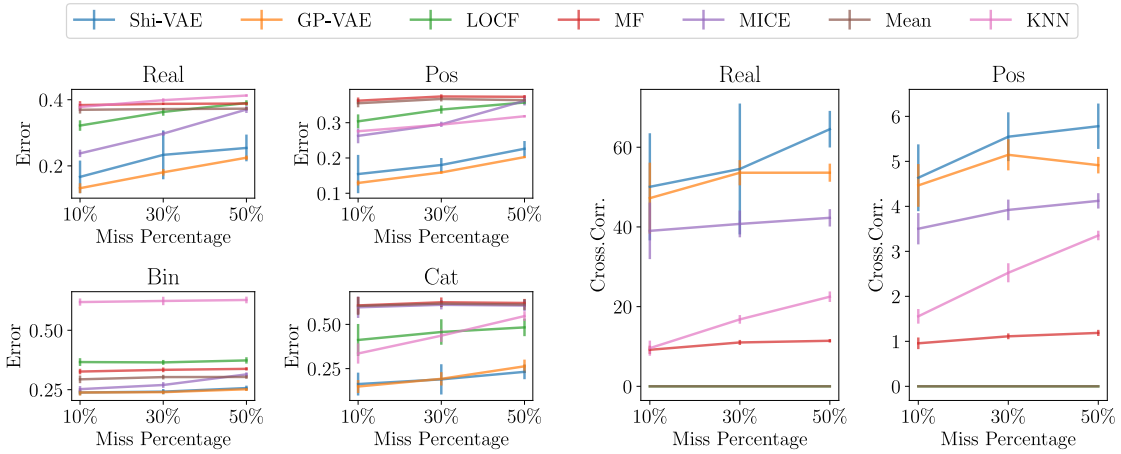

Fig. 5: Results for the synthetic data set. On the left part we show the imputation error for each variable and on the right part the cross correlation for the continuous variables at different missing rates.

Model	Avg. Error	Cross. Corr
Shi-VAE	**0.064 ± 0.003**	**38.061 ± 5.000**
GP-VAE	**0.060 ± 0.002**	31.414 ± 1.016

TABLE III: Physionet database results on the test set. For average error, lower is better. For cross correlation, larger is better.

C. Physionet

In this section, we compare both GP-VAE and Shi-VAE over the Physionet database [18]. The data set contains 12,000 patients which were monitored on the intensive care unit (ICU) for 48 hours each. Each signal is sampled once an hour, hence their length is $T = 48$. At each hour, there is a measurement of 35 different variables[3] (heart rate, blood pressure, etc.), any number of which might be missing. We further introduce artificial burst missing data up to an overall fraction of 10%. Note that the dataset already contains a large fraction of missing values.

In Table III we report GP-VAE and Shi-VAE average reconstruction error and average cross correlation. Observe that, as in the previous case, GP-VAE slightly improves the Shi-VAE in terms of average imputation error. However, Shi-VAE achieves a larger cross-correlation with respect to the ground-truth. To illustrate why reconstruction error can be a missleading metric when it comes to missing bursts, in the first row of Figure 6 we display the imputation of both methods for different missing bursts located at different Physionet attributes. Missing values are indicated by markers in the true signal. Observe that, while GP-VAE tends to impute missing burst with smooth solutions, Shi-VAE imputations certainly follow the true dynamics of the signal. But this discrepancy is not reflected in the average reconstruction error. In addition, observe that the Shi-VAE uncertainty (shaded area around the imputed signal) is informative and varies along time, allowing to identify regions of large and small uncertainty. On the other hand, the GP-VAE uncertainty does not show such a desired behaviour.

Similar conclusions can be drawn from the next related experiment. In Figure 7 we show one signal from the Physionet dataset, the real signal in green, the imputation from the Shi-VAE in blue and the imputation from the GP-VAE in orange. The first column on the right of each signal shows the distribution of the data with a histogram for the observed values of the signal, and the second column the distribution for the missing values (shown with black markers on the plot). The first row corresponds to the histogram that Shi-VAE produces, obtained by sampling from the model at each point. In the second row we do the same for the GP-VAE. We use the average of 10 samples produced by the models for the results. Below each histogram we show the corresponding average RMSE between the real samples and the imputed samples for each model. Observe that, while GP-VAE struggles to fit the real distribution even in the observed values, Shi-VAE provides a reasonably better result, being able to fit the two modes of the real distribution. This issue is not clearly reflected in the RMSE metric, which is not indeed very different between both models. On the contrary, the temporal correlation metric clearly shows the superior performance of Shi-VAE.

D. Human Monitoring Database

Finally, we reproduce the experiment for the human monitoring database described in Section II. The average fraction of artificially introduced missing rate per attribute is 15%. In this case the length of the temporal sequences for each patient is different. For Shi-VAE and GP-VAE, we pad with zeros to the right those sequences with a length smaller than the maximum sequence length in a batch. As described in Section III-E, the GP-VAE complexity badly scales with the sequence length. To run GP-VAE in reasonably time, any sequence larger than 50

[3]The list and definition of the attributes can be found in [18].

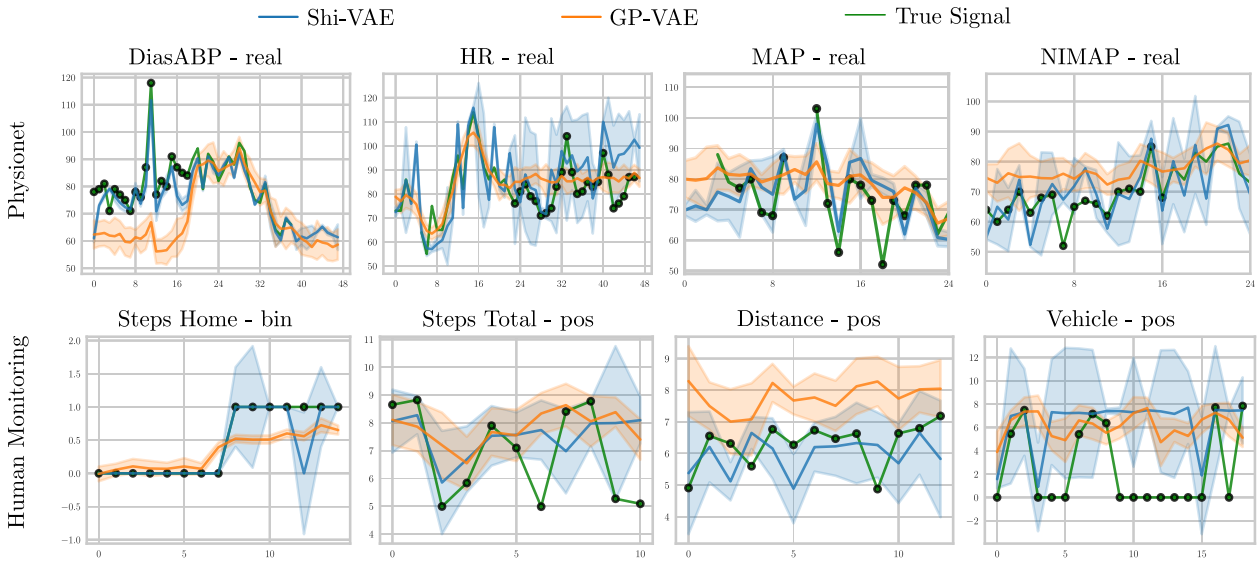

Fig. 6: Shi-VAE and GP-VAE example reconstruction for different attributes over the Physionet dataset (upper row) and the human monitoring database (bottom row). Missing values are indicated by markers in the true signal.

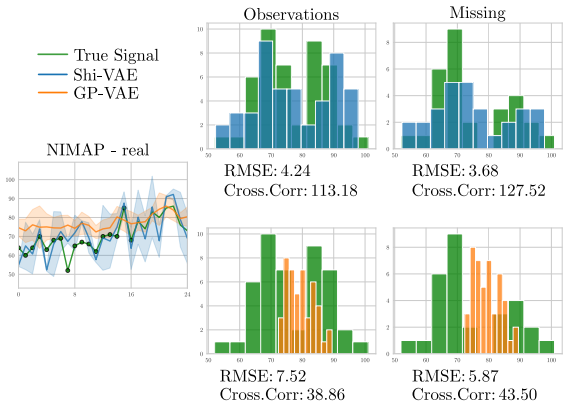

Fig. 7: Comparison of Shi-VAE and GP-VAE using histograms and the evaluation metrics on missing and observed data. For RMSE (lower is better) and for cross-correlation metric (larger is better).

Variable	Model	Error	Cross Correlation
Average	Shi-VAE	**0.200 ± 0.038**	**0.369± 0.140**
	GP-VAE	**0.184 ± 0.022**	0.157 ± 0.031
Distance	Shi-VAE	**0.201 ± 0.012**	**0.783 ± 0.249**
	GP-VAE	**0.205 ± 0.014**	0.389 ± 0.092
Steps home	Shi-VAE	**0.170 ± 0.054**	**0.010 ± 0.009**
	GP-VAE	**0.151 ± 0.016**	**0.011 ± 0.009**
Steps total	Shi-VAE	**0.269 ± 0.046**	**0.444 ± 0.181**
	GP-VAE	**0.268 ± 0.044**	0.205 ± 0.038
App usage	Shi-VAE	**0.113 ± 0.014**	**0.088 ± 0.045**
	GP-VAE	**0.115 ± 0.013**	**0.039 ± 0.008**
Sport	Shi-VAE	**0.216 ± 0.086**	**0.013 ± 0.005**
	GP-VAE	**0.121 ± 0.030**	**0.009 ± 0.004**
Sleep	Shi-VAE	**0.063 ± 0.010**	**0.034 ± 0.016**
	GP-VAE	**0.059 ± 0.010**	0.013 ± 0.003
Vehicle	Shi-VAE	**0.372 ± 0.043**	**1.215 ± 0.477**
	GP-VAE	**0.370 ± 0.028**	0.436 ± 0.064

TABLE IV: Results for each variable for the human monitoring data set.

time steps is subsampled to fit this maximum length. Note that Shi-VAE does not suffer from such penalization with respect to sequence length.

In Table IV, we report the error and cross correlation per attribute (seven of them, as described in Table I), and the overall average values. Observe that, systematically, Shi-VAE achieves the largest correlation per attribute. In the second row of Figure 6 we show the imputation of both methods for different missing bursts located at different attributes. The robustness of the Shi-VAE can be observed in terms of the correlation between the imputed signal and the true one and in terms of the uncertainty along time, which tends to be larger for those points in time for which the Shi-VAE mode is far from the true value. Again, such a behavior is not provided by GP-VAE.

V. Discussion

In this work we propose Shi-VAE, a deep generative model that handles temporal and heterogeneous streams of data in the presence of missing data. While GP-VAE badly scales with long time series, Shi-VAE handles long term dependencies by encapsulating the temporal information into the continuous latent code $\mathbf{z}$ by using RNN architectures. Having a hierarchi-

cal latent model with an additional discrete latent embedding s provides a more flexible understanding of the data and benefits the latter process of modeling the heterogeneous distributions.

We have shown with a synthetic data set and two real-world medical data sets that standard error metrics are not completely informative to fully assess the performance of temporal models. We remark the importance of analyzing the temporal correlation in these type of studies by using sequences of missing data along time instead of fully random missing masks as it is normally done in similar works. In this scenario, Shi-VAE emerges as a robust solution to impute missing data bursts and perform dimensionality reduction.

REFERENCES

[1] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to diagnose with LSTM recurrent neural networks," in *Proc. ICLR*, 2016, pp. 1–18.

[2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[3] Z. C. Lipton, D. C. Kale, R. Wetzel *et al.*, "Modeling missing data in clinical time series with rnns," *Machine Learning for Healthcare*, vol. 56, 2016.

[4] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific reports*, vol. 8, no. 1, pp. 1–12, 2018.

[5] W. Cao, D. Wang, J. Li, H. Zhou, L. Li, and Y. Li, "Brits: Bidirectional recurrent imputation for time series," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/734e6bfcd358e25ac1db0a4241b95651-Paper.pdf

[6] A. Nazabal, P. M. Olmos, Z. Ghahramani, and I. Valera, "Handling incomplete heterogeneous data using VAEs," *Pattern Recognition*, p. 107501, 2020.

[7] C. Ma, S. Tschiatschek, R. Turner, J. M. Hernández-Lobato, and C. Zhang, "VAEM: a deep generative model for heterogeneous mixed type data," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 11 237–11 247. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/8171ac2c5544a5cb54ac0f38bf477af4-Paper.pdf

[8] P.-A. Mattei and J. Frellsen, "Miwae: Deep generative modelling and imputation of incomplete data sets," in *International Conference on Machine Learning*, 2019, pp. 4413–4423.

[9] C. Ma, S. Tschiatschek, K. Palla, J. M. Hernández-Lobato, S. Nowozin, and C. Zhang, "Eddi: Efficient dynamic discovery of high-value information with partial VAE," *arXiv preprint arXiv:1809.11142*, 2018.

[10] M. Collier, A. Nazabal, and C. K.I. Williams, "VAEs in the presence of missing data," in *International Conference on Machine Learning 2020 Workshop Art of Learning with Mising Values (Artemiss)*.

[11] Y. L. Qiu, H. Zheng, and O. Gevaert, "Genomic data imputation with variational auto-encoders," *GigaScience*, vol. 9, no. 8, 08 2020.

[12] V. Fortuin, D. Baranchuk, G. Rätsch, and S. Mandt, "GP-VAE: Deep probabilistic time series imputation," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 1651–1661.

[13] J. Yoon, J. Jordon, and M. van der Schaar, "GAIN: Missing data imputation using generative adversarial nets," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 80. PMLR, 2018, pp. 5689–5698.

[14] Y. Luo, X. Cai, Y. ZHANG, J. Xu, and Y. xiaojie, "Multivariate time series imputation with generative adversarial networks," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/96b9bff013acedfb1d140579e2fbeb63-Paper.pdf

[15] S. C.-X. Li, B. Jiang, and B. Marlin, "Misgan: Learning from incomplete data with generative adversarial networks," in *International Conference on Learning Representations*, 2019.

[16] C. Shang, A. Palmer, J. Sun, K.-S. Chen, J. Lu, and J. Bi, "Vigan: Missing view imputation with generative adversarial networks," in *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 2017, pp. 766–775.

[17] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015. [Online]. Available: https://proceedings.neurips.cc/paper/2015/file/b618c3210e934362ac261db280128c22-Paper.pdf

[18] I. Silva, G. Moody, D. J. Scott, L. A. Celi, and R. G. Mark, "Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012," in *2012 Computing in Cardiology*. IEEE, 2012, pp. 245–248.

[19] R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong, "Moodscope: Building a mood sensor from smartphone usage patterns," in *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, 2013, pp. 389–402.

[20] A. Mehrotra, F. Tsapeli, R. Hendley, and M. Musolesi, "Mytraces: Investigating correlation and causation between users' emotional states and mobile phone interaction," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, pp. 1–21, 2017.

[21] L. Canzian and M. Musolesi, "Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis," in *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, 2015, pp. 1293–1304.

[22] M. De Choudhury, S. Counts, and E. Horvitz, "Predicting postpartum changes in emotion and behavior via social media," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2013, pp. 3267–3276.

[23] H. Lu, D. Frauendorfer, M. Rabbi, M. S. Mast, G. T. Chittaranjan, A. T. Campbell, D. Gatica-Perez, and T. Choudhury, "Stresssense: Detecting stress in unconstrained acoustic environments using smartphones," in *Proceedings of the 2012 ACM conference on ubiquitous computing*, 2012, pp. 351–360.

[24] "Evidence base behavior (eb2)," https://eb2.tech/.

[25] N. Dilokthanakul, P. A. M. Mediano, M. Garnelo, M. C. H. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan, "Deep unsupervised clustering with gaussian mixture variational autoencoders." *CoRR*, vol. abs/1611.02648, 2016.

[26] J. Tomczak and M. Welling, "VAE with a vampprior," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2018, pp. 1214–1223.

[27] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *CoRR*, vol. abs/1406.1078, 2014. [Online]. Available: http://arxiv.org/abs/1406.1078

[28] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[29] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," 2016, cite arxiv:1611.01144. [Online]. Available: http://arxiv.org/abs/1611.01144

[30] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. MIT Press, 2006.

[31] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[32] I. R. White, P. Royston, and A. M. Wood, "Multiple imputation using chained equations: issues and guidance for practice," *Statistics in medicine*, vol. 30, no. 4, pp. 377–399, 2011.

[33] "Autoimpute." [Online]. Available: https://kearnz.github.io/autoimpute-tutorials/

[34] A. Rubinsteyn and S. Feldman, "fancyimpute: An imputation library for python." [Online]. Available: https://github.com/iskandr/fancyimpute

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.