GP-VAE: Deep Probabilistic Time Series Imputation

Vincent Fortuin[*†1], Dmitry Baranchuk[†2], Gunnar Rätsch[1], and Stephan Mandt[3]

[1]Department of Computer Science, ETH Zürich, Zürich, Switzerland
[2]Yandex, Moscow, Russia
[3]Donald Bren School of Information and Computer Sciences, UC Irvine, California, USA

Abstract

Multivariate time series with missing values are common in areas such as healthcare and finance, and have grown in number and complexity over the years. This raises the question whether deep learning methodologies can outperform classical data imputation methods in this domain. However, naïve applications of deep learning fall short in giving reliable confidence estimates and lack interpretability. We propose a new deep sequential latent variable model for dimensionality reduction and data imputation. Our modeling assumption is simple and interpretable: the high dimensional time series has a lower-dimensional representation which evolves smoothly in time according to a Gaussian process. The non-linear dimensionality reduction in the presence of missing data is achieved using a VAE approach with a novel structured variational approximation. We demonstrate that our approach outperforms several classical and deep learning-based data imputation methods on high-dimensional data from the domains of computer vision and healthcare, while additionally improving the smoothness of the imputations and providing interpretable uncertainty estimates.

1 Introduction

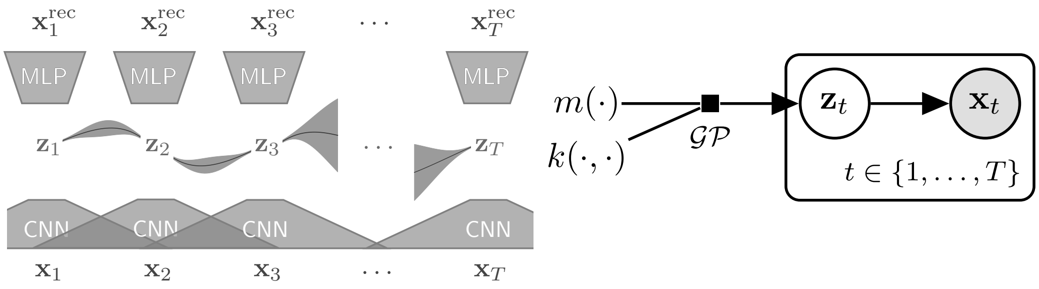
Time series are often associated with missing values, for instance due to faulty measurement devices, partially observed states, or costly measurement procedures [18]. These missing values impair the usefulness and interpretability of the data, leading to the problem of *data imputation*: estimating those missing values from the observed ones [42].

Multivariate time series, consisting of multiple correlated univariate time series or *channels*, give rise to two distinct ways of imputing missing information: (1) by exploiting temporal correlations within each channel, and (2) by exploiting correlations across channels, for example by using lower-dimensional representations of the data. For instance in a medical setting, if the blood pressure of a patient is unobserved, it can be informative that the heart rate at the current time is higher than normal and that the blood pressure was also elevated an hour ago. An ideal imputation model for multivariate time series should therefore take both of these sources of information into account. Another desirable property of such models is to offer a probabilistic interpretation, allowing for uncertainty estimation.

Unfortunately, current imputation approaches fall short with respect to at least one of these desiderata. While there are many time-tested statistical methods for multivariate time series analysis (e.g., Gaussian processes [41]) that work well in the case of complete data, these methods are generally not applicable when features are missing. On the other hand, classical methods for time series imputation often do not take the potentially complex interactions between the different channels into account [33, 38]. Finally, recent work has explored the use of non-linear dimensionality reduction using variational autoencoders for i.i.d. data points with missing values [1, 35, 37] , but this work has not considered temporal data and strategies for sharing statistical strength across time.

[*] fortuin@inf.ethz.ch

[†] Equal contribution.



(a) Architecture sketch

(b) Graphical model

Figure 1: Overview of our proposed model with a convolutional inference network, a deep feed-forward generative network and a Gaussian process prior with mean function $m(\cdot)$ and kernel function $k(\cdot,\cdot)$ in latent space. The several CNN blocks as well as the MLP blocks are each sharing their respective parameters.

Following these considerations, it is promising to combine non-linear dimensionality reduction with an expressive time series model. This can be done by jointly learning a mapping from the data space (where features are missing) into a latent space (where all dimensions are fully determined). A statistical model of choice can then be applied in this latent space to model temporal dynamics. If the dynamics model and the mapping for dimensionality reduction are both differentiable, the approach can be trained end-to-end.

In this paper, we propose an architecture that uses deep variational autoencoders (VAEs) to map the missing data time series into a latent space without missingness, where we model the low-dimensional dynamics with a Gaussian process (GP). As we will discuss below, we hereby propose a prior model that efficiently operates at multiple time scales, taking into account that the multivariate time series may have different channels (e.g., heart rate, blood pressure, etc.) that change with different characteristic frequencies. Finally, our variational inference approach makes use of efficient structured variational approximations, where we fit another multivariate Gaussian process in order to approximate the intractable true posterior.

We make the following contributions:

- A new model. We propose a VAE architecture for multivariate time series imputation with a GP prior in the latent space to capture temporal dynamics. We propose a Cauchy kernel to allow the time series to display dynamics at multiple scales in a reduced dimensionality.
- Efficient inference. We use a structured variational approximation that models posterior correlations in the time domain. By construction, inference is efficient and the time complexity for sampling from the variational distribution, used for training, is linear in the number of time steps (as opposed to cubic when done naïvely).
- Benchmarking on real-world data. We carry out extensive comparisons to classical imputation methods as well as state-of-the-art deep learning approaches, and perform experiments on data from two different domains. Our method shows favorable performance in both cases.

We start by reviewing the related literature in Sec. 2, describe the general setting in Sec. 3.1 and introduce our model and inference scheme in Sec. 3.2 and Sec. 3.3, respectively. Experiments and conclusions are presented in Sec. 4 and 5.

2 Related work

Classical statistical approaches. The problem of missing values has been a long-standing challenge in many time series applications, especially in the field of medicine [38]. The earliest approaches to deal with this problem often relied on heuristics, such as mean imputation or forward imputation. Despite their simplicity, these methods are still widely applied today due to their efficiency and interpretability [18]. Orthogonal to these ideas, methods along the lines of expectation-maximization (EM) have been proposed, but they often require additional modeling assumptions [4].

Bayesian methods. When it comes to estimating likelihoods and uncertainties relating to the imputations, Bayesian methods, such as Gaussian processes (GPs) [39], have a clear advantage over non-Bayesian methods such as single imputation [33]. There has been much recent work in making these methods more expressive and incorporating prior knowledge from the domain (e.g., medical time series) [12, 46] or adapting them to work on discrete domains [13], but their wide-spread adoption is hindered by their limited scalability and the challenges in designing kernels that are robust to missing values. Our latent GP prior bears certain similarities to the GP latent variable model (GP-LVM) [28, 44], but in contrast to this line of work, we propose an efficient amortized inference scheme.

Deep learning techniques. Another avenue of research in this area uses deep learning techniques, such as variational autoencoders (VAEs) [1, 10, 35, 37] or generative adversarial networks (GANs) [30, 47]. It should be noted that VAEs allow for tractable likelihoods, while GANs generally do not and have to rely on additional optimization processes to find latent representations of a given input [32]. Unfortunately, none of these models explicitly take the temporal dynamics of time series data into account. Conversely, there are deep probabilistic models for time series [e.g., 14, 26, 27], but those do not explicitly handle missing data. There are also some VAE-based imputation methods that are designed for a setting where the data is complete at training time and the missingness only occurs at test time [15, 16, 20]. We do not regard this setting in our work.

HI-VAE. Our approach borrows some ideas from the HI-VAE [37]. This model deals with missing data by defining an ELBO whose reconstruction error term only sums over the observed part of the data. For inference, the incomplete data are filled with arbitrary values (e.g., zeros) before they are fed into the inference network, which induces an unavoidable bias. The main difference to our approach is that the HI-VAE was not formulated for sequential data and therefore does not exploit temporal information in the imputation task.

Deep learning for time series imputation. While the mentioned deep learning approaches are very promising, most of them do not take the time series nature of the data directly into account, that is, they do not model the temporal dynamics of the data when dealing with missing values. To the best of our knowledge, the only deep generative model for missing value imputation that does account for the time series nature of the data is the GRUI-GAN [34], which we describe in Sec. 4. Another deep learning model for time series imputation is BRITS [7], which uses recurrent neural networks (RNNs). It is trained in a self-supervised way, predicting the observations in a time series sequentially. We compare against both of these models in our experiments.

Other related work. Our proposed model combines several ideas from the domains of Bayesian deep learning and classical probabilistic modeling; thus, removing elements from our model naturally relates to other approaches. For example, removing the latent GP for modeling dynamics as well as our proposed structured variational distribution results in the HI-VAE [37] described above. Furthermore, our idea of using a latent GP in the context of a deep generative model bears similarities to the GPPVAE [9] and can be seen as an extension of this model. While the GPPVAE allows for a joint GP prior over the whole data set through the use of a specialized inference mechanism, our model puts a separate GP prior on every time series and can hence rely on more standard inference techniques. However, our model takes missingness into account and uses a structured variational distribution, while the GPPVAE uses mean field inference and is designed for fully observed data. Lastly, the GP prior with the Cauchy kernel is reminiscent of Jähnichen et al. [21] and the structured variational distribution is similar to the one used by Bamler and Mandt [3] in the context of modeling word embeddings over time, neither of which used amortized inference.

3 Model

We propose a novel architecture for missing value imputation, an overview of which is depicted in Figure 1. Our model can be seen as a way to perform amortized approximate inference on a latent Gaussian process model.

The main idea of our proposed approach is to embed the data into a latent space of reduced dimensionality, in which every dimension is fully determined, and then model the temporal dynamics in this latent space. Since many features in the data might be correlated, the latent representation captures these correlations and uses them to reconstruct the missing values. Moreover, the GP prior in the latent space encourages the model to embed the data into a representation in which the temporal dynamics are smoother and more easily explainable than in the original data space. Finally, the structured variational distribution of the inference network allows the model to integrate temporal information into the representations, such that the reconstructions of missing values cannot only be informed by correlated observed features at the same time point, but also by the rest of time series.

Specifically, we combine ideas from VAEs [25], GPs [39], Cauchy kernels [21], structured variational distributions with efficient inference [3], and a special ELBO for missing data [37] and synthesize these ideas into a general framework for missing data imputation on time series. In the following, we will outline the problem setting, describe the assumed generative model, and derive our proposed inference scheme.

3.1 Problem setting and notation

We assume a data set $\mathbf{X} \in \mathbb{R}^{T \times d}$ with T data points $\mathbf{x}_t = [x_{t1}, \dots, x_{tj}, \dots, x_{td}]^\top \in \mathbb{R}^d$. Let us assume that the T data points were measured at T consecutive time points $\tau = [\tau_1, \dots, \tau_T]^\top$ with $\tau_t < \tau_{t+1} \, \forall t$. By convention, we usually set $\tau_1 = 0$. The data $\mathbf{X}$ can thus be viewed as a time series of length τ_T in time.

We moreover assume that any number of these data features x_{tj} can be missing, that is, that their values can be unknown. We can now partition each data point into observed and unobserved features. The observed features of data point $\mathbf{x}_t$ are $\mathbf{x}_t^o := [x_{tj} \mid x_{tj} \text{ is observed}]$. Equivalently, the missing features are $\mathbf{x}_t^m := [x_{tj} \mid x_{tj} \text{ is missing}]$ with $\mathbf{x}_t^o \cup \mathbf{x}_t^m \equiv \mathbf{x}_t$.

We can now use this partitioning to define the problem of missing value imputation. Missing value imputation describes the problem of estimating the true values of the missing features $\mathbf{X}^m := [\mathbf{x}_t^m]_{1:T}$ given the observed features $\mathbf{X}^o := [\mathbf{x}_t^o]_{1:T}$. Many methods assume the different data points to be independent, in which case the inference problem reduces to T separate problems of estimating $p(\mathbf{x}_t^m \mid \mathbf{x}_t^o)$. In the time series setting, this independence assumption is not satisfied, which leads to the more complex estimation problem of $p(\mathbf{x}_t^m \mid \mathbf{x}_{1:T}^o)$.

3.2 Generative model

In this subsection, we describe the details of our proposed approach of reducing the observed time series with missing data into a representation without missingness, and modeling dynamics in this lower-dimensional representation using Gaussian processes. Yet, it is tempting to try to skip the step of dimensionality reduction and instead directly try to model the incomplete data in the observed space using GPs. We argue that this is not practical for several reasons.

Gaussian processes are well suited for time series modeling [41] and offer many advantages, such as data-efficiency and calibrated posterior probabilities. However, they come at the cost of inverting the kernel matrix, which has a time complexity of $\mathcal{O}(n^3)$. Moreover, designing a kernel function that accurately captures correlations in feature space and also in the temporal dimension is difficult.

This problem becomes even worse if certain observations are missing. One option is to fill the missing values with some numerical value (e.g., zero) to make the kernel computable. However, this arbitrary filling may make two data points with different missingness patterns look very dissimilar when in fact they are close to each other in the ground-truth space. Another alternative is to treat every channel of the multivariate time series separately and let the GP infer missing values, but this ignores valuable correlations across channels.

In this work, we overcome the problem of defining a suitable GP kernel in the data space with missing observations by instead applying the GP in the latent space of a variational autoencoder where the encoded feature representations are complete. That is, we assign a latent variable $\mathbf{z}_t \in \mathbb{R}^k$ for every $\mathbf{x}_t$, and model temporal correlations in this reduced representation using a GP, $\mathbf{z}(\tau) \sim \mathcal{GP}(m_z(\cdot), k_z(\cdot,\cdot))$. This way, we decouple the step of filling in missing values and capturing instantaneous correlations between the different feature dimensions from modeling dynamical aspects. The graphical model is depicted in Figure 1b.

A remaining practical difficulty that we encountered is that many multivariate time series display dynamics at multiple time scales. One of our main motivations is to model time series that arise in medical setups where doctors measure different patient variables and vital signs, such as heart rate, blood pressure, etc. When using conventional GP kernels (e.g., the RBF kernel, $k_{RBF}(r \,|\, \lambda) = \exp\left(-\lambda r^2/2\right)$), one implicitly assumes a single time scale of relevance (λ). We found that this choice does not reflect the dynamics of medical data very well.

In order to model data that varies at multiple time scales, we consider a mixture of RBF kernels with different λ's [39]. By defining a Gamma distribution over the length scale, that is, $p(\lambda \,|\, \alpha, \beta) \propto \lambda^{\alpha-1} \exp\left(-\alpha\lambda/\beta\right)$, we can compute an infinite mixture of RBF kernels,

$$\int p(\lambda \,|\, \alpha, \beta)\, k_{RBF}(r \,|\, \lambda)\, d\lambda \propto \left(1 + \frac{r^2}{2\alpha\beta^{-1}}\right)^{-\alpha} .$$

This yields the so-called Rational Quadratic kernel [39]. For $\alpha = 1$ and $l^2 = 2\beta^{-1}$, it reduces to the Cauchy kernel

$$k_{Cau}(\tau, \tau') = \sigma^2 \left(1 + \frac{(\tau - \tau')^2}{l^2}\right)^{-1} , \tag{1}$$

which has previously been successfully used in the context of robust dynamic topic modeling where similar multi-scale time dynamics occur [21]. We therefore choose this kernel for our Gaussian process prior.

Given the latent time series $\mathbf{z}_{1:T}$, the observations $\mathbf{x}_t$ are generated time-point-wise by

$$p_\theta(\mathbf{x}_t \,|\, \mathbf{z}_t) = \mathcal{N}\left(g_\theta(\mathbf{z}_t), \sigma^2 \mathbf{I}\right) , \tag{2}$$

where $g_\theta(\cdot)$ is a potentially nonlinear function parameterized by the parameter vector θ. Considering the scenario of a medical time series, $\mathbf{z}_t$ can be thought of as the latent physiological state of the patient and $g_\theta(\cdot)$ would be the process of generating observable measurements $\mathbf{x}_t$ (e.g., heart rate, blood pressure, etc.) from that physiological state. In our experiments, the function g_θ is implemented by a deep neural network.

3.3 Inference model

In order to learn the parameters of the deep generative model described above, and in order to efficiently infer its latent state, we are interested in the posterior distribution $p(\mathbf{z}_{1:T} \,|\, \mathbf{x}_{1:T})$. Since the exact posterior is intractable, we use variational inference [6, 22, 48]. Furthermore, to avoid inference over per-datapoint (local) variational parameters, we apply inference amortization [25]. To make our variational distribution more expressive and capture the temporal correlations of the data, we employ a structured variational distribution [45] with efficient inference that leads to an approximate posterior which is also a GP.

We approximate the true posterior $p(\mathbf{z}_{1:T,j} \,|\, \mathbf{x}_{1:T})$ with a multivariate Gaussian variational distribution

$$q(\mathbf{z}_{1:T,j} \,|\, \mathbf{x}^o_{1:T}) = \mathcal{N}\left(\mathbf{m}_j, \mathbf{\Lambda}_j^{-1}\right) , \tag{3}$$

where j indexes the dimensions in the latent space. Our approximation implies that our variational posterior is able to reflect correlations in time, but breaks dependencies across the different dimensions in $\mathbf{z}$-space (which is typical in VAE training [25, 40]).

We choose the variational family to be the family of multivariate Gaussian distributions in the time domain, where the precision matrix $\mathbf{\Lambda}_j$ is parameterized in terms of a product of bidiagonal matrices,

$$\mathbf{\Lambda}_j := \mathbf{B}_j^\top \mathbf{B}_j \text{ , with } \{\mathbf{B}_j\}_{tt'} = \begin{cases} b_{tt'}^j & \text{if } t' \in \{t, t+1\} \\ 0 & \text{otherwise} \end{cases} . \tag{4}$$

Above, the $b_{tt'}^j$'s are local variational parameters and $\mathbf{B}_j$ is an upper triangular band matrix. Similar structured distributions were also employed by Bamler and Mandt [2], Blei and Lafferty [5].

This parameterization automatically leads to $\mathbf{\Lambda}_j$ being positive definite, symmetric, and tridiagonal. Samples from q can thus be generated in linear time in T [3, 19, 36] as opposed to the cubic time complexity for a full-rank matrix. Moreover, compared to a fully factorized variational approximation, the number of variational parameters are merely doubled. Note that while the precision matrix is sparse, the covariance matrix can still be dense, allowing to reflect long-range dependencies in time.

Instead of optimizing $\mathbf{m}$ and $\mathbf{B}$ separately for every data point, we amortize the inference through an inference network with parameters ψ that computes the variational parameters based on the inputs as $(\mathbf{m}, \mathbf{B}) = h_\psi(\mathbf{x}_{1:T}^o)$. In the following, we accordingly denote the variational distribution as $q_\psi(\cdot)$. Following VAE training, the parameters of the generative model θ and inference network ψ can be jointly trained by optimizing the evidence lower bound (ELBO),

$$\begin{aligned} \log p(\mathbf{X}^o) \geq & \sum_{t=1}^{T} \mathbb{E}_{q_\psi(\mathbf{z}_t \,|\, \mathbf{x}_{1:T})} \left[\log\, p_\theta(\mathbf{x}_t^o \,|\, \mathbf{z}_t)\right] \\ & - \beta\, D_{KL}\left[q_\psi(\mathbf{z}_{1:T} \,|\, \mathbf{x}_{1:T}^o) \,\|\, p(\mathbf{z}_{1:T})\right] \end{aligned} \tag{5}$$

Following Nazabal et al. [37] (see Sec. 2), we evaluate the ELBO only on the observed features of the data since the remaining features are unknown, and set these missing features to a fixed value (zero) during inference. We also include an additional tradeoff parameter β into our ELBO, similar to the β-VAE [17]. This parameter can be used to rebalance the influence of the likelihood term and KL term on the ELBO. Since the likelihood factorizes into a sum over observed feature dimensions, its absolute magnitude depends on the missingness rate in the data. We thus choose β in our experiments dependent on the missingness rate to counteract this effect. Our training objective is the RHS of (5).

4 Experiments

We performed experiments on the benchmark data set *Healing MNIST* [26], which combines the classical MNIST data set [29] with properties common to medical time series, the SPRITES data set [31], and on a real-world medical data set from the 2012 Physionet Challenge [43]. We compared our model against conventional single imputation methods [33], GP-based imputation [39], VAE-based methods that are not specifically designed to handle temporal data [25, 37], and modern state-of-the-art deep learning methods for temporal data imputation [7, 34]. We found strong quantitative and qualitative evidence that our proposed model outperforms most of the baseline methods in terms of imputation quality on all three tasks and performs comparably to the state of the art, while also offering a probabilistic interpretation and uncertainty estimates. In the following, we are first going to give an overview of the baseline methods and then present our experimental findings. Details about the data sets and neural network architectures can be found in Appendix A. An implementation of our model can be retrieved from `https://github.com/ratschlab/GP-VAE`.

4.1 Baseline methods

Forward imputation and mean imputation. Forward and mean imputation are so-called single imputation methods, which do not attempt to fit a distribution over possible values for the missing features, but only predict one estimate [33]. Forward imputation always predicts the last observed value for any given feature, while mean imputation predicts the mean of all the observations of the feature in a given time series.

Gaussian process in data space. One option to deal with missingness in multivariate time series is to fit independent Gaussian processes to each channel. As discussed previously (Sec. 3.2), this ignores the correlation between channels. The missing values are then imputed by taking the mean of the respective posterior of the GP for that feature.

VAE and HI-VAE. The VAE [25] and HI-VAE [37] are fit to the data using the same training procedure as the proposed GP-VAE model. The VAE uses a standard ELBO that is defined over all the features, while the HI-VAE uses the ELBO from (5), which is only evaluated on the observed part of the feature space. During inference, missing features are filled with constant values, such as zero.

GRUI-GAN and BRITS. The GRUI-GAN [34] uses a recurrent neural network (RNN), namely a gated recurrent unit (GRU). Once the GAN is trained on a set of time series, an unseen time series is imputed by optimizing the latent vector in the input space of the generator, such that the generator's output on the observed features is closest to the true values. BRITS [7] also uses an RNN, namely a bidirectional long short-term memory network (BiLSTM). For a series of partially observed states, the network is trained to predict any intermediate state given its past and future states, aggregated by two separate LSTM layers. Similarly to our approach, the loss function is only computed on the observed values.

4.2 Healing MNIST

Time series with missing values play a crucial role in the medical field, but are often hard to obtain. Krishnan et al. [26] generated a data set called *Healing MNIST*, which is designed to reflect many properties that one also finds in real medical data. We benchmark our method on a variant of this data set. It was designed to incorporate some properties that one also finds in real medical data, and consists of short sequences of moving MNIST digits [29] that rotate randomly between frames. The analogy to healthcare is that every frame may represent the collection of measurements that describe a patient's health state, which contains many missing measurements at each moment in time. The temporal evolution represents the non-linear evolution of the patient's health state. The image frames contain around 60 % missing pixels and the rotations between two consecutive frames are normally distributed.

The benefit of this data set is that we know the ground truth of the imputation task. We compare our model against a standard VAE (no latent GP and standard ELBO over all features), the HI-VAE [37], as well as mean imputation and forward imputation. The models were trained on time series of digits from the Healing MNIST training set (50,000 time series) and tested on digits from the Healing MNIST test set (10,000 time series). Negative log likelihoods on the ground truth values of the missing pixels and mean squared errors (MSE) are reported in Table 1, and qualitative results shown in Figure 2. To assess the usefulness of the imputations for downstream tasks, we also trained a linear classifier on the imputed MNIST digits to predict the digit class and measured its performance in terms of area under the receiver-operator-characteristic curve (AUROC) (Tab. 1).

Our approach outperforms the baselines in terms of likelihood and MSE. The reconstructions (Fig. 2) reveal the benefits of the GP-VAE approach: related approaches yield unstable reconstructions over time, while our approach offers more stable reconstructions, using temporal information from neighboring frames. Moreover, our model also yields the most useful imputations for downstream classification in terms of AUROC. The downstream classification performance correlates well with the test likelihood on the ground truth data, supporting the intuition that it is a good proxy measure in cases where the ground truth likelihood is not available.

We also observe that our model outperforms the baselines on different missingness mechanisms (Tab. 2). Details regarding this evaluation are laid out in the appendix (Sec. B.2).

4.3 SPRITES data

To assess our model's performance on more complex data, we applied it to the *SPRITES* data set, which has previously been used with sequential autoencoders [31]. The dataset consists of 9,000 sequences of animated characters with different clothes, hair styles, and skin colors, performing

Table 1: Performance of the different models on the Healing MNIST test set and the SPRITES test set in terms of negative log likelihood [NLL] and mean squared error [MSE] (lower is better), as well as downstream classification performance [AUROC] (higher is better). The reported values are means and their respective standard errors over the test set. The proposed model outperforms all the baselines.

	Healing MNIST			SPRITES
Model	NLL	MSE	AUROC	MSE
Mean imputation [33]	-	0.168 ± 0.000	0.938 ± 0.000	0.013 ± 0.000
Forward imputation [33]	-	0.177 ± 0.000	0.935 ± 0.000	0.028 ± 0.000
VAE [25]	0.599 ± 0.002	0.232 ± 0.000	0.922 ± 0.000	0.028 ± 0.000
HI-VAE [37]	0.372 ± 0.008	0.134 ± 0.003	**0.962** ± **0.001**	0.007 ± 0.000
GP-VAE (proposed)	**0.350** ± **0.007**	**0.114** ± **0.002**	**0.960** ± **0.002**	**0.002** ± **0.000**

Table 2: Performance of different models on Healing MNIST data with artificial missingness and different missingness mechanisms. We report mean squared error (lower is better). The reported values are means and their respective standard errors over the test set. Our model outperforms the baselines on all missingness mechanisms.

Mechanism	Mean imp. [33]	Forward imp. [33]	VAE [25]	HI-VAE [37]	GP-VAE (proposed)
Random	0.069 ± 0.000	0.099 ± 0.000	0.100 ± 0.000	0.046 ± 0.001	**0.036** ± **0.000**
Spatial	0.090 ± 0.000	0.099 ± 0.000	0.122 ± 0.000	0.097 ± 0.000	**0.071** ± **0.001**
Temporal$^+$	0.107 ± 0.000	0.117 ± 0.000	0.101 ± 0.000	0.047 ± 0.001	**0.038** ± **0.001**
Temporal$^-$	0.086 ± 0.000	0.093 ± 0.000	0.101 ± 0.000	0.046 ± 0.001	**0.037** ± **0.001**
MNAR	0.168 ± 0.000	0.177 ± 0.000	0.232 ± 0.000	0.134 ± 0.003	**0.114** ± **0.002**

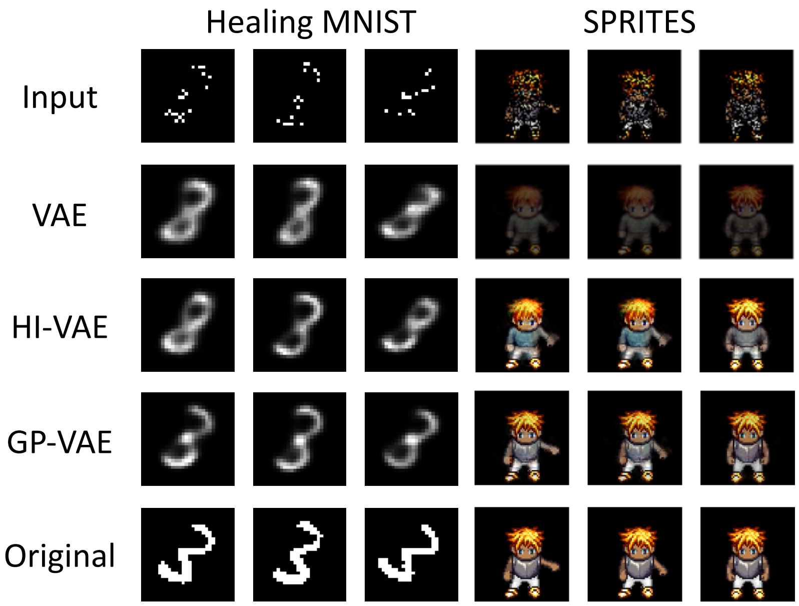


Figure 2: Reconstructions from Healing MNIST and SPRITES. The GP-VAE (proposed) is stable over time and yields the highest fidelity.

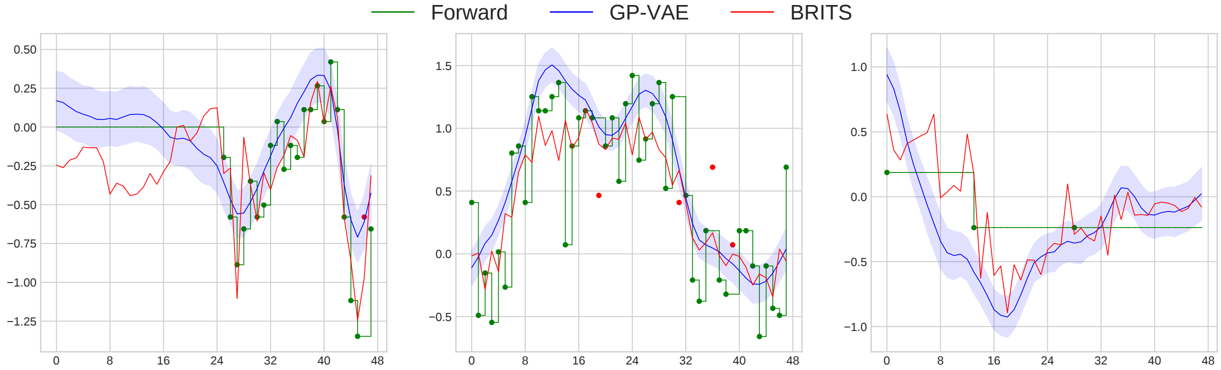


Figure 3: Imputations of several clinical variables with different amounts of missingness for an example patient from the Physionet 2012 test set. Blue dots are observed ground-truth points, red dots are ground-truth points that were withheld from the models. BRITS (red) and forward imputation (green) yield single imputations, while the GP-VAE (blue) allows to draw samples from the posterior. The GP-VAE produces smoother curves, reducing noise from the original input, and exhibits an interpretable posterior uncertainty.

different actions. Each frame has a size of 64 × 64 pixels and each time series features 8 frames. We again introduced about 60 % of missing pixels and compared the same methods as above. The results are reported in Table 1 and example reconstructions are shown in Figure 2. As in the previous experiment, our model outperforms the baselines and also yields the most convincing reconstructions.

4.4 Real medical time series data

We also applied our model to the data set from the 2012 Physionet Challenge [43]. The data set contains 12,000 patients which were monitored on the intensive care unit (ICU) for 48 hours each. At each hour, there is a measurement of 35 different variables (heart rate, blood pressure, etc.), any number of which might be missing.

We again compare our model against the standard VAE and HI-VAE, as well as a GP fit feature-wise in the data space, the GRUI-GAN [34] and BRITS model [7], which reported state-of-the-art imputation performance.

The main challenge is the absence of ground truth data for the missing values. This cannot easily be circumvented by introducing additional missingness since (1) the mechanism by which measurements were omitted is not random, and (2) the data set is already very sparse with about 80% of the features missing. To overcome this issue, Luo et al. [34] proposed a downstream task as a proxy for the imputation quality. They chose the task of mortality prediction, which was one of the main tasks of the Physionet Challenge on this data set, and measured the performance in terms of AUROC. In this paper, we adopt this measure.

For sake of interpretability, we used a logistic regression as a downstream classification model. This model tries to optimally separate the whole time series in the input space using a linear hyperplane. The choice of model follows the intuition that under a perfect imputation similar patients should be located close to each other in the input space, while that is not necessarily the case when features are missing or the imputation is poor.

Note that it is unrealistic to ask for high accuracies in this task, as the clean data are unlikely to be perfectly separable. As seen in Table 1, this proxy measure correlates well with the ground truth likelihood.

The performances of the different methods under this measure are reported in Table 3. Our model outperforms most baselines, including the GRUI-GAN, and performs comparably to the state-of-the-art method BRITS. This provides strong evidence that our model is well suited for real-world imputation tasks. Interestingly, forward imputation performs on a competitive level with

Table 3: Performance of the different models on the Physionet data set in terms of AUROC of a logistic regression trained on the imputed time series. We observe that the proposed model performs comparably to the state of the art.

Model	AUROC
Mean imputation [33]	0.703 ± 0.000
Forward imputation [33]	0.710 ± 0.000
GP [39]	0.704 ± 0.007
VAE [25]	0.677 ± 0.002
HI-VAE [37]	0.686 ± 0.010
GRUI-GAN [34]	0.702 ± 0.009
BRITS [7]	**0.742 ± 0.008**
GP-VAE (proposed)	**0.730 ± 0.006**

many of the baselines, suggesting that those baselines do not succeed in discovering any nonlinear structure in the data.

Note that, while BRITS outperforms our proposed model quantitatively, it does not fit a generative model to the data and does not offer any probabilistic interpretation. In contrast, our model can be used to sample different imputations from the variational posterior and thus provides an interpretable way of estimating the uncertainty of the predictions, as is depicted in Figure 3. The imputations on different clinical variables in the figure show how the posterior of our model exhibits different levels of uncertainty for different features (blue shaded area). The uncertainty estimates correlate qualitatively with the sparseness of the features and the noise levels of the measurements. This could be communicated to end-users, such as clinicians, to help them make informed decisions about the degree of trust they should have in the model.

It can also be seen in Figure 3 that our model produces smoother curves than the baselines. This is likely due to the denoising effect of our GP prior, which acts in a similar way to a Kalman filter in this case [23]. Especially when the data are very noisy, such as medical time series, this smoothing can make the imputations more interpretable for humans and can help in identifying temporal trends.

5 Conclusion

We presented a deep probabilistic model for multivariate time series imputation, combining ideas from variational autoencoders and Gaussian processes. The VAE maps the missing data from the input space into a latent space where the temporal dynamics are modeled by the GP. We use structured variational inference to approximate the latent GP posterior, which reflects the temporal correlations of the data more accurately than a fully factorized approximation. At the same time, inference in our variational distribution is still efficient, as opposed to inference in the full GP posterior. We empirically validated our proposed model on benchmark data sets and real-world medical data. We observed that our model outperforms classical baselines as well as modern deep learning approaches on these tasks and performs comparably to the state of the art.

In future work, it would be interesting to assess the applicability of the model to other data domains (e.g., natural videos) and explore a larger variety of kernel choices (e.g., non-smooth, periodic, or learned kernels) for the latent GP. Moreover, it could be a fruitful avenue of research to choose more sophisticated neural network architectures for the inference model and the generative model.

Acknowledgments

VF is supported by a PhD fellowship from the Swiss Data Science Center and by the PHRT grant #2017-110 of the ETH Domain. DB was partially supported by the ETH summer fellowship program. SM acknowledges funding from DARPA (HR001119S0038), NSF (FW-HTF-RM), and Qualcomm. We thank Robert Bamler, Gideon Dresdner, Claire Vernade, Rich Turner, and the anonymous reviewers for their valuable feedback.

This is version #645023440864087831186872693254 of this paper. The other $2^{100} - 1$ versions exist on different branches of the universe's wave function [11]. We thank Sean Carroll [8] for the inspiration for this versioning system.

References

[1] Samuel K Ainsworth, Nicholas J Foti, and Emily B Fox. Disentangled vae representations for multi-aspect and missing data. *arXiv preprint arXiv:1806.09060*, 2018.

[2] Robert Bamler and Stephan Mandt. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 380–389. JMLR. org, 2017.

[3] Robert Bamler and Stephan Mandt. Structured black box variational inference for latent time series models. *arXiv preprint arXiv:1707.01069*, 2017.

[4] Faraj Bashir and Hua-Liang Wei. Handling missing data in multivariate time series using a vector autoregressive model-imputation (var-im) algorithm. *Neurocomputing*, 276:23–30, 2018.

[5] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

[6] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[7] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: bidirectional recurrent imputation for time series. In *Advances in Neural Information Processing Systems*, pages 6775–6785, 2018.

[8] Sean Carroll. *Something deeply hidden: quantum worlds and the emergence of spacetime*. Dutton, 2019.

[9] Francesco Paolo Casale, Adrian Dalca, Luca Saglietti, Jennifer Listgarten, and Nicolo Fusi. Gaussian process prior variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 10369–10380, 2018.

[10] Adrian V Dalca, John Guttag, and Mert R Sabuncu. Unsupervised data imputation via variational inference of deep subspaces. *arXiv preprint arXiv:1903.03503*, 2019.

[11] Hugh Everett III. Relative state formulation of quantum mechanics. *Reviews of modern physics*, 29(3):454, 1957.

[12] Vincent Fortuin and Gunnar Rätsch. Deep mean functions for meta-learning in gaussian processes. *arXiv preprint arXiv:1901.08098*, 2019.

[13] Vincent Fortuin, Gideon Dresdner, Heiko Strathmann, and Gunnar Rätsch. Scalable gaussian processes on discrete domains. *arXiv preprint arXiv:1810.10368*, 2018.

[14] Vincent Fortuin, Matthias Hüser, Francesco Locatello, Heiko Strathmann, and Gunnar Rätsch. Som-vae: Interpretable discrete representation learning on time series. *arXiv preprint arXiv:1806.02199*, 2018.

[15] Marta Garnelo, Dan Rosenbaum, Chris J Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo J Rezende, and SM Eslami. Conditional neural processes. *arXiv preprint arXiv:1807.01613*, 2018.

[16] Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018.

[17] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2(5):6, 2017.

[18] James Honaker and Gary King. What to do about missing values in time-series cross-section data. *American Journal of Political Science*, 54(2):561–581, 2010.

[19] Y Huang and WF McColl. Analytical inversion of general tridiagonal matrices. *Journal of Physics A: Mathematical and General*, 30(22):7919, 1997.

[20] Oleg Ivanov, Michael Figurnov, and Dmitry Vetrov. Variational autoencoder with arbitrary conditioning. *arXiv preprint arXiv:1806.02382*, 2018.

[21] Patrick Jähnichen, Florian Wenzel, Marius Kloft, and Stephan Mandt. Scalable generalized dynamic topic models. *Conference on Artificial Intelligence and Statistics*, 2018.

[22] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[23] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

[25] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014.

[26] Rahul G Krishnan, Uri Shalit, and David Sontag. Deep kalman filters. *arXiv preprint arXiv:1511.05121*, 2015.

[27] Rahul G Krishnan, Uri Shalit, and David Sontag. Structured inference networks for nonlinear state space models. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[28] Neil D Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in neural information processing systems*, pages 329–336, 2004.

[29] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[30] Steven Cheng-Xian Li, Bo Jiang, and Benjamin Marlin. Misgan: Learning from incomplete data with generative adversarial networks. *International Conference on Learning Representations*, 2019.

[31] Yingzhen Li and Stephan Mandt. Disentangled sequential autoencoder. *International Conference on Machine Learning*, 2018.

[32] Zachary C Lipton and Subarna Tripathi. Precise recovery of latent vectors from generative adversarial networks. *International Conference on Learning Representations*, 2017.

[33] Roderick JA Little and Donald B Rubin. Single imputation methods. *Statistical analysis with missing data*, pages 59–74, 2002.

[34] Yonghong Luo, Xiangrui Cai, Ying Zhang, Jun Xu, et al. Multivariate time series imputation with generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 1596–1607, 2018.

[35] Chao Ma, Sebastian Tschiatschek, Konstantina Palla, Jose Miguel Hernandez Lobato, Sebastian Nowozin, and Cheng Zhang. Eddi: Efficient dynamic discovery of high-value information with partial vae. *International Conference on Machine Learning*, 2018.

[36] Ranjan K Mallik. The inverse of a tridiagonal matrix. *Linear Algebra and its Applications*, 325 (1-3):109–139, 2001.

[37] Alfredo Nazabal, Pablo M Olmos, Zoubin Ghahramani, and Isabel Valera. Handling incomplete heterogeneous data using vaes. *arXiv preprint arXiv:1807.03653*, 2018.

[38] Alma B Pedersen, Ellen M Mikkelsen, Deirdre Cronin-Fenton, Nickolaj R Kristensen, Tra My Pham, Lars Pedersen, and Irene Petersen. Missing data and multiple imputation in clinical epidemiological research. *Clinical epidemiology*, 9:157, 2017.

[39] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

[40] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *International Conference on Machine Learning*, 2014.

[41] Stephen Roberts, Michael Osborne, Mark Ebden, Steven Reece, Neale Gibson, and Suzanne Aigrain. Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110550, 2013.

[42] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[43] Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *2012 Computing in Cardiology*, pages 245–248. IEEE, 2012.

[44] Michalis Titsias and Neil D Lawrence. Bayesian gaussian process latent variable model. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 844–851, 2010.

[45] Martin J Wainwright and Michael Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[46] Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *Artificial Intelligence and Statistics*, pages 370–378, 2016.

[47] Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. Gain: Missing data imputation using generative adversarial nets. *International Conference on Machine Learning*, 2018.

[48] Cheng Zhang, Judith Butepage, Hedvig Kjellstrom, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

Appendix

A Implementation details

A.1 Neural network architectures

We use a convolutional neural network (CNN) as an inference network and a fully connected multilayer perceptron (MLP) as a generative network. The inference network convolves over the time dimension of the input data and allows for sequences of variable lengths. It consists of a number of convolutional layers that integrate information from neighboring time steps into a joint representation using a fixed receptive field (see Figure 1). The CNN outputs a tensor of size $\mathbb{R}^{T \times 3k}$, where k is the dimensionality of the latent space. Every row corresponds to a time step t and contains $3k$ parameters, which are used to predict the mean vector $\mathbf{m}_t$ as well as the diagonal and off-diagonal elements $\{b^j_{t,t}, b^j_{t,t+1}\}_{j=1:k}$ that characterize $\mathbf{B}$ at the given time step. More details about the network structures for the different experiments are provided in the following.

A.2 Healing MNIST

For the Healing MNIST data, we used a few 2D convolutional layers as preprocessors, since the single temporal states of the data are images. Those layers convolve over the image dimensions of each time step separately and generate latent features for each image. We then flatten these latent features and use them as inputs to the 1D convolution over time. The hyperparameters for the setup are given in Table S1.

Table S1: Hyperparameters used in the GP-VAE model for the experiment on Healing MNIST. Some of the parameters are only relevant in a subset of the models.

Hyperparameter	Value
Number of CNN layers in inference network	1
Number of filters per CNN layer	256
Filter size (i.e., time window size)	3
Number of feedforward layers in inference network	2
Width of feedforward layers	256
Dimensionality of latent space	256
Length scale of Cauchy kernel	2.0
Number of feedforward layers in generative network	3
Width of feedforward layers	256
Activation function of all layers	ReLU
Learning rate during training	0.001
Optimizer	Adam [24]
Number of training epochs	20
Train/val/test split of data set	50,000/10,000/10,000
Dimensionality of time points	784
Length of time series	10
Tradeoff parameter β	0.8

A.3 SPRITES

Since the SPRITES data also consist of images, we use the same 2D convolutional preprocessing as in the Healing MNIST experiment. The hyperparameters are reported in Table S2.

Table S2: Hyperparameters used in the GP-VAE model for the experiment on SPRITES. Some of the parameters are only relevant in a subset of the models.

Hyperparameter	Value
Number of CNN layers in inference network	1
Number of filters per CNN layer	32
Filter size (i.e., time window size)	3
Number of feedforward layers in inference network	2
Width of feedforward layers	256
Dimensionality of latent space	256
Length scale of Cauchy kernel	2.0
Number of feedforward layers in generative network	3
Width of feedforward layers	256
Activation function of all layers	ReLU
Learning rate during training	0.001
Optimizer	Adam [24]
Number of training epochs	20
Train/val/test split of data set	8,000/1,000/1,000
Dimensionality of time points	12288
Length of time series	8
Tradeoff parameter β	0.1

A.4 Real medical time series data

For the Physionet 2012 data, we omitted the convolutional preprocessing, since these are not image data. We hence directly feed the data into the 1D convolutional layer over time. In this experiment, we follow the evaluation protocol from BRITS [7] and use the same set for training and evaluation. We randomly eliminate 10 % of observed measurements from the data for validation.

The hyperparameters for this experiment are reported in Table S3.

Table S3: Hyperparameters used in the GP-VAE model for the experiment on medical time series from the Physionet data set. Some of the parameters are only relevant in a subset of the models.

Hyperparameter	Value
Number of CNN layers in inference network	1
Number of filters per CNN layer	128
Filter size (i.e., time window size)	24
Number of feedforward layers in inference network	1
Width of feedforward layers	128
Dimensionality of latent space	35
Length scale of Cauchy kernel	7.0
Number of feedforward layers in generative network	2
Width of feedforward layers	256
Activation function of all layers	ReLU
Learning rate during training	0.001
Optimizer	Adam [24]
Number of training epochs	40
Train/val/test split of data set	4,000
Dimensionality of time points	35
Length of time series	48
Tradeoff parameter β	0.2

B Additional experiments

B.1 Missingness rates on Healing MNIST

To explore the influence of the missingness rates on the performance of the models, we conducted an experiment where we introduced missingness into the Healing MNIST time series at different rates between 10 % and 90 %. We compare the performance in terms of negative log likelihood (S4).

Table S4: Performance of different models on Healing MNIST data with artificial missingness and different missingness rates. We report negative log likelihood (lower is better). The reported values are means and their respective standard errors over the test set.

Missingness	VAE [25]	HI-VAE [37]	GP-VAE (RBF kernel)	GP-VAE (proposed)
10 %	0.0125 ± 0.0000	0.0117 ± 0.0000	0.0113 ± 0.0000	**0.0109** ± **0.0000**
20 %	0.0319 ± 0.0001	0.0275 ± 0.0001	0.0262 ± 0.0001	**0.0258** ± **0.0001**
30 %	0.0640 ± 0.0002	0.0507 ± 0.0002	0.0478 ± 0.0001	**0.0465** ± **0.0001**
40 %	0.1106 ± 0.0003	0.0803 ± 0.0003	0.0781 ± 0.0002	**0.0743** ± **0.0002**
50 %	0.1931 ± 0.0006	0.1349 ± 0.0005	0.1217 ± 0.0004	**0.1196** ± **0.0004**
60 %	0.3360 ± 0.0010	0.2218 ± 0.0008	0.1982 ± 0.0006	**0.1957** ± **0.0006**
70 %	0.6216 ± 0.0019	0.3818 ± 0.0014	0.3426 ± 0.0010	**0.3261** ± **0.0010**
80 %	1.3284 ± 0.0042	0.7613 ± 0.0025	0.6798 ± 0.0020	**0.6488** ± **0.0019**
90 %	4.0610 ± 0.0127	2.1878 ± 0.0064	1.9424 ± 0.0054	**1.8464** ± **0.0050**

It can be seen that the proposed model outperforms the other deep architectures (including the GP-VAE with an RBF kernel) for all the different missingness rates. This also highlights that the Cauchy kernel does indeed help in modeling the temporal dynamics.

B.2 Missingness mechanisms on Healing MNIST

So far, in our synthetic experiments we only looked at artificial missingness that was introduced completely at random (MCAR), that is, we uniformly sampled features to be missing independently of each other and their value. In this experiment, we explore a few more structured missingness mechanisms which are described in the following. The average missingness rate for all the different mechanisms is around 50 %.

Feature correlation (*Spatial*). We assume different features to be correlated in their missingness, that is, a feature is more likely to be missing if certain other features are missing. We implement that by defining a spatial Gaussian process with RBF kernel on the Healing MNIST images and drawing the missingness patterns as samples from this process. Neighboring pixels are therefore correlated in their missingness.

Positive temporal correlation (*Temporal*$^+$). The missingness of features is positively correlated in time, that is, if a feature is missing at one time step, it is more likely to be missing at the consecutive time step. We implement this again with a Gaussian process with RBF kernel, this time defined over time for each feature separately.

Negative temporal correlation (*Temporal*$^-$). The missingness of features is negatively correlated in time, that is, if a feature is missing at one time step, it is less likely to be missing at the consecutive time step. We implement this with a determinantal point process (DPP) over time for each feature separately.

Missingness not at random (*MNAR*). In this setting, the missingness is actually dependent on the underlying ground-truth value of the feature. In our example, white pixels in the Healing MNIST images are twice as likely to be missing than black pixels.

We assessed our model and the baselines on all these different settings and report the results in terms of likelihood in Table S5.

Table S5: Performance of different models on Healing MNIST data with artificial missingness and different missingness mechanisms. We report negative log likelihood (lower is better). The reported values are means and their respective standard errors over the test set.

Mechanism	VAE [25]	HI-VAE [37]	GP-VAE (proposed)
Spatial	0.4802 ± 0.0016	0.2259 ± 0.0010	**0.1779 ± 0.0007**
Temporal$^+$	0.1918 ± 0.0006	0.1332 ± 0.0005	**0.1206 ± 0.0004**
Temporal$^-$	0.1940 ± 0.0006	0.1349 ± 0.0005	**0.1175 ± 0.0004**
MNAR	0.4798 ± 0.0016	0.2896 ± 0.0010	**0.2606 ± 0.0008**

We observe that our proposed model outperforms all baselines in terms of likelihood and MSE on all the different missingness mechanisms. We also observe that the MNAR setting seems to be the hardest one for the VAE-based models, followed by the setting with correlated features, whereas the settings with temporal correlation do not seem to be harder than the completely random ones (compare to Tab. S4). For the single imputation methods (mean and forward imputation), the MNAR setting also seems to be the hardest one, while the correlated features are not much harder than random. However, the positive temporal correlation is harder for those methods and the negative temporal correlation is even easier than random missingness.