2nd Symposium on Advances in Approximate Bayesian Inference, 2019 1–15

Deep Amortized Variational Inference for Multivariate Time Series Imputation with Latent Gaussian Process Models

Vincent Fortuin* FORTUIN@INF.ETHZ.CH
Department of Computer Science
ETH Zürich
Zürich, Switzerland

Dmitry Baranchuk* DMITRYBARANCHUK@GMAIL.COM
Yandex
Moscow, Russia

Gunnar Rätsch RAETSCH@INF.ETHZ.CH
Department of Computer Science
ETH Zürich
Zürich, Switzerland

Stephan Mandt MANDT@UCI.EDU
Donald Bren School for Information & Computer Sciences
University of California, Irvine
Irvine, USA

1. Introduction and related work

Multivariate medical 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. An ideal imputation model for medical time series should 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 (Roberts et al., 2013)), 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 (Little and Rubin, 2002; Pedersen et al., 2017). Finally, recent work has explored the use of non-linear dimensionality reduction using variational autoencoders for i.i.d. data points with missing values (Ainsworth et al., 2018; Ma et al., 2018; Nazabal et al., 2018) , but this work has not considered temporal data and strategies for sharing statistical strength across time. A more comprehensive analysis of existing approaches and their shortcomings is deferred to the appendix (Sec. A).

* Equal contribution.

In this paper, we propose an architecture that combines deep variational autoencoders (VAEs) with Gaussian process (GP) to efficiently model the latent dynamics at multiple time scales. Moreover, our 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.
- Efficient inference. We use a structured variational approximation that models posterior correlations in the time domain.
- 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 different domains.

2. Model

We propose a novel architecture for missing value imputation in medical time series. Our model can be seen as a way to perform amortized approximate inference on a latent Gaussian process model.

Specifically, we combine ideas from VAEs (Kingma and Welling, 2014), GPs (Rasmussen, 2003), Cauchy kernels (Jähnichen et al., 2018), structured variational distributions with efficient inference (Bamler and Mandt, 2017b), and a special ELBO for missing data (Nazabal et al., 2018) and synthesize these ideas into a general framework for missing data imputation on time series. In the following, we will outline the assumed generative model and derive our proposed inference scheme. We use standard notation (similar to (Nazabal et al., 2018)), which is detailed in the appendix (Sec. B.1).

2.1. Generative model

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 the appendix (Fig. S2).

In order to model data that varies at multiple time scales, we consider the Cauchy kernel, which has previously been successfully used in the context of robust dynamic topic modeling where similar multi-scale time dynamics occur (Jähnichen et al., 2018). It corresponds to an infinite mixture of RBF kernels with different length scales (Rasmussen, 2003).

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

$$p_\theta(\mathbf{x}_t \,|\, \mathbf{z}_t) = \mathcal{N}\left(g_\theta(\mathbf{z}_t), \sigma^2 \mathbf{I}\right) \ , \tag{1}$$

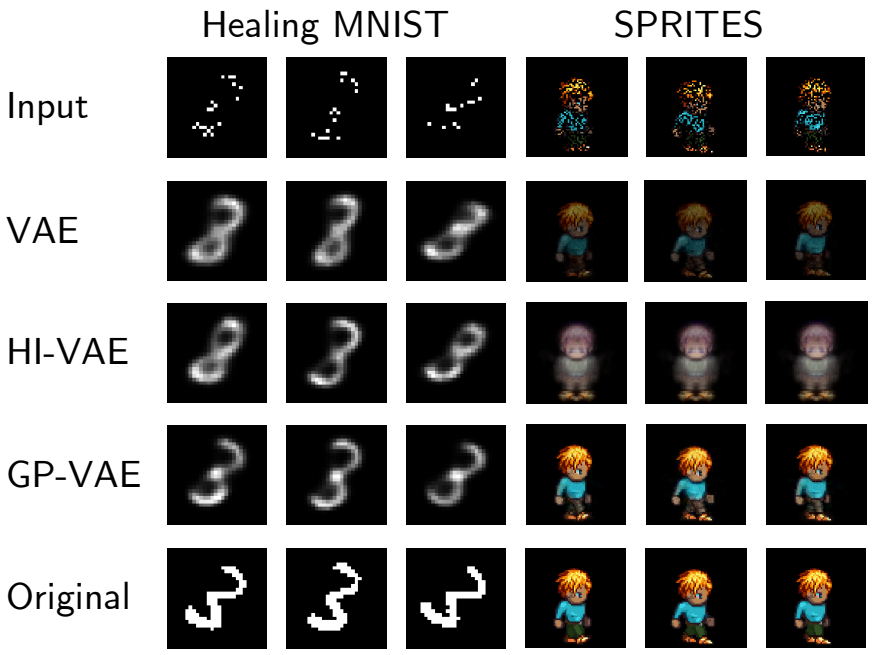

Figure 1: Reconstructions of Healing MNIST and SPRITES. The GP-VAE (proposed) is stable over time and yields the highest fidelity.

where $g_\theta(\cdot)$ is a potentially nonlinear function parameterized by the parameter vector θ. In our experiments, the function g_θ is implemented by a deep neural network.

2.2. 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 (Blei et al., 2017; Jordan et al., 1999; Zhang et al., 2018). Furthermore, to avoid inference over per-datapoint (local) variational parameters, we apply inference amortization (Kingma and Welling, 2014). To make our variational distribution more expressive and capture the temporal correlations of the data, we employ a structured variational distribution (Wainwright and Jordan, 2008) 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

$$q(\mathbf{z}_{1:T,j} \,|\, \mathbf{x}^o_{1:T}) = \mathcal{N}\left(\mathbf{m}_j, \mathbf{\Lambda}_j^{-1}\right) \,, \tag{2}$$

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 (Kingma and Welling, 2014; Rezende et al., 2014)).

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 as a tridiagonal matrix.

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 proposed model outperforms all the baselines.

	Healing MNIST			SPRITES
Model	NLL	MSE	AUROC	MSE
Mean imputation (Little and Rubin, 2002)	-	0.168 ± 0.000	0.938 ± 0.000	0.013 ± 0.000
Forward imputation (Little and Rubin, 2002)	-	0.177 ± 0.000	0.935 ± 0.000	0.028 ± 0.000
VAE (Kingma and Welling, 2014)	0.599 ± 0.002	0.232 ± 0.000	0.922 ± 0.000	0.034 ± 0.000
HI-VAE (Nazabal et al., 2018)	0.372 ± 0.008	0.134 ± 0.003	**0.962 ± 0.001**	0.035 ± 0.000
GP-VAE (proposed)	**0.341 ± 0.007**	**0.117 ± 0.002**	**0.960 ± 0.002**	**0.002 ± 0.000**

Samples from q can thus be generated in $\mathcal{O}(T)$ time (Bamler and Mandt, 2017b; Huang and McColl, 1997; Mallik, 2001) as opposed to the $\mathcal{O}(T^3)$ time complexity for a full-rank matrix. Moreover, compared to a fully factorized variational approximation, the number of variational parameters is 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.

We amortize the inference over $\mathbf{m}_j$ and $\mathbf{\Lambda}_j$ using an inference network $q_\psi(\cdot)$. As in standard VAE training, the parameters of the generative model and of the inference network can be jointly trained by optimizing the evidence lower bound (ELBO),

$$\begin{aligned} \log p(\mathbf{X}^o) \geq \sum_{t=1}^{T} \mathbb{E}_{q_\psi(\mathbf{z}_t \,|\, \mathbf{x}_{1:T})} \left[\log\, p_\theta(\mathbf{x}_t^o \,|\, \mathbf{z}_t)\right] \\ - \beta\, D_{KL}\left[q_\psi(\mathbf{z}_{1:T} \,|\, \mathbf{x}_{1:T}^o) \,\|\, p(\mathbf{z}_{1:T})\right] \end{aligned} \tag{3}$$

Following Nazabal et al. (2018) (see Sec. A), 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 (Higgins et al., 2017). This parameter takes care of balancing the influence between the likelihood on the observed data features and the latent prior. Our training objective is thus the RHS of (3).

3. Experiments

We performed experiments on the benchmark data set *Healing MNIST* (Krishnan et al., 2015), which combines the classical MNIST data set (LeCun et al., 1998) with properties common to medical time series, the SPRITES data set (Li and Mandt, 2018), and on a real-world medical data set from the 2012 Physionet Challenge (Silva et al., 2012). We compared our model against conventional single imputation methods (Little and Rubin, 2002), GP-based imputation (Rasmussen, 2003), VAE-based methods that are not specifically designed to handle temporal data (Kingma and Welling, 2014; Nazabal et al., 2018), and modern state-of-the-art deep learning methods for temporal data imputation (Cao et al., 2018; Luo et al., 2018).

We found strong quantitative (Tab. 1, 2) and qualitative (Fig. 1, 2) evidence that our proposed model outperforms most baseline methods in terms of imputation quality on all

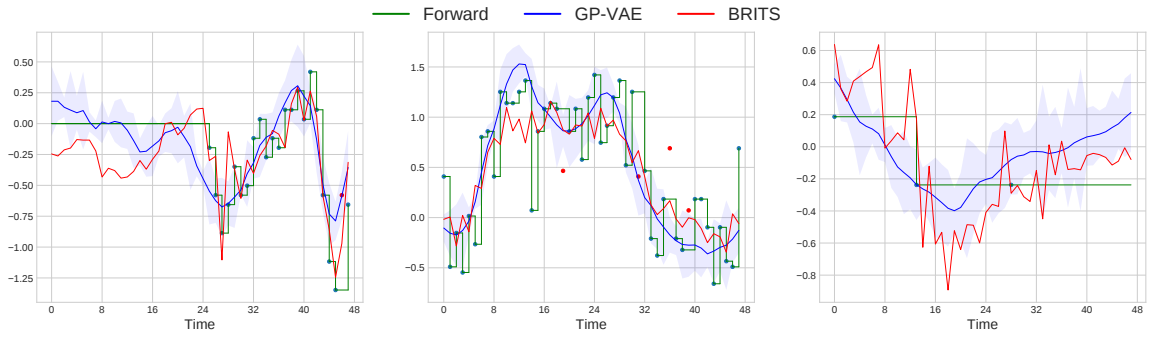

Figure 2: Imputations of several clinical variables with different amounts of missingness. BRITS (red) and forward imputation (green) yield single imputations, while the GP-VAE (blue) allows to draw samples from the posterior. The GP-VAE produces smoother curves, reducing noise from the original input, and exhibits an interpretable posterior uncertainty.

Table 2: 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 (Little and Rubin, 2002)	0.703 ± 0.000
Forward imputation (Little and Rubin, 2002)	0.710 ± 0.000
GP (Rasmussen, 2003)	0.704 ± 0.007
VAE (Kingma and Welling, 2014)	0.677 ± 0.002
HI-VAE (Nazabal et al., 2018)	0.686 ± 0.010
GRUI-GAN (Luo et al., 2018)	0.702 ± 0.009
BRITS (Cao et al., 2018)	**0.742 ± 0.008**
GP-VAE (proposed)	**0.730 ± 0.006**

three tasks and performs comparable to the state of the art (BRITS) on the medical data. This extends even to different missingness mechanisms, as is described in the appendix (Tab. S1).

For the real medical time series task, no ground-truth data exists, so we cannot report the mean squared error (MSE) or the negative log likelihoood (NLL). Following (Luo et al., 2018), we instead use a downstream classifier as a proxy measure. We use a linear SVM to predict mortality based on the imputed time series, since this was also one of the original tasks in the 2012 Physionet challenge (Silva et al., 2012). We find that this proxy measure correlates well with the likelihood in cases where ground-truth data is available (see Healing MNIST AUROC in Tab. 1), lending credence to the metric. More details about these experiments can be found in the appendix (Sec. C).

References

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.

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, 2017a.

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

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.

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.

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.

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.

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.

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

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

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

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*, 2018b.

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*, 2018a.

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

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.

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.

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

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

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

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.

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

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

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.

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.

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.

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

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

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

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

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.

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.

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

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

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.

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

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

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.

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.

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.

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.

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

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

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

Appendix

Appendix A. 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 (Pedersen et al., 2017). 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 (Honaker and King, 2010). Orthogonal to these ideas, methods along the lines of expectation-maximization (EM) have been proposed, but they often require additional modeling assumptions (Bashir and Wei, 2018).

Bayesian methods. When it comes to estimating likelihoods and uncertainties relating to the imputations, Bayesian methods, such as Gaussian processes (GPs) (Rasmussen, 2003), have a clear advantage over non-Bayesian methods such as single imputation (Little and Rubin, 2002). There has been much recent work in making these methods more expressive and incorporating prior knowledge from the domain (e.g., medical time series) (Fortuin and Rätsch, 2019; Wilson et al., 2016) or adapting them to work on discrete domains (Fortuin et al., 2018a), 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) (Lawrence, 2004; Titsias and Lawrence, 2010), 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) (Ainsworth et al., 2018; Dalca et al., 2019; Ma et al., 2018; Nazabal et al., 2018) or generative adversarial networks (GANs) (Li et al., 2019; Yoon et al., 2018). 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 (Lipton and Tripathi, 2017). 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., Fortuin et al., 2018b; Krishnan et al., 2015, 2017), 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 (Garnelo et al., 2018a,b; Ivanov et al., 2018). We do not regard this setting in our work.

HI-VAE. Our approach borrows some ideas from the HI-VAE (Nazabal et al., 2018). 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

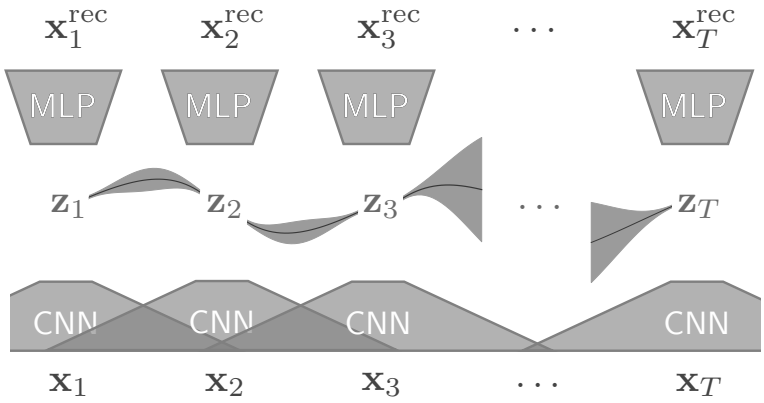

Figure S1: Architecture sketch of the model.

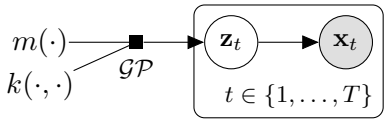

Figure S2: Graphical model.

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 (Luo et al., 2018), which we describe in Sec. 3. Another deep learning model for time series imputation is BRITS (Cao et al., 2018), 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 (Nazabal et al., 2018) described above. Furthermore, our idea of using a latent GP in the context of a deep generative model bears similarities to the GPPVAE (Casale et al., 2018), but note that the GPPVAE was not proposed to model time series data and does not take missing values into account. Lastly, the GP prior with the Cauchy kernel is reminiscent of Jähnichen et al. (2018) and the structured variational distribution is similar to the one used by Bamler and Mandt (2017b) in the context of modeling word embeddings over time, none of which used amortized inference.

Appendix B. Model details

B.1. Notation

We assume a data set $\mathbf{X} \in \mathbb{R}^{T \times d}$ with T data points $\mathbf{x}_t = [x_{t1}, \ldots, x_{tj}, \ldots, x_{td}]^\top \in \mathbb{R}^d$. Let us assume that the T data points were measured at T consecutive time points $\tau = [\tau_1, \ldots, \tau_T]^\top$

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 $\mathbf{x}_t^o := [x_{tj} \,|\, x_{tj} \text{ is observed}]$. Equivalently, the missing features are $\mathbf{x}_t^m := [x_{tj} \,|\, x_{tj} \text{ is missing}]$ 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 $p(\mathbf{x}_t^m \,|\, \mathbf{x}_t^o)$. In the time series setting, this independence assumption is not satisfied, which leads to the more complex estimation problem of $p(\mathbf{x}_t^m \,|\, \mathbf{x}_{1:T}^o)$.

B.2. Generative model

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 (Roberts et al., 2013) 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 order to model data that varies at multiple time scales, we consider a mixture of RBF kernels with different λ's (Rasmussen, 2003). By defining a Gamma distribution over the length scale, that is, $p(\lambda \,|\, \alpha, \beta) \propto \lambda^{\alpha-1} \exp(-\alpha\lambda/\beta)$, 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 (Rasmussen, 2003). For $\alpha = 1$ and $l^2 = 2\beta^{-1}$, it reduces to the Cauchy kernel

$$k_{Cau}(\tau, \tau') = \sigma^2 \left(1 + \frac{(\tau - \tau')^2}{l^2}\right)^{-1}, \tag{4}$$

which has previously been successfully used in the context of robust dynamic topic modeling where similar multi-scale time dynamics occur (Jähnichen et al., 2018). We therefore choose this kernel for our Gaussian process prior.

B.3. Inference

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,

$$\mathbf{\Lambda}_j := \mathbf{B}_j^\top \mathbf{B}_j \ , \quad \text{with} \quad \{\mathbf{B}_j\}_{tt'} = \begin{cases} b_{tt'}^j & \text{if } t' \in \{t, t+1\} \\ 0 & \text{otherwise} \end{cases} . \tag{5}$$

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 (2017a); Blei and Lafferty (2006).

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 (Bamler and Mandt, 2017b; Huang and McColl, 1997; Mallik, 2001) 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 $(\mathbf{m}, \mathbf{B}) = h_\psi(\mathbf{x}^o_{1:T})$. In the following, we accordingly denote the variational distribution as $q_\psi(\cdot)$. Following standard VAE training, the parameters of the generative model θ and of the inference network ψ can be jointly trained by optimizing the evidence lower bound (ELBO).

Following Nazabal et al. (2018) (see Sec. A), 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. Our training objective is thus the RHS of (3).

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 S1). 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 $\{b_{t,t}^j, b_{t,t+1}^j\}_{j=1:k}$ that characterize $\mathbf{B}$ at the given time step. More details about the network structure are given in the appendix (Sec. C).

Appendix C. Experimental details

C.1. Baseline methods

Forward imputation and mean imputation. Forward and mean imputation are so-called single imputation methods, which means that they do not attempt to fit a distribution over possible values for the missing features, but only predict one estimate (Little and Rubin,

2002). 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. 2.1), 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 (Kingma and Welling, 2014) and HI-VAE (Nazabal et al., 2018) 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 (3), 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. The GRUI-GAN (Luo et al., 2018) uses a recurrent neural network (RNN), namely a gated recurrent unit (GRU). Once the network is trained, a 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.

C.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. (2015) 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 (LeCun et al., 1998) 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 (Nazabal et al., 2018), 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 1. 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. 1) reveal the benefits of the GP-VAE approach: related approaches yield unstable reconstructions over time, while our approach offers more stable reconstructions,

Table S1: 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.

Mechanism	Mean imp.	Forward imp.	VAE	HI-VAE	GP-VAE (proposed)
Random	0.069 ± 0.000	0.099 ± 0.000	0.066 ± 0.000	0.042 ± 0.000	**0.037 ± 0.000**
Spatial	0.069 ± 0.000	0.099 ± 0.000	0.101 ± 0.000	0.060 ± 0.000	**0.052 ± 0.000**
Temporal^{+}	0.091 ± 0.000	0.116 ± 0.000	0.065 ± 0.000	0.042 ± 0.000	**0.037 ± 0.000**
Temporal^{-}	0.064 ± 0.000	0.093 ± 0.000	0.066 ± 0.000	0.042 ± 0.000	**0.037 ± 0.000**
MNAR	0.178 ± 0.000	0.174 ± 0.000	0.152 ± 0.001	0.088 ± 0.000	**0.078 ± 0.000**

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. S1).

C.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 (Li and Mandt, 2018). The dataset consists of 9,000 sequences of animated characters with different clothes, hair styles, and skin colors, performing 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 1. As in the previous experiment, our model outperforms the baselines in terms of likelihood and MSE and also yields the most convincing reconstructions. The HI-VAE seems to suffer from posterior collapse in this setting, which might be due to the large dimensionality of the input data.

C.4. Real medical time series data

We also applied our model to the data set from the 2012 Physionet Challenge (Silva et al., 2012). 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 36 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 and the GRUI-GAN model (Luo et al., 2018), 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 90 % of the features missing. To overcome this issue, Luo et al. (2018) 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 linear support vector machine (SVM) 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 when 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 2. Our model outperforms all baselines, including the GRUI-GAN, which provides strong evidence that our model is well suited for real-world medical time series imputations.