Longitudinal Variational Autoencoder

Siddharth Ramchandran[1] **Gleb Tikhonov**[1] **Kalle Kujanpää**[1]
Miika Koskinen[2,3] **Harri Lähdesmäki**[1]

[1]Department of Computer Science, Aalto University, Finland
[2]HUS Helsinki University Hospital, Finland
[3]Faculty of Medicine, University of Helsinki, Finland
siddharth.ramchandran@aalto.fi

Abstract

Longitudinal datasets measured repeatedly over time from individual subjects, arise in many biomedical, psychological, social, and other studies. A common approach to analyse high-dimensional data that contains missing values is to learn a low-dimensional representation using variational autoencoders (VAEs). However, standard VAEs assume that the learnt representations are i.i.d., and fail to capture the correlations between the data samples. We propose the Longitudinal VAE (L-VAE), that uses a multi-output additive Gaussian process (GP) prior to extend the VAE's capability to learn structured low-dimensional representations imposed by auxiliary covariate information, and derive a new KL divergence upper bound for such GPs. Our approach can simultaneously accommodate both time-varying shared and random effects, produce structured low-dimensional representations, disentangle effects of individual covariates or their interactions, and achieve highly accurate predictive performance. We compare our model against previous methods on synthetic as well as clinical datasets, and demonstrate the state-of-the-art performance in data imputation, reconstruction, and long-term prediction tasks.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

1 Introduction

Longitudinal datasets naturally arise in a wide variety of fields and applications, such as biomedicine, sociology, psychology, and many others. Such datasets include, for example, healthcare records, social media behaviour, consumer behaviour, etc., all collected repeatedly over time for each individual. Most longitudinal datasets contain both dependent and independent variables. For example, in biomedical data, dependent variables can comprise of lab tests and other measurements of the patient, whereas independent variables contain auxiliary descriptors of the patient, such as age, sex, time to disease event, etc. Analysing longitudinal datasets collected in such studies is challenging as they often involve time-varying covariates, high-dimensional correlated measurements, and missing values.

While non-linear, high-dimensional generative models are capable of learning complex data distributions, the statistical inference for such models is generally highly non-trivial. Auto-Encoding Variational Bayes (AEVB) (Kingma and Welling, 2014) is a powerful deep learning technique for efficient inference of latent variable models. The variational autoencoder (VAE) (Kingma and Welling, 2014; Rezende et al., 2014), the most popular exemplification of AEVB, learns a low-dimensional latent code of the dataset using two complementary deep neural networks (DNNs) to encode the high-dimensional data and decode the latent distribution, respectively. However, VAEs usually ignore the possible correlations (e.g. temporal correlations) between the learnt latent embeddings.

Related work Numerous extensions to achieve correlations in the latent space, model temporal data, and enhance the expressiveness of posterior distributions have been proposed for VAEs. Kulkarni et al. (2015) had proposed to group samples with specific proper-

ties in mini-batches to induce structure on the latent space. The conditional VAE (CVAE) introduced in (Sohn et al., 2015) incorporated the auxiliary covariate information directly in the inference and generative networks. However, CVAE fails to model the subject-specific temporal structure and does not explicitly constrain the latent space to achieve a low dimensional representation that evolves smoothly in time. Casale et al. (2018) proposed the GPPVAE to incorporate view and object information in a Gaussian process (GP) prior, to model the VAE's latent space structure. GPPVAE can account for temporal covariances between samples, but its ability to model subject-specific temporal structure is limited by the restrictive nature of the view-object GP product kernel. This compromises the applicability of GPPVAE in longitudinal study designs. Moreover, GPPVAE's pseudo-minibatch stochastic gradient descent (SGD) training scheme lacks the ability to scale to large data, as each training step requires a pass over the full data (epoch). Fortuin et al. (2020) built upon the idea of using a latent GP in VAEs, and proposed the GP-VAE that assumes an independent GP prior on each subject's time-series. Though GP-VAE is especially designed for time-series data, it can neither capture shared temporal structure across all data points nor make use of available auxiliary covariate information other than time.

Limitations of the expressiveness of posterior approximations in VAEs has been addressed by using normalising flows (NF) (Rezende and Mohamed, 2015) implemented with RealNVPs (Dinh et al., 2017), continuous-time NFs (Chen et al., 2018), inverse autoregressive flows (Kingma et al., 2016), and importance sampling (Müller et al., 2019). Methods have also been proposed to handle the disentanglement of dimensions in the latent space (Ainsworth et al., 2018a,b; Higgins et al., 2017) and improve latent representations (Alemi et al., 2018; Zhao et al., 2019). All these methods, however, assume independence across samples.

From the deep neural networks perspective, recurrent architectures (RNN) have been found to be particularly well-suited for temporal data analysis (Pearlmutter, 1989; Giles et al., 1994), including multi-outcome modelling problems. For example, Chung et al. (2015) proposed the variational RNN (VRNN) which extends the VAE into a recurrent framework for modelling highly structured sequential data. The VRNN models the temporal dependencies between the latent random variables across time steps. However, Chung et al. (2015) do not propose a way to handle and impute missing values. Also, VRNNs neither makes use of auxiliary covariates nor takes into account differing

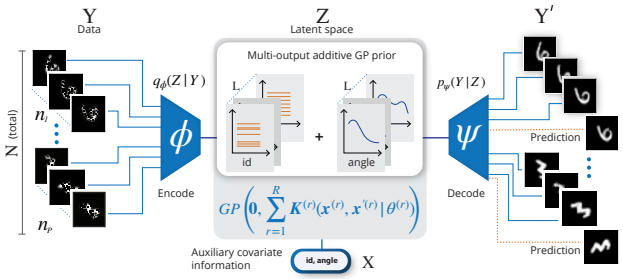

Figure 1: L-VAE overview.

time steps. BRITS (Cao et al., 2018) makes use of bi-directional LSTM-type RNNs (Schuster and Paliwal, 1997), and can efficiently impute missing values while accounting for the irregularities in the sampling times. However, BRITS is not a generative model which can limit the applicability of the trained model. Moreover, it is not straightforward to incorporate auxiliary information. Generative adversarial networks (GANs) can also be used for time-series data imputation and modelling (Goodfellow et al., 2016; Guo et al., 2019; Luo et al., 2018). GRUI-GAN (Luo et al., 2018) is an RNN-based method that makes use of adversarial training. This recurrent model suffers from similar pitfalls as BRITS with the added complexity of adversarial training. Table 1 contrasts the features of our proposed model to the key related methods.

Contributions In this paper, we propose a novel deep generative model that extends the capabilities of a VAE with a multi-output additive GP prior over the latent encodings domain, that models the correlation structure between the samples w.r.t. auxiliary information. Our L-VAE model is conceptualised in Fig. 1. Our model probabilistically encodes the longitudinal measurements with missing values (missing completely at random) onto a low-dimensional latent space with no missing values. The structured, low-dimensional latent dynamics are modelled using a multi-output additive GP that utilises the auxiliary covariates, followed by decoding back to the data domain. Such a GP prior introduces computational challenges for which we derive a novel divergence bound that leverages the commonly used inducing point formalism (Quiñonero-Candela and Rasmussen, 2005; Titsias, 2009) for efficient model inference. Our contributions can be summarised as follows:

- We introduce a VAE for longitudinal data with auxiliary covariate information, that can model the structured latent space dynamics with a multi-output additive GP prior.
- We derive an efficient mini-batch based GP infer-

Table 1: Comparison of related methods.

Models	Shared temporal structure	Individual temporal structure	Other covariates	Minibatching	Temporal irregularities	Generative	Reference
VAE	✗	✗	✗	✓	✗	✓	Kingma and Welling (2014)
CVAE	✓	✗	✓	✓	✗	✓	Sohn et al. (2015)
GPPVAE	✓	Limited	Limited	Pseudo	✗	✓	Casale et al. (2018)
GP-VAE	✗	✓	✗	✓	✗	✓	Fortuin et al. (2020)
VRNN	✓	✗	✗	✓	✗	✓	Chung et al. (2015)
BRITS	✓	✗	✗	✓	✓	✗	Cao et al. (2018)
GRUI-GAN	✓	✗	✗	✓	✓	✓	Luo et al. (2018)
L-VAE	✓	✓	✓	✓	✓	✓	Our work

ence scheme by exploiting the natural structural properties of the additive GP covariance functions (CF) used for longitudinal modelling.

- We compare L-VAE's performance against competing methods on several datasets and report the state-of-the-art performance in imputation, reconstruction, and long-term prediction.

The source code is available at `https://github.com/SidRama/Longitudinal-VAE` .

2 Methods

Problem setting Let D be the dimensionality of the observed data, P be the number of unique instances (e.g. unique patients, unique handwritten digit styles), n_p be the number of longitudinal samples from instance p, and $N = \sum_{p=1}^{P} n_p$ be the total number of samples (e.g. the total number of measurements for all patients, total number of images). Longitudinal samples for instance p are denoted as $Y_p = [\boldsymbol{y}_1^p, \ldots, \boldsymbol{y}_{n_p}^p]^T$, where each sample $\boldsymbol{y}_t^p \in \mathcal{Y}$. In this work, we assume $\mathcal{Y} = \mathbb{R}^D$. We represent the auxiliary covariate information for the instance p as $X_p = [\boldsymbol{x}_1^p, \ldots, \boldsymbol{x}_{n_p}^p]^T$, where covariates for each sample $\boldsymbol{x}_t^p \in \mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_Q$, $\mathcal{X}_q$ is the domain of the q^{th} covariate, and Q is the number of covariates. For example in Electronic Health Record (EHR) data, covariate information can include patient information such as age, sex, weight, time since remission, etc. Collectively, instance-specific samples and covariates form the full longitudinal data matrix $Y = [Y_1^T, \ldots, Y_P^T]^T = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N]^T$ and covariate matrix $X = [X_1^T, \ldots, X_P^T]^T = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^T$, respectively. We denote the low-dimensional latent space as $\mathcal{Z} = \mathbb{R}^L$ and a latent embedding for all N samples as $Z = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N]^T \in \mathbb{R}^{N \times L}$. We also index Z across L dimensions as $Z = [\bar{\boldsymbol{z}}_1, \ldots, \bar{\boldsymbol{z}}_L]$, where $\bar{\boldsymbol{z}}_l = [z_{1l}, \ldots, z_{Nl}]^T$ is a vector that contains the l^{th} dimension of the latent embedding for all N samples.

Auto-encoding variational Bayes Consider the joint generative model $p_\omega(\boldsymbol{y}, \boldsymbol{z}) = p_\psi(\boldsymbol{y}|\boldsymbol{z})p_\theta(\boldsymbol{z})$ parameterised by $\omega = \{\psi, \theta\}$, and assume we are interested to infer the latent variable $\boldsymbol{z}$ given $\boldsymbol{y}$. The posterior distribution $p_\omega(\boldsymbol{z}|\boldsymbol{y}) = p_\psi(\boldsymbol{y}|\boldsymbol{z})p_\theta(\boldsymbol{z})/p_\omega(\boldsymbol{y})$ is generally intractable due to the marginalisation over the latent space $p_\omega(\boldsymbol{y}) = \int p_\psi(\boldsymbol{y}|\boldsymbol{z})p_\theta(\boldsymbol{z})dz$. Auto-Encoding Variational Bayes (AEVB) (Kingma and Welling, 2014) is a powerful deep learning technique for latent variable models that factorise across samples as $p_\omega(Y, Z) = p_\psi(Y|Z)p_\theta(Z) = \prod_{n=1}^{N} p_\psi(\boldsymbol{y}_n|\boldsymbol{z}_n)p_\theta(\boldsymbol{z}_n)$. AEVB introduces an inference model (also called probabilistic encoder) $q_\phi(\boldsymbol{z}|\boldsymbol{y})$, parameterised by ϕ, that seeks to approximate the true posterior. The most well-known AEVB model is the variational autoencoder (VAE), where the inference model as well as the probabilistic decoder $p_\psi(\boldsymbol{y}|\boldsymbol{z})$ are parameterised by DNNs. Instead of optimising sample-specific variational parameters, as in standard variational inference (VI), AEVB uses amortised VI that exploits the inference model $q_\phi(\boldsymbol{z}|\boldsymbol{y})$ to obtain approximate distributions for each $\boldsymbol{z}_n$ as a function of the corresponding sample $\boldsymbol{y}_n$. Then, the approximate inference problem is fitted by maximising the evidence lower bound (ELBO) of the marginal log-likelihood w.r.t. ϕ:

$$\begin{aligned}&\log p_\omega(Y) \geq \mathcal{L}(\phi, \psi, \theta; Y)\\ &\triangleq \mathbb{E}_{q_\phi}[\log p_\psi(Y|Z)] - D_{\text{KL}}(q_\phi(Z|Y)||p_\theta(Z)) \rightarrow \max_\phi,\end{aligned}$$

where D_{KL} denotes the KL divergence. In practice, the approximate inference is typically conducted simultaneously alongside learning the generative model's parameters via solving the joint optimisation problem $\mathcal{L}(\phi, \psi, \theta; Y) \longrightarrow \max_{\phi,\psi,\theta}$.

2.1 Longitudinal variational autoencoder

The standard VAE model assumes that the joint distribution factorises across samples. Therefore, it can neither capture the potentially non-trivial structure of the data across the samples nor exploit that structure while making predictions. GP-VAE (Fortuin et al., 2020) and GPPVAE (Casale et al., 2018) address this issue for multivariate time-series data by assuming a GP prior on Z, whose CF depends on time or is factorised into feature and view components, respectively.

However, neither of these approaches coherently accommodates population-level structure by accounting for the potentially available additional covariates (such as patient sex or genotype) with instance-specific variability. In contrast, our method naturally incorporates both using an additive multi-output GP prior.

Building upon the work of Casale et al. (2018) and Fortuin et al. (2020), we propose to enhance the ability of VAEs to learn meaningful low-dimensional representations of Y, with a non-i.i.d. model for the low-dimensional latent space. This would enable our proposed method to effectively capture the structure of the data across the observed samples w.r.t. the auxiliary information X. Specifically, our generative model, parameterised by $\omega = \{\psi, \theta\}$, is formulated as

$$\begin{aligned} p_\omega(Y|X) &= \int_Z p_\psi(Y|Z,X)p_\theta(Z|X)dZ \\ &= \int_Z \prod_{n=1}^{N} p_\psi(\boldsymbol{y}_n|\boldsymbol{z}_n)p_\theta(Z|X)dZ, \end{aligned} \quad (1)$$

where the probabilistic decoder for normally distributed data,

$$p_\psi(\boldsymbol{y}_n|\boldsymbol{z}_n) = \mathcal{N}\left(\boldsymbol{y}_n|\boldsymbol{g}_\psi(\boldsymbol{z}_n), \mathrm{diag}(\sigma_{y1}^2, \ldots, \sigma_{yD}^2)\right) \quad (2)$$

is parameterised by a neural network ψ (variance parameters σ_{yd}^2 are included in ψ) and $p_\theta(Z|X)$ is defined by a multi-output GP prior that regulates the joint structure of Z with auxiliary variables X.

2.1.1 Multi-output additive GP prior

Consider a multi-output function $\boldsymbol{f} : \mathcal{X} \to \mathcal{Z} = \mathbb{R}^L$ where $L > 1$, denoted as $\boldsymbol{f}(\boldsymbol{x}) = [f_1(\boldsymbol{x}), \ldots, f_L(\boldsymbol{x})]^T$. Following Álvarez et al. (2012), we denote that $\boldsymbol{f}$ follows a multi-output Gaussian process prior as $\boldsymbol{f}(\boldsymbol{x}) \sim GP(\boldsymbol{\mu}(\boldsymbol{x}), \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}'|\theta))$, where $\boldsymbol{\mu}(\boldsymbol{x}) \in \mathbb{R}^L$ is the mean (which we assume as $\mathbf{0}$) and $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}'|\theta)$ is a matrix-valued positive definite cross-covariance function (CCF) whose entries define the covariances between the output dimensions for any $\boldsymbol{x}, \boldsymbol{x}'$. For any finite collection of inputs $X = [\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_N^T]^T$, the corresponding function values $\boldsymbol{f}(X) = [\boldsymbol{f}(\boldsymbol{x}_1)^T, \ldots, \boldsymbol{f}(\boldsymbol{x}_N)^T]^T \in \mathbb{R}^{NL\times 1}$ have a joint multivariate Gaussian distribution $p(\boldsymbol{f}(X)) = \mathcal{N}(\mathbf{0}, \boldsymbol{K}_{XX}(\theta))$, where the covariance matrix $\boldsymbol{K}_{XX}(\theta)$ is a block-partitioned matrix of size $NL \times NL$ with $L \times L$ blocks, so that block $[\boldsymbol{K}_{XX}(\theta)]_{i,j} = \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j|\theta)$.

In this work we consider multi-output GPs that factorise across the output dimensions, which is equivalent to diagonal CCF

$$\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}'|\theta) = \mathrm{diag}\left(k_1(\boldsymbol{x}, \boldsymbol{x}'|\theta_1), \ldots, k_L(\boldsymbol{x}, \boldsymbol{x}'|\theta_L)\right),$$

where $k_l(\boldsymbol{x}, \boldsymbol{x}'|\theta_l) = \mathrm{cov}(f_l(\boldsymbol{x}), f_l(\boldsymbol{x}'))$ is the CF for the l^{th} latent dimension. From the perspective of the generative model in eq. (1), such a choice is completely nonrestrictive compared to commonly used linear-type CCFs (e.g. linear model of coregionalisation), since the output of the multivariate GP is multiplied with the weights of the first layer in the neural network ψ. Moreover, diagonal CCFs enable us to completely characterise the multi-output GP prior for multivariate latent embedding Z in terms of univariate GP priors for its univariate column-components $\bar{\boldsymbol{z}}_l$, and to derive an efficient inference algorithm elucidated in Sec. 2.2.

Further, we will assume an additive structure for each CF, where each component depends on only a single covariate or a pair of covariates. Such CFs naturally provide a way to handle heterogeneous inputs, interpretability, and disentanglement for different covariates. Specifically, for the l^{th} dimension we assume that

$$\begin{aligned} f_l(\boldsymbol{x}) &= f_l^{(1)}(\boldsymbol{x}^{(1)}) + \ldots + f_l^{(R)}(\boldsymbol{x}^{(R)}), \\ f_l^{(r)}(\boldsymbol{x}^{(r)}) &\sim GP\left(\mathbf{0}, k_l^{(r)}(\boldsymbol{x}^{(r)}, \boldsymbol{x}^{(r)\prime}|\theta_l^{(r)})\right), \end{aligned}$$

where R refers to the number of additive kernels/components, each $f_l^{(r)}(\boldsymbol{x}^{(r)})$ is a separate GP with specific parameters $\theta_l^{(r)}$, and $\boldsymbol{x}^{(r)} \in \mathcal{X}^{(r)} \subseteq \mathcal{X}$. The covariance function of the additive GP is a sum of its components' covariances $f_l(\boldsymbol{x}) \sim GP\left(\mathbf{0}, \sum_{r=1}^{R} k_l^{(r)}(\boldsymbol{x}^{(r)}, \boldsymbol{x}^{(r)\prime}|\theta_l^{(r)})\right)$ (Rasmussen and Williams, 2006).

Assuming the l^{th} dimension of latent embedding $\bar{\boldsymbol{z}}_l$ is perturbed by i.i.d. zero-mean Gaussian noise σ_{zl}^2, the multi-output GP marginalised likelihood follows:

$$\begin{aligned} p(Z|X,\theta) &= \prod_{l=1}^{L} p(\bar{\boldsymbol{z}}_l|X, \theta_l, \sigma_{zl}^2) \\ &= \prod_{l=1}^{L} \mathcal{N}\left(\bar{\boldsymbol{z}}_l|\mathbf{0}, \sum_{r=1}^{R} K_{XX}^{(l,r)}(\theta_l^{(r)}) + \sigma_{zl}^2 I_N\right), \end{aligned}$$

where θ incorporates all $\theta_l^{(r)}$ and σ_{zl} parameters, and $K_{XX}^{(l,r)}(\theta_l^{(r)})$ is a $N \times N$ covariance matrix for $k_l^{(r)}(\boldsymbol{x}^{(r)}, \boldsymbol{x}^{(r)\prime}|\theta_l^{(r)})$.

Similar to Cheng et al. (2019), we use the following elementary covariance functions to construct the additive GP components within our framework: a) the effects of continuous covariates are modelled with the squared exponential CF; b) categorical covariates are modelled either alone with the categorical CF or together with the time (or other continuous) covariate using an interaction CF, which is the product of the categorical and squared exponential CFs; and c) the product of the squared exponential CF and the binary CF is used to model covariates that are defined for a subset of

samples (an example of such a covariate in a biomedical context is the time to the disease onset, which is defined only for those individuals who get a disease). All CFs for different latent dimensions and additive components have separate parameters $\theta_l^{(r)}$ (LR in total). Moreover, each additive component has an output variance parameter, which is a scale factor that is learnt alongside the other parameters. We constrain the scale parameters to positive values and fix the likelihood noise σ_{zl}^2 to 1, for better identifiability. Restricting to this family of CFs enabled us to devise accurate approximate computational strategies to overcome the typically cubic scaling of GPs as described in Sec. 2.2. A detailed description of the elementary CFs used is included in Suppl. Sec. 1. To handle missingness in the covariates, we set each CF $k_l^{(r)}(\boldsymbol{x}^{(r)}, \boldsymbol{x}^{(r)\prime}|\theta_l^{(r)}) \doteq 0$ for those sample pairs $\boldsymbol{x}$ and $\boldsymbol{x}'$ that contain at least one missing value for their respective covariate(s) in $\mathcal{X}^{(r)}$. This ensures that the contribution of the missing values to the target variable is 0.

2.1.2 Auto-Encoding Variational Bayes for L-VAE

We approximate the true posterior of Z with the product of multivariate Gaussian distributions across samples, each of which has a diagonal covariance matrix:

$$\begin{aligned} q_\phi(Z|Y) &= \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{z}_n|\boldsymbol{\mu}_\phi(\boldsymbol{y}_n), \mathrm{diag}(\boldsymbol{\sigma}_\phi^2(\boldsymbol{y}_n))\right) \\ &= \prod_{n=1}^{N}\prod_{l=1}^{L} \mathcal{N}\left(z_{nl}|\mu_{\phi,l}(\boldsymbol{y}_n), \sigma_{\phi,l}^2(\boldsymbol{y}_n)\right). \end{aligned} \quad (3)$$

Here, the probabilistic encoder is represented by neural network functions parameterised by ϕ, $\boldsymbol{\mu}_\phi : \mathbb{R}^D \to \mathbb{R}^L$ and $\boldsymbol{\sigma}_\phi^2 : \mathbb{R}^D \to \mathbb{R}_+^L$, that determine the means and variances of the approximating variational distribution. Following the AEVB approach of Kingma and Welling (2014), we form the ELBO for the L-VAE:

$$\begin{aligned} &\log p_\omega(Y|X) \geq \mathcal{L}(\phi, \psi, \theta; Y, X) \quad (4) \\ &\triangleq \mathbb{E}_{q_\phi(Z|Y)}\left[\log p_\psi(Y|Z)\right] - D_{\mathrm{KL}}(q_\phi(Z|Y)||p_\theta(Z|X)). \end{aligned}$$

From equations (1-2) and (3) which describe the factorised decoder and variational approximation respectively, the first term of the ELBO in eq. (4), called reconstruction loss, consists of additive terms across the samples and observations which can be written as $\mathbb{E}_{q_\phi(Z|Y)}\left[\log p_\psi(Y|Z)\right] = \sum_{n=1}^{N}\sum_{d=1}^{D}\mathbb{E}_{q_\phi(\boldsymbol{z}_n|\boldsymbol{y}_n)}\left[\log p_\psi(y_{nd}|\boldsymbol{z}_n)\right]$. If Y contains missing values, this summation is done only over the non-missing elements.

Given that our multi-output additive GP prior factorises across the latent dimensions $p_\theta(Z|X) = \prod_{l=1}^{L} p(\bar{\boldsymbol{z}}_l|X, \theta_l, \sigma_{zl}^2)$, we can exploit the additive nature of the KL divergence for pairs of independent distributions:

$$\begin{aligned} &D_{\mathrm{KL}}(q_\phi(Z|Y,X)||p_\theta(Z|X)) \\ &\qquad = \sum_{l=1}^{L} D_{\mathrm{KL}}(q_\phi(\bar{\boldsymbol{z}}_l|Y,X)||p_\theta(\bar{\boldsymbol{z}}_l|X)). \end{aligned} \quad (5)$$

Hence, we can completely avoid explicitly dealing with the numerics of multi-output GPs.

We have experimented with, e.g., a structured variational distribution (a tri-diagonal precision matrix per longitudinal instance) motivated by Bamler and Mandt (2017). However, we did not observe any improvement in the resulting performance. We leave the exploration of other approximating variational distributions for future work.

2.2 Efficient KL divergence computation

Optimising the variational objective in eq. (4) involves the computation of the KL divergence in eq. (5), which decomposes into L KL divergences,

$$\begin{aligned} D_{\mathrm{KL}}^{(l)} &= D_{\mathrm{KL}}(q_\phi(\bar{\boldsymbol{z}}_l|Y,X)||p_\theta(\bar{\boldsymbol{z}}_l|X)) \\ &= D_{\mathrm{KL}}(\mathcal{N}(\bar{\boldsymbol{\mu}}_l, W_l)||\mathcal{N}(\boldsymbol{0}, \Sigma_l)), \end{aligned} \quad (6)$$

where $\bar{\boldsymbol{\mu}}_l = [\mu_{\phi,l}(\boldsymbol{y}_1), \ldots, \mu_{\phi,l}(\boldsymbol{y}_N)]^T$, $W_l = \mathrm{diag}(\sigma_{\phi,l}^2(\boldsymbol{y}_1), \ldots, \sigma_{\phi,l}^2(\boldsymbol{y}_N))$, and $\Sigma_l = \sum_{r=1}^{R} K_{XX}^{(l,r)} + \sigma_{zl}^2 I_N$. Each of the KL divergences is available in closed form, but its exact computation requires $\mathcal{O}(N^3)$ flops, which makes it impractical when N exceeds a few thousands. Instead, we introduce a novel strategy to approximately compute this KL divergence at a significantly reduced computational cost. Without a loss of generality, we drop the index l for the remainder of this section.

A closely related problem has been studied by Titsias (2009) who proposed the well-known free-form variational lower bound for a GP marginal log-likelihood $\log \mathcal{N}(\bar{\boldsymbol{z}}|\boldsymbol{0}, \Sigma)$, assuming a set of M inducing locations $S = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_M]$ in $\mathcal{X}$ and inducing function values $\boldsymbol{u} = [f(\boldsymbol{s}_1), \ldots, f(\boldsymbol{s}_M)]^T = [u_1, \ldots, u_M]^T$. In fact, any lower bound for the prior GP marginal log-likelihood induces an upper bound of KL divergence (eq. (6)). The free-form bound of Titsias is known to be tight when M is sufficiently high and the covariance function is smooth enough.

However, longitudinal studies, by definition, always contain a categorical covariate corresponding to observed instances, which makes the covariance function non-continuous. By separating the additive component that corresponds to the interaction between instances and time (or age) from the other additive com-

ponents, the covariance matrix has the following general form $\Sigma = K_{XX}^{(A)} + \hat{\Sigma}$, where $\hat{\Sigma} = \text{diag}(\hat{\Sigma}_1, \ldots, \hat{\Sigma}_P)$, $\hat{\Sigma}_p = K_{X_pX_p}^{(R)} + \sigma_z^2 I_{n_p}$ and $K_{XX}^{(A)} = \sum_{r=1}^{R-1} K_{XX}^{(r)}$ contains all the other $R-1$ components. It is now clear that a reasonable inducing point set-up for the Titsias free-form bound would mandate that $M \geq P$, thus rendering the bound either computationally inefficient once P is large (due to high M) or insufficiently tight. Since the interaction CF is essential for accurate longitudinal modelling, we devised a novel free-form divergence upper bound for this class of GPs:

$$D_{\text{KL}} \leq \frac{1}{2}\Bigg(\text{tr}(\bar{\Sigma}^{-1}W) + \bar{\boldsymbol{\mu}}^T\bar{\Sigma}^{-1}\bar{\boldsymbol{\mu}} - N + \log|\bar{\Sigma}| - \log|W| + \sum_{p=1}^{P}\text{tr}\left(\hat{\Sigma}_p^{-1}\tilde{K}_{X_pX_p}^{(A)}\right)\Bigg) \tag{7}$$

where $\bar{\Sigma} = K_{XS}^{(A)}{K_{SS}^{(A)}}^{-1}K_{SX}^{(A)} + \hat{\Sigma}$ and $\tilde{K}_{X_pX_p}^{(A)} = K_{X_pX_p}^{(A)} - K_{X_pS}^{(A)}{K_{SS}^{(A)}}^{-1}K_{SX_p}^{(A)}$. The computational complexity of such a bound is $\mathcal{O}(\sum_{p=1}^{P} n_p^3 + NM^2)$ flops, which leads to an approximately similar computational complexity as Titsias (2009) bound when $n_p \simeq M \ll N$, but is significantly tighter:

Theorem 1. *For any set of inducing points S, the novel bound in eq.* (7) *is tighter than the one induced by the free-form variational bound of Titsias (2009).*

We provide a detailed derivation of the new bound and the Theorem 1 in Suppl. Sec. 2.

2.3 Mini-batch training

Although the novel KL divergence bound presented in the previous section enables $\mathcal{O}(N)$ scaling of the ELBO computation (see eq. (7)), it may still be prohibitively expensive for large datasets. To address this limitation, we devised a mini-batch training scheme that allows for multiple learning steps per epoch using unbiased stochastic estimates of the ELBO and its gradient. As elaborated above, the reconstruction loss term of eq. (4) trivially enables unbiased estimates. For the KL divergence part, we derived a principled mini-batching-compatible variant of the upper bound computation (eq. (7)) by building upon the SGD training of GPs with natural gradients presented in (Hensman et al., 2013). Specifically, the mini-batch compatible upper bound takes the form (see Suppl. Sec. 3 for an in-depth formulation and technical details)

$$\hat{D}_{\text{KL}} \leq \frac{1}{2}\frac{P}{\hat{P}}\sum_{p\in\mathcal{P}}\Upsilon_p - \frac{N}{2} + D_{\text{KL}}(\mathcal{N}(\boldsymbol{m}, H)||\mathcal{N}(\boldsymbol{0}, K_{SS}^{(A)})), \tag{8}$$

where the summation is over a subset of instances $\mathcal{P} \subset \{1, \ldots, P\}$ with $|\mathcal{P}| = \hat{P}$, Υ_p involves computing a quantity for instance p (see eq. 16 in Suppl. Sec. 3), and $\boldsymbol{m}$ and H are the global variational parameters for $\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{m}, H)$.

Eq. (8) enables us to use the mini-batching technique for a more precise approximate computation of the KL divergence term of L-VAE and its gradients, by approximately distributing an equal number of instances to each batch. Typically, the number of longitudinal samples per instance is small or moderate so that eq. (8) can be used to implement an efficient SGD algorithm. However, if the number of longitudinal samples for one or more instances is too large for an efficient implementation, the mini-batch based bound described above can be further extended beyond iterating over instances by allocating inducing points for each instance.

L-VAE is trained using the Adam optimiser (Kingma and Ba, 2015) on a *training* set, with early-stopping evaluated on a *validation* set. We evaluate imputation on the *training* set and predictive performance on an independent *test* set. See Suppl. Sec. 7 for more details.

2.4 Predictive distribution

Given the training samples Y, covariate information X, and learnt parameters ϕ, ψ, θ, the predictive distribution for the high-dimensional out-of-sample data $\boldsymbol{y}_*$ given covariates $\boldsymbol{x}_*$ follows

$$\begin{aligned} p_\omega(\boldsymbol{y}_*|\boldsymbol{x}_*, Y, X) &\approx \int_{\boldsymbol{z}_*, Z} p_\psi(\boldsymbol{y}_*|\boldsymbol{z}_*)p_\theta(\boldsymbol{z}_*|\boldsymbol{x}_*, Z, X) \\ &\qquad \cdot q_\phi(Z|Y, X)d\boldsymbol{z}_* dZ \\ &= \int_{\boldsymbol{z}_*}\prod_{d=1}^{D}\mathcal{N}\left(y_{*d}|g_{\psi,d}(\boldsymbol{z}_*), \sigma_{yd}^2\right)\prod_{l=1}^{L}\mathcal{N}(z_{*l}|\mu_{*l}, \sigma_{*l}^2)d\boldsymbol{z}_* \end{aligned}$$

where the means of the predictive low-dimensional representation are $\mu_{*l} = K_{\boldsymbol{x}_*X}^{(l)}\Sigma_l^{-1}\bar{\boldsymbol{\mu}}_l$ and variances are $\sigma_{*l}^2 = k_l(\boldsymbol{x}_*, \boldsymbol{x}_*) - K_{\boldsymbol{x}_*X}^{(l)}\Sigma_l^{-1}K_{X\boldsymbol{x}_*}^{(l)} + K_{\boldsymbol{x}_*X}^{(l)}\Sigma_l^{-1}W_l\Sigma_l^{-1}K_{X\boldsymbol{x}_*}^{(l)} + \sigma_{zl}^2$. See Suppl. Sec. 4 for more details.

Computation of the predictive distribution above does not assume any low-rank approximation and thus scales cubically with N. Unlike in the training phase, the required cubic matrix operations have to be conducted only once for any number of predictions, hence for moderate-sized data it is feasible to handle exactly. However, we can also write a scalable predictive distribution for the sparse variational approximation model that uses the inducing points. Detailed expressions are shown in Suppl. Sec. 5.

3 Experiments

We demonstrate the efficacy of our method in capturing the underlying data distribution and learning meaningful latent representations, quantifying performance in missing value imputation, reconstructing from the latent space, and by long-term predictions for synthetic images and a real-world medical time series dataset. We compare the performance of our method with various state-of-the-art methods for the respective datasets. Moreover, we have used similar encoder and decoder network architectures when performing comparisons across methods (see Suppl. Sec. 8). Additional results showing the latent embeddings and more comparisons can be found in Suppl. Sec. 9. In all our experiments, the *id* covariate is used as an identifier of an instance (e.g. patient or unique handwritten digit style) that takes values in $\{1, \ldots, P\}$, whereas other covariates are specific for each dataset (e.g. image rotation *angle* or time relative to disease onset *diseaseAge*). For brevity, we denote the additive components with different covariance functions (CFs) as follows: SE CF $\boldsymbol{f}_{\text{se}}(\cdot)$, categorical CF $\boldsymbol{f}_{\text{ca}}(\cdot)$, binary CF $\boldsymbol{f}_{\text{bi}}(\cdot)$, and their interaction CFs $\boldsymbol{f}_{\text{ca}\times\text{se}}(\cdot \times \cdot)$ and $\boldsymbol{f}_{\text{bi}\times\text{se}}(\cdot \times \cdot)$.

3.1 Rotated MNIST digits

We demonstrate our method on a variant of the MNIST dataset that comprises of 400 unique instances of the digit '3' as proposed in Casale et al. (2018). Each instance in this *training* set is rotated through 16 evenly separated rotation angles in $[0, 2\pi)$. That is, we have two covariates: categorical *id* and continuous *angle*. Moreover, the *validation* set comprises of 40 unique instances of the digit '3' rotated through 16 evenly separated rotation angles like the training set. As proposed in Casale et al. (2018), we created the *test* set (out-of-sample predictions) by completely removing one of the rotation angles for each instance and further removed four randomly selected angles to simulate incomplete data.

Hence, for each sequence, 5 images are not observed (see Fig. 2(a)). Therefore, in this experiment the *training* set parameters are $P = 400$, $N = 4400$ (where each $n_p = 11$), and $Q = 2$. The *test* set comprises of 400 test images of one rotation angle each. Fig. 2(a) demonstrates that our model was able to reconstruct arbitrary rotation angles, including the out-of-sample rotation (Fig. 2(b)). Fig. 2(c) compares the mean squared error (MSE) of the *test* set reconstructions from our method (three different GP variants) and GPPVAE and GP-VAE. The reconstruction loss decreases with increasing latent space dimension L, but our method consistently outperforms both the GPPVAE and GP-VAE.

3.2 Health MNIST

We simulate a longitudinal dataset with missing values using a modified version of the MNIST dataset. The dataset imitates many properties that may be found in actual medical data. In this experiment, we took the digits '3' as well as '6' and assumed that the different digits would represent two biological sexes. To simulate a shared age-related effect, all digit instances where shifted towards the right corner over time. We assume that half of the instances of '3' and '6' remain healthy (*diseasePresence* $= 0$) and half get a disease (*diseasePresence* $= 1$). For the diseased instances, we performed a sequence of 20 rotations with the amount of rotation depending on the time to disease diagnosis (*diseaseAge*).

We also introduced an irrelevant binary covariate, *location* which is set randomly for each unique instance. To every data point, we applied a random rotational jitter to mimic the addition of noise. We also randomly selected 25% of each image's pixels and set them as missing (we use these pixels to assess the imputation capability). Therefore, the simulated *training* dataset comprises of $P = 1000$ unique instances with a total of $N = 20000$ samples such that each $n_p = 20$. Each sample has $Q = 6$ covariates, namely *age*, *id*, *diseasePresence*, *diseaseAge*, *sex*, and *location*. The *validation* set comprised of 200 instances which are not present in the *training*. Additionally, the *training* dataset contained 100 additional instances for which the images of only the first 5 time points are given (as in the 'Data' row of Fig. 3(a)) — we use these to assess prediction capability by computing the MSE as well as by visualising the output of the decoder. As in Fortuin et al. (2020), we try to draw an analogy to healthcare by assuming that each frame of the time series represents a collection of measurements pertaining to a patient's health state and that the temporal evolution represents the non-linear evolution of that patient's health state.

Fig. 3(a) indicates that our approach performs well in reconstructing the temporal trajectory (or disease trajectory as per our analogy) and is able to predict the remaining trajectory, given the corresponding covariates. The benefits of using our model can especially be seen in the time period $[-4, 9]$ as it effectively captures the non-linear transformation about the disease event. GP-VAE is also capable to effectively reconstruct in the time period $[-10, -6]$, but fails completely in future predictions because it can only utilise the *age* covariate. Fig. 3(b) shows that our model also outperforms GP-VAE, GPPVAE, BRITS and GRUI-GAN in imputing the missing values for observed time points. Fig. 3(b) also highlights the robustness of our approach as the irrelevant covariate, *location*, has a very mild

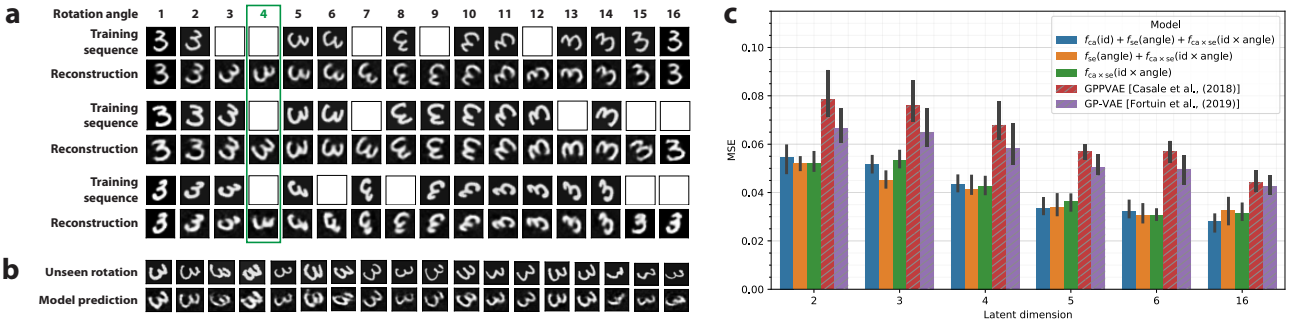

Figure 2: The rotated MNIST experiments. **(a)** Reconstructions obtained from our model using $\boldsymbol{f}_{\text{ca}}(\text{id}) + \boldsymbol{f}_{\text{se}}(\text{angle}) + \boldsymbol{f}_{\text{ca}\times\text{se}}(\text{id} \times \text{angle})$, and 16 latent dimensions. The blank boxes corresponds to the missing images. Rotation angle 4 is completely withheld from all instances. **(b)** Predictions for 18 random draws of the out-of-sample prediction state (i.e. unobserved angle in panel **a**). The first row is the real data and the bottom row is our model's prediction. **(c)** MSE on *test* set. The error bars represent the minimum and maximum values after 10 repetitions. The *training*, *test*, and *validation* sets are re-sampled for each repetition.

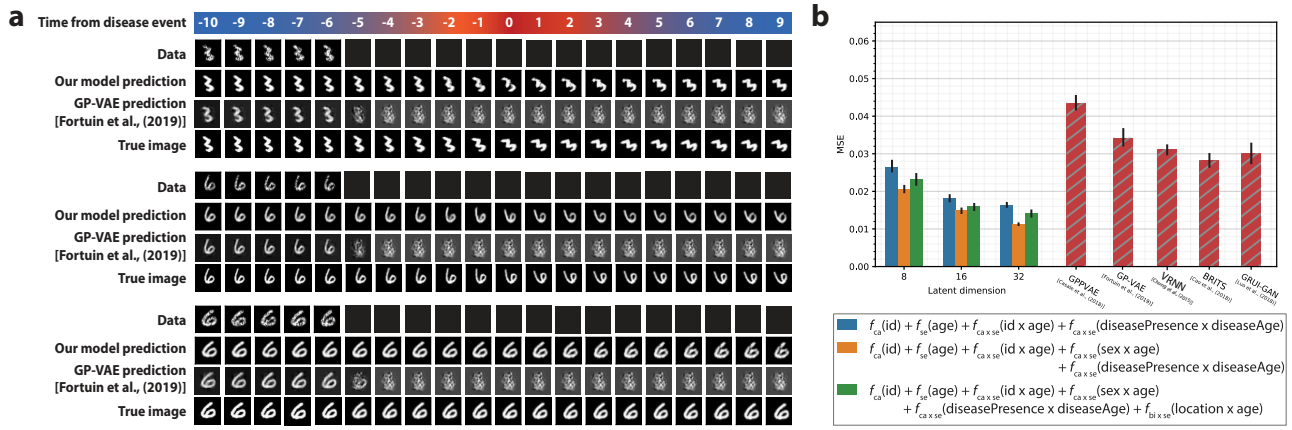

Figure 3: The Health MNIST experiments. **(a)** Reconstructions and predictions obtained from our model using $\boldsymbol{f}_{\text{ca}}(\text{id}) + \boldsymbol{f}_{\text{se}}(\text{age}) + \boldsymbol{f}_{\text{ca}\times\text{se}}(\text{id} \times \text{age}) + \boldsymbol{f}_{\text{ca}\times\text{se}}(\text{sex} \times \text{age}) + \boldsymbol{f}_{\text{ca}\times\text{se}}(\text{diseasePresence} \times \text{diseaseAge})$, and 32 latent dimensions. For the other methods, 64 latent dimensions were used when applicable. **(b)** MSE from imputing the missing values for the observed time points. Three model variants are shown for L-VAE.

detrimental effect on the overall model performance. Finally, Table 2 shows that L-VAE outperforms other methods in performing future predictions. See Suppl. Figs. 3 and 4 for latent space visualisations.

3.3 Healthcare data

We evaluated our model on health-care data from the Physionet Challenge 2012 (Silva et al., 2012). The objective of this challenge was to predict the in-hospital mortality of the patients that were monitored in the Intensive Care Unit (ICU) over a period of 48 hours. We made use of data from 3997 individuals ('set a') for *training* and 1000 individuals ('set c') for *validation*. Additionally, we used 3993 individuals ('set b') for *testing*. As in Cao et al. (2018), we focused on modelling the measurements of 35 different attributes (such as glucose level, blood pressure, body temper-

Table 2: MSE from performing future predictions (i.e., from time [-5, 9]) on the Health MNIST dataset. The values are the means and respective standard errors.

Model	Latent dimension	MSE
GPPVAE	64	0.057 ± 0.003
GP-VAE	64	0.059 ± 0.002
VRNN	64	0.049 ± 0.004
BRITS	N/A	0.047 ± 0.004
GRUI-GAN	64	0.053 ± 0.007
L-VAE	8	0.038 ± 0.003
L-VAE	16	0.033 ± 0.0018
L-VAE	**32**	$\mathbf{0.025 \pm 0.0015}$

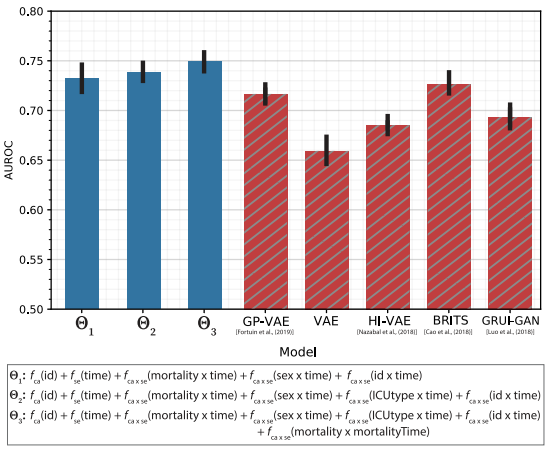

Figure 4: *Test* set AUROC scores for the patient mortality prediction task for the Physionet Challenge 2012 dataset. The number of latent dimensions is 32. Higher score is better. The error bars represent the minimum and maximum values after 10 repetitions.

ature, etc.), approx. 80% of which are missing in the data. We also made use of 7 patient-specific general auxiliary covariates that were made available as a part of the challenge, i.e. patient identifier (*id*), type of ICU unit (*ICUtype*), *height*, *weight*, *age*, *sex*, in-hospital death (*mortality*) as well as measurement hour (*time*), some of which were also missing for some patients. We constructed an additional covariate (time to mortality or *mortalityTime*) based on the provided survival time (see Suppl. Sec. 6 for data pre-processing). For model training, data for all patients ($P = 3997$) is available hourly ($n_p = 48$), so $N = 191856$. We trained our L-VAE model using the *training* samples and used it to build a Bayes classifier aimed at predicting the patient *mortality* for *test* data. Since the *test* data lacks information on *mortality* and *mortalityTime* covariates, for each patient in the *test* set, characterised by a pair of 48 hour attributes time-series Y_* and incomplete auxiliary information X_*, we approximated the marginal log-likelihoods from eq. (4) of the two alternative hypotheses: $L_i = \mathcal{L}(\phi, \psi, \theta; Y_*, X_*, \text{mortality} = i)$ for $i = \{0, 1\}$. Then the predicted mortality probability was computed as $P_1 = \exp(L_1)/(\exp(L_0) + \exp(L_1))$. We provide a detailed explanation of the Bayes classifier and mortality probability P_1 in Suppl. Sec. 6.

To demonstrate the efficacy of our method, we compared the AUROC scores for predicting mortality of the *test* data instances obtained using our method with those obtained from GP-VAE, a standard VAE, HI-VAE (Nazábal et al., 2020), BRITS and GRUI-GAN. These methods either do not use any auxiliary information (HI-VAE and VAE) or use only the time covariate (GP-VAE, BRITS, and GRUI-GAN). The mortality classification procedure of these methods first imputes the missing values with the generative model, and then exploits the imputed values as covariates to train a logistic regression that is finally used for mortality prediction (Fortuin et al., 2020). Fig. 4 shows that our L-VAE approach achieves higher AUROC scores. The performance with other additive GP covariance functions can be seen in Suppl. Fig. 5.

4 Discussion

In this paper, we introduced a novel deep generative model, L-VAE, that incorporates auxiliary covariate information to model the structured latent space dynamics for longitudinal datasets with missing values. We also introduced a novel computationally efficient inference strategy that exploits the structure of the additive GP covariance functions resulting in a novel lower bound. Moreover, the derived bound is theoretically guaranteed to be tighter than the free-form variational bound of Titsias (2009). We further developed this bound to allow mini-batch SGD training for computational efficiency. We demonstrated the efficacy of our method on synthetic as well as real-world datasets by showing that L-VAE achieves better out-of-sample prediction performance and missing value imputation than competing methods. Given the flexibility of our model and the state-of-the-art results, we expect L-VAE to become a useful tool for high-dimensional longitudinal data analysis.

Acknowledgements

We would like to acknowledge the computational resources provided by Aalto Science-IT, Finland. We would also like to thank Charles Gadd for the helpful discussions. This work was supported by the Academy of Finland [335436, 311584], Business Finland [2383/31/2015], and institutional HUS research funding.

References

S. K. Ainsworth, N. J. Foti, and E. B. Fox. Disentangled VAE representations for multi-aspect and missing data. *arXiv preprint arXiv:1806.09060*, 2018a.

S. K. Ainsworth, N. J. Foti, A. K. C. Lee, and E. B. Fox. oi-vae: Output interpretable vaes for nonlinear group factor analysis. In *Proceedings of the 35th International Conference on Machine Learning, ICML*. PMLR, 2018b.

A. A. Alemi, B. Poole, I. Fischer, J. V. Dillon, R. A. Saurous, and K. Murphy. Fixing a broken ELBO.

In *Proceedings of the 35th International Conference on Machine Learning, ICML*. PMLR, 2018.

W. Cao, D. Wang, J. Li, H. Zhou, L. Li, and Y. Li. BRITS: bidirectional recurrent imputation for time series. In *Advances in Neural Information Processing Systems, NeurIPS*, 2018.

F. P. Casale, A. V. Dalca, L. Saglietti, J. Listgarten, and N. Fusi. Gaussian process prior variational autoencoders. In *Advances in Neural Information Processing Systems, NeurIPS*, 2018.

C. Chen, C. Li, L. Chen, W. Wang, Y. Pu, and L. Carin. Continuous-time flows for efficient inference and density estimation. In *Proceedings of the 35th International Conference on Machine Learning, ICML*. PMLR, 2018.

L. Cheng, S. Ramchandran, T. Vatanen, N. Lietzén, R. Lahesmaa, A. Vehtari, and H. Lähdesmäki. An additive gaussian process regression model for interpretable non-parametric analysis of longitudinal data. *Nature Communications*, 2019.

J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio. A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems, NeurIPS*, 2015.

L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real NVP. In *5th International Conference on Learning Representations, ICLR*, 2017.

V. Fortuin, D. Baranchuk, G. Rätsch, and S. Mandt. GP-VAE: deep probabilistic time series imputation. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS*. PMLR, 2020.

C. L. Giles, G. M. Kuhn, and R. J. Williams. Dynamic recurrent neural networks: Theory and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 1994.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

Z. Guo, Y. Wan, and H. Ye. A data imputation method for multivariate time series based on generative adversarial network. *Neurocomputing*, 2019.

J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, UAI*. AUAI Press, 2013.

I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR*, 2017.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR*, 2015.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR*, 2014.

D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems, NeurIPS*, 2016.

T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. B. Tenenbaum. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems, NeurIPS*, 2015.

Y. Luo, X. Cai, Y. Zhang, J. Xu, and X. Yuan. Multivariate time series imputation with generative adversarial networks. In *Advances in Neural Information Processing Systems, NeurIPS*, 2018.

T. Müller, B. McWilliams, F. Rousselle, M. Gross, and J. Novák. Neural importance sampling. *ACM Transactions on Graphics*, 2019.

A. Nazábal, P. M. Olmos, Z. Ghahramani, and I. Valera. Handling incomplete heterogeneous data using VAEs. *Pattern Recognition*, 2020.

B. A. Pearlmutter. Learning state space trajectories in recurrent neural networks. *Neural Computation*, 1989.

J. Quiñonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 2005.

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

D. J. Rezende and S. Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning, ICML*, 2015.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31th International Conference on Machine Learning, ICML*, 2014.

M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 1997.

I. Silva, G. Moody, D. J. Scott, L. A. Celi, and R. G. Mark. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *2012 Computing in Cardiology*. IEEE, 2012.

K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems, NeurIPS*, 2015.

M. K. Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS*, 2009.

S. Zhao, J. Song, and S. Ermon. Infovae: Balancing learning and inference in variational autoencoders. In *Proceedings of the AAAI conference on artificial intelligence*. AAAI Press, 2019.

M. A. Álvarez, L. Rosasco, and N. D. Lawrence. Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning*, 2012.