
SCALABLE GAUSSIAN PROCESS VARIATIONAL AUTOENCODERS

Metod Jazbec
ETH Zürich
Zürich, Switzerland
jazbec.metod@gmail.com

Vincent Fortuin
ETH Zürich
Zürich, Switzerland
fortuin@inf.ethz.ch

Michael Pearce
Warwick University
Warwick, United Kingdom
scrambledpie@gmail.com

Stephan Mandt
University of California, Irvine
Irvine, United States of America
mandt@uci.edu

Gunnar Rätsch
ETH Zürich
Zürich, Switzerland
raetsch@inf.ethz.ch

ABSTRACT

Conventional variational autoencoders fail in modeling correlations between data points due to their use of factorized priors. Amortized Gaussian process inference through GP-VAEs has led to significant improvements in this regard, but is still inhibited by the intrinsic complexity of exact GP inference. We improve the scalability of these methods through principled sparse inference approaches. We propose a new scalable GP-VAE model that outperforms existing approaches in terms of runtime and memory footprint, is easy to implement, and allows for joint end-to-end optimization of all components.

1 Introduction

Variational autoencoders (VAEs) are among the most widely used models in representation learning and generative modeling (Kingma and Welling, 2013, 2019; Rezende et al., 2014). As VAEs typically make use of factorized priors, they fall short when modeling correlations between different data points. However, more expressive priors that capture correlations enable useful applications. Casale et al. (2018), for instance, showed that by modeling prior correlations between the data, one could generate a digit’s rotated image based on rotations of the same digit at different angles.

Gaussian process VAEs (GP-VAEs) have been designed to overcome this shortcoming (Casale et al., 2018). These models introduce a Gaussian process (GP) prior over the latent variables that correlates pairs of latent variables through a kernel function. While GP-VAEs have outperformed standard VAEs on many tasks (Casale et al., 2018; Fortuin et al., 2020; Pearce, 2020), combining the GPs and VAEs brings along fundamental computational challenges. On the one hand, neural networks reveal their full power in conjunction with large datasets, making mini-batching a practical necessity. GPs, on the other hand, are traditionally restricted to medium-scale datasets due to their unfavorable scaling. In GP-VAEs, these contradictory demands must be reconciled, preferably by reducing the $\mathcal{O}(N^3)$ complexity of GP inference, where N is the number of data points.

Despite recent attempts to improve the scalability of GP-VAE models by using specifically designed kernels and inference methods (Casale et al., 2018; Fortuin et al., 2020), a generic way to scale these models, regardless of data type or kernel choice, has remained elusive. This limits current GP-VAE implementations to small-scale datasets. In this work, we introduce the first generically scalable method for training GP-VAEs based on inducing points. We thereby improve the computational complexity from $\mathcal{O}(N^3)$ to $\mathcal{O}(bm^2 + m^3)$, where m is the number of inducing points and b is the batch size.

We show that applying the well-known inducing point approaches (Hensman et al., 2013; Titsias, 2009) to GP-VAEs is a non-trivial task: existing sparse GP approaches cannot be used off-the-shelf within GP-VAE models as they either necessitate having the entire dataset in the memory or do not lend themselves to being amortized. To address this issue, we propose a novel hybrid sparse GP method that is amenable to both mini-batching and amortization.

We make the following contributions:

- We propose the first scalable GP-VAE framework based on sparse GP inference (Sec. 3). In contrast to existing methods, our model is agnostic to the kernel choice, makes no assumption on the structure of the data at hand and allows for joint optimization of all model components.
- We provide theoretical motivations for the proposed method and introduce a hybrid sparse GP model that accommodates a crucial demand of GP-VAEs for simultaneous amortization and batching.
- We show empirically that the proposed approximation scheme maintains a high accuracy while being much more scalable and efficient (Sec. 4). Importantly from a practitioner’s point of view, our model is easy to implement as it requires no special modification of the training procedure.

2 Related Work

Sparse Gaussian processes. There has been a long line of work on sparse Gaussian process approximations, dating back to Snelson and Ghahramani (2006), Quiñero-Candela and Rasmussen (2005), and others. Most of these sparse methods rely on a summarizing set of points referred to as *inducing points* and mainly differ in the exact way of selecting those. Variational learning of inducing points was first considered in Titsias (2009) and was shown to lead to significant performance gains. Instead of optimizing an approximate marginal GP likelihood as done in non-variational sparse models, a lower bound on the exact GP marginal likelihood is derived and used as a training objective. Another approach relevant for our work is the stochastic variational approach from Hensman et al. (2013), where the authors proposed a sparse model that can, in addition to reducing the GP complexity, also be trained in mini-batches, enabling the use of GP models on (extremely) large datasets.

Improving VAEs. Extending the expressiveness and representational power of VAEs can be roughly divided into two (orthogonal) approaches. The first one focuses on increasing the flexibility of the approximate posterior (Rezende and Mohamed, 2015; Kingma et al., 2016), while the second one consists of imposing a richer prior distribution on the latent space. Various extensions to the standard Gaussian prior have been proposed, including a Gaussian mixture prior (Dilokthanakul et al., 2016; Kopf et al., 2019) and a von Mises-Fisher distribution prior (Davidson et al., 2018). GP-VAE models are part of this second group and, contrary to other work on extending VAE priors, aim to relax the *iid* assumption between data points. Moreover, GP-VAEs are also related to approaches that aim to learn more structured and interpretable representations of the data by incorporating auxiliary information, such as time or viewpoints (Sohn et al., 2015; Lin et al., 2018; Johnson et al., 2016).

Gaussian process VAEs. As mentioned above, the most related approaches to our work are the GP-VAE models of Casale et al. (2018) and Pearce (2020). However, neither of these are scalable for generic kernel choices and data types. The model from Pearce (2020) relies on exact GP inference, while Casale et al. (2018) exploit a (partially) linear structure of their GP kernel and use a Taylor approximation of the ELBO to get around computational challenges. Another GP-VAE model is proposed in Fortuin et al. (2020) where it is used for multivariate time series imputation. Their model is indeed scalable (even in linear time complexity), but it works exclusively on time series data since it exploits the Markov assumption. Additionally, it does not support a joint optimization of GP parameters, but assumes a fixed GP kernel. In parallel work, Ashman et al. (2020) have developed a framework for sparse GP-VAEs, but they focus on a setting with lower-dimensional data, several GPs per dataset, and presence of missing data (more similar to the setting in Fortuin et al. (2020)). Their work and ours are thus complementary.

3 Scalable SVGP-VAE

This work’s main contribution is the sparsification of the GP-VAE using the sparse GP approaches mentioned above. To this end, two separate variational approximation problems have to be solved jointly: an outer amortized inference procedure from the high-dimensional space to the latent space, and the inner sparse variational inference scheme on the GP. To motivate our proposed solution, we begin by pointing out the problems that arise when naïvely combining the two objectives.

3.1 Problem setting and notation

In this work, we consider high-dimensional data $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_N]^\top \in \mathbb{R}^{N \times K}$. Each data point has a corresponding low-dimensional auxiliary data entry, summarized as $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]^\top \in \mathcal{X}^N$, $\mathcal{X} \subseteq \mathbb{R}^D$. For example, \mathbf{y}_i could be a video frame and \mathbf{x}_i the corresponding time stamp. Our goal is to train a conditional generative model for generating \mathbf{Y} conditioned on \mathbf{X} .

To this end, we adopt a latent GP approach, summarized below. First, we need to model a prior distribution over the collection of latent variables $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_N]^\top \in \mathbb{R}^{N \times L}$, each latent variable \mathbf{z}_i living in an L -dimensional latent space. To model their joint distribution, we assume L independent latent functions $f^1, \dots, f^L \sim GP(0, k_\theta)$ with kernel parameters θ that result in \mathbf{Z} when being evaluated on \mathbf{X} . More precisely, $\mathbf{z}_i = [f^1(\mathbf{x}_i), \dots, f^L(\mathbf{x}_i)]$. By construction, the l^{th} latent channel of all latent variables $\mathbf{z}_{1:N}^l \in \mathbb{R}^N$ (the l^{th} column of \mathbf{Z}) has a correlated Gaussian prior with covariance $\mathbf{K}_{NN} = k_\theta(\mathbf{X}, \mathbf{X})$. Setting $\mathbf{K}_{NN} = I$ recovers the fully factorized prior commonly used in standard VAEs.

As in regular VAEs, each $\mathbf{z}_i \in \mathbb{R}^L$ is then “decoded” to parameterize the distribution over observations $\mathbf{y}_i = \mu_\psi(\mathbf{z}_i) + \epsilon_i$ where $\mu_\psi : \mathbb{R}^L \rightarrow \mathbb{R}^K$ is a network with parameters ψ and $\epsilon_i \sim \mathcal{N}(\mathbf{0}, \sigma_y^2 \mathbf{I}_K)$. Mathematically, the full generative model is given by

$$p_\theta(\mathbf{Z}|\mathbf{X}) = \prod_{l=1}^L \mathcal{N}(\mathbf{z}_{1:N}^l | \mathbf{0}, \mathbf{K}_{NN}),$$

$$p_\psi(\mathbf{Y}|\mathbf{Z}) = \prod_{i=1}^N p_\psi(\mathbf{y}_i|\mathbf{z}_i) = \prod_{i=1}^N \mathcal{N}(\mathbf{y}_i | \mu_\psi(\mathbf{z}_i), \sigma_y^2 \mathbf{I}_K).$$

The joint distribution is $p_{\psi,\theta}(\mathbf{Y}, \mathbf{Z}|\mathbf{X}) = p_\psi(\mathbf{Y}|\mathbf{Z})p_\theta(\mathbf{Z}|\mathbf{X})$. The true posterior for the latent variables $p_{\psi,\theta}(\mathbf{Z}|\mathbf{Y}, \mathbf{X}) = p_{\psi,\theta}(\mathbf{Y}, \mathbf{Z}|\mathbf{X})/p_{\psi,\theta}(\mathbf{Y}|\mathbf{X})$ is intractable due to the denominator which requires integrating over \mathbf{Z} . Hence, approximate inference methods are required to infer the unobserved \mathbf{Z} given the observed \mathbf{X} and \mathbf{Y} .

3.2 Amortized variational inference

Amortization in the typical VAE architecture uses a second (recognition) network from the high-dimensional data \mathbf{y}_i to the mean and variance of a fully factorized Gaussian distribution over $\mathbf{z}_i \in \mathbb{R}^L$. We denote it as $\tilde{q}_\phi(\mathbf{z}_i|\mathbf{y}_i) = \mathcal{N}(\mathbf{z}_i | \mu_\phi(\mathbf{y}_i), \text{diag}(\sigma_\phi^2(\mathbf{y}_i)))$ and it has network parameters ϕ . In Casale et al. (2018), this Gaussian distribution is used directly to approximate the posterior, $p_{\psi,\theta}(\mathbf{Z}|\mathbf{Y}) \approx \prod_i \tilde{q}_\phi(\mathbf{z}_i|\mathbf{y}_i)$. While this approach mirrors classical VAE design, the approximate posterior for a latent variable \mathbf{z}_i only depends on \mathbf{y}_i and ignores \mathbf{x}_i . This is in stark contrast to traditional Gaussian processes where latent function values $f(x)$ are informed by all y values according to the similarity of the corresponding x values.

Building on this model, Pearce (2020) instead proposed to use the inference network $\tilde{q}_\phi(\mathbf{z}_i|\mathbf{y}_i)$ to replace only the intractable likelihood $p_\psi(\mathbf{y}_i|\mathbf{z}_i)$ in the posterior. By combining \tilde{q}_ϕ with tractable terms, the approximate posterior could be explicitly normalized as

$$q(\mathbf{Z}|\mathbf{Y}, \mathbf{X}, \phi, \theta) := \prod_{l=1}^L \frac{\prod_{i=1}^N \tilde{q}_\phi(\mathbf{z}_i^l|\mathbf{y}_i) p_\theta(\mathbf{z}_{1:N}^l|\mathbf{X})}{Z_{\phi,\theta}^l(\mathbf{Y}, \mathbf{X})}, \quad (1)$$

where the normalizing constant $Z_{\phi,\theta}^l(\mathbf{Y}, \mathbf{X})$ can be computed analytically. Noting the symmetry of the Gaussian distribution, $\mathcal{N}(z|\mu, \sigma) = \mathcal{N}(\mu|z, \sigma)$, the approximate posterior for channel l is mathematically equivalent to the (exact) GP posterior in the traditional GP regression with inputs

\mathbf{X} and outputs $\tilde{\mathbf{y}}_l := \mu_\phi^l(\mathbf{Y})$ with heteroscedastic noise $\tilde{\sigma}_l := \sigma_\phi^l(\mathbf{Y})$. We therefore refer to each $\{\mathbf{X}, \tilde{\mathbf{y}}_l, \tilde{\sigma}_l\}$ as the *latent dataset* for the l^{th} channel. Each normalizing constant of Equation 1 is also the GP marginal likelihood of the l^{th} latent dataset. The parameters $\{\psi, \phi, \theta\}$ are learnt by maximizing the evidence lower bound (ELBO) in the Pearce model,

$$\mathcal{L}_P(\psi, \phi, \theta) = \sum_{i=1}^N \mathbb{E}_{q(\mathbf{z}_i|\cdot)} \left[\log p_\psi(\mathbf{y}_i|\mathbf{z}_i) - \log \tilde{q}_\phi(\mathbf{z}_i|\mathbf{y}_i) \right] + \sum_{l=1}^L \log Z_{\phi, \theta}^l(\mathbf{Y}, \mathbf{X}). \quad (2)$$

The first term is the difference between the true likelihood and recognition network approximate likelihood, the second term is the sum over GP marginal likelihoods of each latent dataset.

The computational challenges of $\mathcal{L}_P(\cdot)$ are twofold. First, for the latent GP regression, an inverse and a log-determinant of the kernel matrix $\mathbf{K}_{NN} \in \mathbb{R}^{N \times N}$ must be computed, resulting in $\mathcal{O}(N^3)$ time complexity. Second, the ELBO does not decompose as a sum over data points, so the entire dataset $\{\mathbf{X}, \mathbf{Y}\}$ is needed for one evaluation of $\mathcal{L}_P(\cdot)$.

Given the latent dataset, at first glance, we may simply apply sparse GP regression techniques instead of traditional regression. We next look at two widely used methods (Titsias (2009) and Hensman et al. (2013)) and highlight their drawbacks for this task. We then propose a new hybrid approach solving these issues.

3.3 Latent Sparse GP Regression

To simplify the notation, we focus on a single channel and suppress l , resulting in $\tilde{\mathbf{y}}$ and $\tilde{\sigma}$, $\log Z_{\theta, \phi}(\cdot)$ and f . Given an (amortized latent) regression dataset $\mathbf{X}, \tilde{\mathbf{y}}, \tilde{\sigma}$, sparse Gaussian process methods assume that there exists a set of $m \ll N$ inducing points with inputs $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_m] \in \mathcal{X}^m$ and outputs $\mathbf{f}_m := f(\mathbf{U}) \sim \mathcal{N}(f(\mathbf{U}) | \boldsymbol{\mu}, \mathbf{A})$ that summarize the regression dataset. $\mathbf{U}, \boldsymbol{\mu}, \mathbf{A}$ are parameters to be learnt. Given a (test) set of r new inputs \mathbf{X}_r , the sparse approximate (predictive) distribution over outputs $\mathbf{f}_r = f(\mathbf{X}_r)$ is

$$q_S(\mathbf{f}_r | \mathbf{X}_r, \mathbf{U}, \boldsymbol{\mu}, \mathbf{A}, \theta) = \mathcal{N}(\mathbf{f}_r | \mathbf{K}_{rm} \mathbf{K}_{mm}^{-1} \boldsymbol{\mu}, \mathbf{K}_{rr} - \mathbf{K}_{rm} \mathbf{K}_{mm}^{-1} \mathbf{K}_{mr} + \mathbf{K}_{rm} \mathbf{K}_{mm}^{-1} \mathbf{A} \mathbf{K}_{mm}^{-1} \mathbf{K}_{mr}), \quad (3)$$

where kernel matrices are $\mathbf{K}_{mm} = k_\theta(\mathbf{U}, \mathbf{U})$, $\mathbf{K}_{rr} = k_\theta(\mathbf{X}_r, \mathbf{X}_r)$, and $\mathbf{K}_{mr} = \mathbf{K}_{rm}^\top = k_\theta(\mathbf{U}, \mathbf{X}_r)$. By introducing inducing points, the cost of learning the model is reduced from $\mathcal{O}(N^3)$ in $\log Z_{\phi, \theta}(\cdot)$ to $\mathcal{O}(Nm^2)$ in a modified objective.

We next describe two of the most popular ways to learn the variational parameters $\mathbf{U}, \boldsymbol{\mu}, \mathbf{A}$ that are based on a second *inner* variational approximation for the Gaussian process regression that lower bounds $\log Z_{\phi, \theta}(\cdot)$. For this second inner variational inference, we aim to learn a cheap $q_S(\cdot)$ (Equation 3) that closely approximates the expensive $q(\cdot)$ (Equation 1).

Titsias (2009). Let $\mathbf{z} = \mathbf{z}_{1:N}^l$, then the parameters $\mathbf{U}, \boldsymbol{\mu}, \mathbf{A}$ may be learnt by minimizing $\text{KL}(q_S(\mathbf{z} | \cdot) || q(\mathbf{z} | \cdot))$, or equivalently by maximizing a lower bound to the marginal likelihood of the latent dataset $\log Z_{\phi, \theta}^l(\cdot)$. Let $\boldsymbol{\Sigma} := \mathbf{K}_{mm} + \mathbf{K}_{mN} \text{diag}(\tilde{\sigma}^{-2}) \mathbf{K}_{Nm}$, then the optimal $\boldsymbol{\mu}$ and \mathbf{A} may be found analytically:

$$\boldsymbol{\mu}_T = \mathbf{K}_{mm} \boldsymbol{\Sigma}^{-1} \mathbf{K}_{mN} \text{diag}(\tilde{\sigma}^{-2}) \tilde{\mathbf{y}}, \quad (4)$$

$$\mathbf{A}_T = \mathbf{K}_{mm} \boldsymbol{\Sigma}^{-1} \mathbf{K}_{mm}, \quad (5)$$

where $\mathbf{K}_{mN} = k_\theta(\mathbf{U}, \mathbf{X})$. Plugging $\boldsymbol{\mu}_T$ and \mathbf{A}_T back into the appropriate evidence lower bound yields the final lower bound for learning \mathbf{U} in the Titsias model

$$\begin{aligned} \mathcal{L}_T(\mathbf{U}, \phi, \theta) = & \log \mathcal{N}(\tilde{\mathbf{y}} | \mathbf{0}, \mathbf{K}_{Nm} \mathbf{K}_{mm}^{-1} \mathbf{K}_{mN} + \text{diag}(\tilde{\sigma}^2)) \\ & - \frac{1}{2} \text{Tr}(\text{diag}(\tilde{\sigma}^{-2}) (\mathbf{K}_{NN} - \mathbf{K}_{Nm} \mathbf{K}_{mm}^{-1} \mathbf{K}_{mN})). \end{aligned} \quad (6)$$

Note that the bound is a function of $\tilde{\mathbf{y}}$ and $\tilde{\sigma}$ which depend on the recognition network with parameters ϕ and the kernel matrices which depend upon θ hence we make these arguments explicit. In the full GP-VAE ELBO $\mathcal{L}_P(\cdot)$, substituting $q_S(\cdot)$, $\mathcal{L}_T(\cdot)$ in place of $q(\cdot)$, $\log Z_{\phi, \theta}(\cdot)$ yields a sparse

GP-VAE ELBO that can be readily used to reduce computational complexity of existing GP-VAE methods for a generic dataset and an arbitrary GP kernel function.¹

However, observe from Equations 4, 5 and 6 that the entire dataset $\{\mathbf{X}, \mathbf{Y}\}$ enters through \mathbf{K}_{NN} and $\tilde{\mathbf{y}}, \tilde{\boldsymbol{\sigma}}$ respectively. Therefore, this ELBO is not amenable to mini-batching and has large memory requirements.

Hensman et al. (2013). In order to make variational sparse GP regression amenable to mini-batching, Hensman et al. (2013) proposed an ELBO that lower bounds \mathcal{L}_T and, more importantly, decomposes as a sum of terms over data points. Adopting our notation with explicit parameters, the Hensman ELBO is given by

$$\mathcal{L}_H(\mathbf{U}, \boldsymbol{\mu}, \mathbf{A}, \phi, \theta) = -\text{KL}(q_S(\mathbf{f}_m|\cdot) \parallel p_\theta(\mathbf{f}_m|\cdot)) + \sum_{i=1}^N \left\{ \log \mathcal{N}(\tilde{y}_i | \mathbf{k}_i \mathbf{K}_{mm}^{-1} \boldsymbol{\mu}, \tilde{\sigma}_i^{-2}) - \frac{1}{2\tilde{\sigma}_i^2} (\tilde{k}_{ii} + \text{Tr}(\mathbf{A} \Lambda_i)) \right\}. \quad (7)$$

Above, \mathbf{k}_i is the i -th row of \mathbf{K}_{Nm} , $\Lambda_i = \mathbf{K}_{mm}^{-1} \mathbf{k}_i \mathbf{k}_i^\top \mathbf{K}_{mm}^{-1}$ and \tilde{k}_{ii} is the i -th diagonal element of the matrix $\mathbf{K}_{NN} - \mathbf{K}_{Nm} \mathbf{K}_{mm}^{-1} \mathbf{K}_{mN}$. Due to the decomposition over data points, the gradients $\nabla \mathcal{L}_H(\cdot)$ in stochastic or mini-batch gradient descent are unbiased and only the data in the current batch are needed in memory for the gradient updates. Consequently, with batch size b the GP complexity is further reduced to $\mathcal{O}(bm^2 + m^3)$. Note that for $\boldsymbol{\mu} = \boldsymbol{\mu}_T$, $\mathbf{A} = \mathbf{A}_T$ and $b = N$, $\mathcal{L}_H(\cdot)$ recovers $\mathcal{L}_T(\cdot)$ (Hensman et al., 2013).

While this method may seem to meet our requirements, it has a fatal drawback. Firstly, it is not amortized as $\boldsymbol{\mu}$ and \mathbf{A} are not functions of the observed data $\{\mathbf{X}, \mathbf{Y}\}$ but instead need to be optimized once for each dataset. Secondly, as a consequence, in the full GP-VAE ELBO $\mathcal{L}_P(\cdot)$, substituting $q_S(\cdot)$, $\mathcal{L}_H(\cdot)$ in place of $q(\cdot)$, $\log Z_{\phi, \theta}(\cdot)$ and simplifying yields the following expression

$$\mathcal{L}_{PH}(\mathbf{U}, \psi, \theta, \boldsymbol{\mu}^{1:L}, \mathbf{A}^{1:L}) = \sum_{i=1}^N \mathbb{E}_{q_S} \left[\log p_\psi(\mathbf{y}_i | \mathbf{z}_i) \right] + \sum_{l=1}^L \text{KL}(q_S^l(\mathbf{f}_m|\cdot) \parallel p_\theta^l(\mathbf{f}_m|\cdot)) \quad (8)$$

where $q_S^l(\mathbf{f}_m|\cdot) = \mathcal{N}(\mathbf{f}_m | \boldsymbol{\mu}^l, \mathbf{A}^l)$.

Note that the ELBO above is not a function of the recognition network parameters ϕ (for the full derivation, we refer to Appendix B.1). The sparse approximate posterior is parameterized by $\mathbf{U}, \boldsymbol{\mu}, \mathbf{A}, \theta$ which are all treated as free parameters to be optimized, that is, they are not functions of the latent dataset or the recognition network. Maximizing the full GP-VAE ELBO is equivalent to minimizing the KL divergence from the approximate to the true posterior and neither of these depend upon the latent dataset or the inference network. Therefore, using the Hensman sparse GP within an amortized GP-VAE model causes the ELBO to be independent of the inference network parameters. Hence, this method also cannot be used as-is to amortize the sparse GP-VAE with mini-batches.

3.4 The best of both ELBOs

Recall our goal to make GP-VAE models amenable to large datasets. This requires avoiding the large memory requirements and being able to amortize inference. To alleviate these problems, Casale et al. (2018) propose to use a Taylor approximation of the GP prior term in their ELBO. However, this significantly increases implementation complexity and gives rise to potential risks in ignoring curvature. We take a different approach utilising sparse GPs. We desire a model that can scale to large datasets, like Hensman et al. (2013), while also being able to directly compute variational parameters from the latent regression dataset, like Titsias (2009). To this end, we take a mini-batch of the data, $\mathbf{X}_b \subset \mathbf{X}$, $\mathbf{Y}_b \subset \mathbf{Y}$, and with the network $\tilde{q}_\phi(\cdot)$ create a mini-batch of the latent dataset \mathbf{X}_b , $\tilde{\mathbf{y}}_b$, $\tilde{\boldsymbol{\sigma}}_b$. Following Titsias (2009), with Equations 4 and 5 for the optimal $\boldsymbol{\mu}_T$ and \mathbf{A}_T , we analytically

¹As an aside, this sparse GP-VAE ELBO may also be derived in the standard way using $\text{KL}(q_S(\mathbf{Z}|\cdot) \parallel p_{\psi, \theta}(\mathbf{Z}|\mathbf{Y}, \mathbf{X}))$, see Appendix B.4.

compute stochastic estimates for each latent channel l given by

$$\begin{aligned}\Sigma_b^l &:= \mathbf{K}_{mm} + \frac{N}{b} \mathbf{K}_{mb} \text{diag}(\tilde{\sigma}_b^{-2}) \mathbf{K}_{bm}, \\ \mu_b^l &:= \frac{N}{b} \mathbf{K}_{mm} (\Sigma_b^l)^{-1} \mathbf{K}_{mb} \text{diag}(\tilde{\sigma}_b^{-2}) \tilde{\mathbf{y}}_b^l, \\ \mathbf{A}_b^l &:= \mathbf{K}_{mm} (\Sigma_b^l)^{-1} \mathbf{K}_{mm}.\end{aligned}\tag{9}$$

where $\mathbf{K}_{mb} = k_\theta(\mathbf{U}, \mathbf{X}_b) \in \mathbb{R}^{m \times b}$. For a full derivation of these estimators, see Appendix B.2. All these estimators are consistent, so they converge to the true values for $b \rightarrow N$. However, while Σ_b^l is an unbiased estimator for Σ^l , the same does not hold for μ_b^l and \mathbf{A}_b^l . We investigate the magnitude of the bias in Appendix C.4 finding that it is generally small in practice. We believe this result to be in line with sparse Gaussian process approximations that assume the whole dataset may be summarized by a set of inducing points. Alternatively, this may be interpreted as assuming that the dataset contains redundancy, that is, that we have more than enough data to learn the latent function. In such a case, (cheaply) learning an average of latent functions of multiple mini-batches would closely approximate (expensively) learning one latent function using the full dataset.

μ_b^l and \mathbf{A}_b^l parameterize the approximate posterior $q_S(\cdot)$ which is, therefore, a direct function of the data $\mathbf{X}_b, \mathbf{Y}_b$ and hence it is an amortized approximate posterior. By taking a mini-batch of data, one may assume that we may also compute $\mathcal{L}_T(\cdot)$ of the mini-batch latent dataset. However, note that such an $\mathcal{L}_T(\cdot)$ is a lower bound for $\log Z_{\phi, \theta}(\cdot)$ of the mini-batch latent dataset, not a lower bound for the full latent dataset. Instead, we use μ_b^l and \mathbf{A}_b^l along with \mathbf{U} and θ to compute the GP evidence lower bound of Hensman et al. (2013) given in Equation 7, which is also suitable to mini-batching and lower bounds the marginal likelihood of the full latent dataset. Finally, the evidence lower bound of our Sparse (Variational) Gaussian Process Variational Autoencoder, for a single mini-batch $\mathbf{X}_b, \mathbf{Y}_b$, is thus

$$\begin{aligned}\mathcal{L}_{SVGP-VAE}(\mathbf{U}, \psi, \phi, \theta) &:= \sum_{i=1}^b \mathbb{E}_{q_S} \left[\log p_\psi(\mathbf{y}_i | \mathbf{z}_i) - \log \tilde{q}_\phi(\mathbf{z}_i | \mathbf{y}_i) \right] \\ &\quad + \frac{b}{N} \sum_{l=1}^L \mathcal{L}_H^l(\mathbf{U}, \phi, \theta, \mu_b^l, \mathbf{A}_b^l),\end{aligned}\tag{10}$$

where each $\mathcal{L}_H^l(\cdot)$ is computed using the mini-batch of the latent dataset $\mathbf{X}_b, \tilde{\mathbf{y}}_b^l, \tilde{\sigma}_b^l$. By naturally combining well known approaches, we arrive at a sparse GP-VAE that is both amortized and can be trained using mini-batches. The VAE parameters ϕ, ψ , inducing points \mathbf{U} , and the GP kernel θ can all be optimized jointly in an end-to-end fashion as we show in the next section.

Also note that during training, μ_b^1, \dots, μ_b^L and $\mathbf{A}_b^1, \dots, \mathbf{A}_b^L$ are computed from a mini-batch $\mathbf{X}_b, \mathbf{Y}_b$. However at test time, given a new dataset, all available data \mathbf{X}, \mathbf{Y} may be used to compute the μ^1, \dots, μ^L and $\mathbf{A}^1, \dots, \mathbf{A}^L$. The Gaussian process structure places no theoretical restriction upon the number of observations that are incorporated into the approximate posterior parameters, any amount of data can be pooled simply according to the kernel operations. In contrast, neural networks typically assume fixed input and output sizes and pooling data in a principled way requires much more attention.

While we have treated the auxiliary data \mathbf{X} as observed throughout this section, our model can also be used when \mathbf{X} is not given (or is only partly observed). In such cases, we make use of the Gaussian Process Latent Variable Model (GP-LVM) introduced by Lawrence (2004) to learn the missing part of \mathbf{X} , similar to what is done in Casale et al. (2018). In SVGP-VAE, (missing parts of) \mathbf{X} can be learned jointly with the rest of the model parameters.

4 Experiments

We compared our proposed model with existing approaches measuring both performance and scalability on some simple synthetic data and large high-dimensional benchmark datasets. Implementation details can be found in Appendix A and additional experiments in Appendix C. The implementation of our model as well as our experiments are publicly available at <https://github.com/ratschlab/SVGP-VAE>.

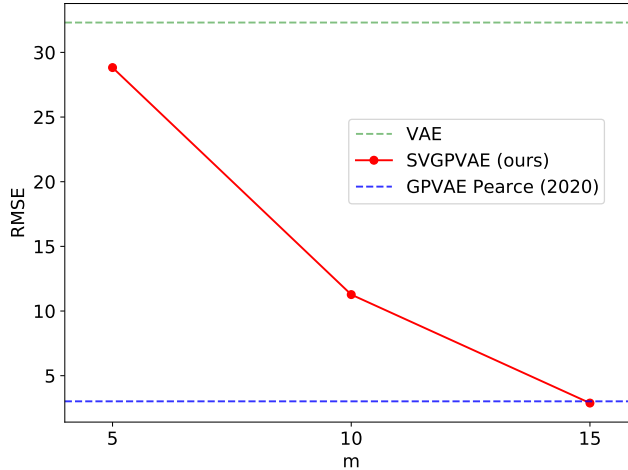


Figure 1: Performance of our SVGP-VAE models as a function of the number of inducing points. We see that as we increase the number of inducing points, the performance gracefully approaches the one of the exact GP-VAE baseline model.

4.1 Synthetic moving ball data

The moving ball data was utilized in Pearce (2020). It consists of black-and-white videos of a moving circle, where the 2D trajectory is sampled from a GP with radial basis function (RBF) kernel. The goal is to reconstruct the correct underlying trajectory in the two-dimensional latent space from the frames in pixel space. Since the videos are short (30 frames), full GP inference is still feasible in this setting, such that we can compare our sparse approach against the gold standard. Note that due to the small dataset size we do not perform mini-batching within each video here.

Scaling behavior. We see in Figure 1 that as we increase the number of inducing points our method uses, its performance in terms of root mean squared error (RMSE) approaches the performance of the full GP baselines. It reaches the baseline performance already with 15 inducing points, which is half the number of data points in the trajectory and therefore four times less computationally intensive than the baseline. The reconstructions of the trajectories also qualitatively agree with the baseline, as can be seen in Figure 2.

Optimization of kernel parameters. Another advantage of our proposed method over the previous approaches is that it is agnostic to the kernel choice and even allows to optimize the kernel parameters (and thereby learn a better kernel) jointly during training. In Pearce (2020), joint optimization of kernel parameters was not considered, while in Casale et al. (2018) a special training regime is deployed where VAE and GP parameters are optimized at different stages. Since the moving ball data is generated by a GP, we know the optimal kernel length scale for the RBF kernel in this case, which is namely the one of the generating process. We optimized the length scale of our SVGP-VAE kernel and found that when using a sufficient number of inducing points, we indeed recover the true length scale almost perfectly (Fig. 3). Note that when too few inducing points are used, the *effective* length scale of the observed process in the subspace spanned by these inducing points is indeed larger, since some of the variation in the data will be orthogonal to that subspace. It is thus to be expected that our model would also choose a larger length scale to model the observations in this subspace.

Optimization of inducing points. When working with sparse Gaussian processes, the selection of inducing point locations can often be crucial for the quality of the approximation (Titsias, 2009; Fortuin et al., 2018; Burt et al., 2019). In our model, we can optimize these inducing point locations jointly with the other components. On the moving ball data, since the trajectories are generated from stationary GPs, the optimal inducing point locations should be roughly equally spaced along

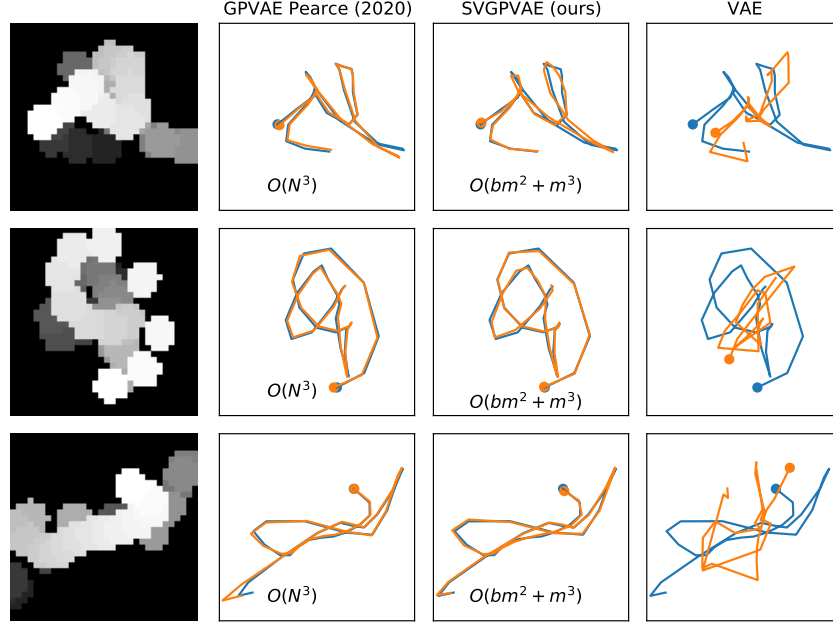


Figure 2: Reconstructions of the latent trajectories for the moving ball data. Frames of each test video are overlaid and shaded by time in the first column. Ground truth trajectories are depicted in blue, while predicted trajectories are shown in orange. We can see that the standard VAE fails to model the trajectories faithfully, while the GP-VAE models (including our sparse approximation) match them closely. Note that $b = N$ in SVGPP-VAE for this experiment. For SVGPP-VAE, the number of inducing points was set to $m = 15$.

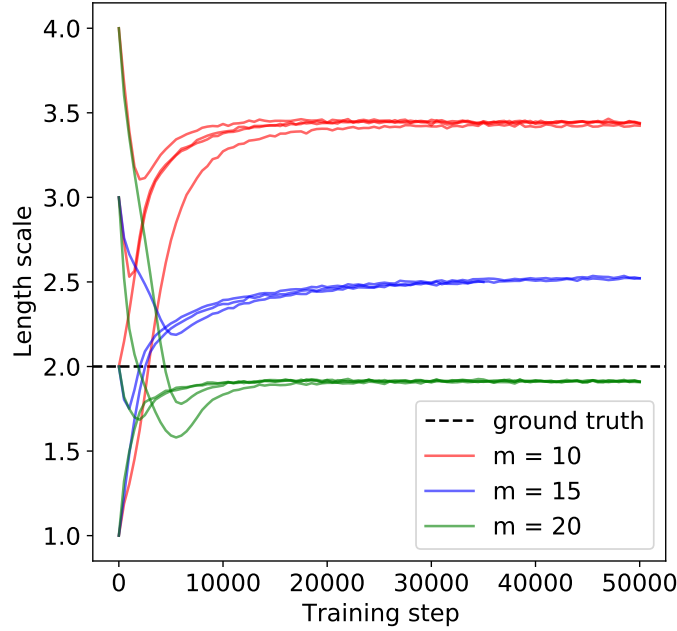


Figure 3: Optimized length scales of our SVGPP-VAE model during training on the moving ball data. With sufficiently many inducing points, the model recovers the true length scale of the generating process.

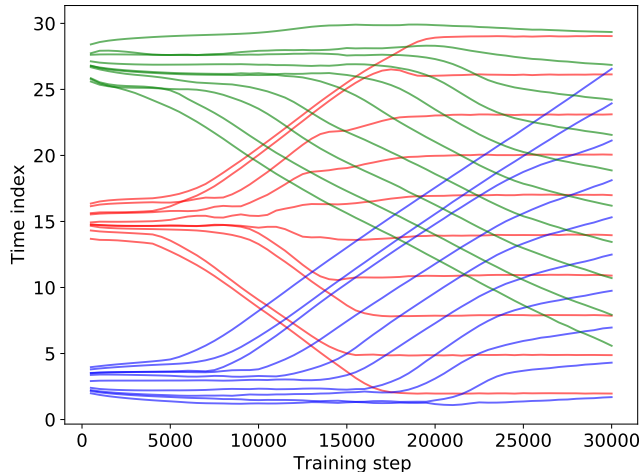


Figure 4: Optimized inducing points of our SVGP-VAE model during training on the moving ball data for three different (suboptimal) initializations. We can see that the model correctly learns to spread the inducing points evenly over the time series, which should be expected as a stationary GP kernel is used in the data generating process.

Table 1: Results on the rotated MNIST digit 3 dataset. Reported here are mean values together with standard deviations based on 5 runs. We see that our proposed model outperforms the baselines while still being more scalable than the Casale et al. (2018) model.

	MSE	GP complexity	Time/epoch [s]
CVAE (Sohn et al., 2015)	0.0796 ± 0.0023	-	0.39 ± 0.01
GP-VAE (Casale et al., 2018)	0.0370 ± 0.0012	$\mathcal{O}(NH^2)$	19.10 ± 0.66
SVGP-VAE (ours)	0.0251 ± 0.0005	$\mathcal{O}(bm^2 + m^3)$	1.90 ± 0.02

the time dimension. When we adversarially initialize the inducing points in a small region of the time series, we see that the model pushes them apart over the course of training and converges to this optimal spacing (Fig. 4). Together with the previous experiment, these observations suggest that the model is able to choose close-to-optimal inducing points and kernel functions in a data-driven way during the normal training process.

4.2 Conditional generation of rotated MNIST digits

To benchmark our model against existing scalable GP-VAE approaches, we follow the experimental setup from Casale et al. (2018) and use rotated MNIST digits (LeCun et al., 1998) in a conditional generation task. The task is to condition on a number of digits that have been rotated at different angles and to generate an image of one of these digits rotated at an unseen angle. In the original work, they consider 400 images of the digit 3, each rotated at multiple angles in $[0, 2\pi]$. Using identical architectures, kernel and dataset ($N = 4050$), we report results for both the GP-VAE of Casale et al. (2018) and our SVGP-VAE. The full GP-VAE model from Pearce (2020) can not be applied to this size of data, hence it is omitted. As an alternative baseline, we report results for a conditional VAE (CVAE) (Sohn et al., 2015). We use the GECO algorithm (Rezende and Viola, 2018) to train our SVGP-VAE model, which greatly improves the stability of the training procedure.

Performance of conditional generation. We see in Table 1 that our proposed model outperforms the baselines in terms of MSE, while still being computationally more efficient than the model from

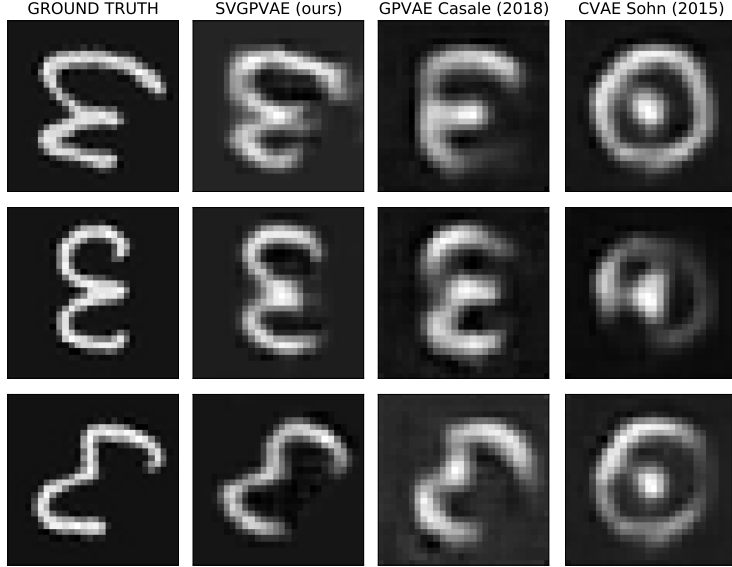


Figure 5: Conditionally generated rotated MNIST images. The generations of our proposed model are qualitatively more faithful to the ground truth. For more examples see Appendix C.3.

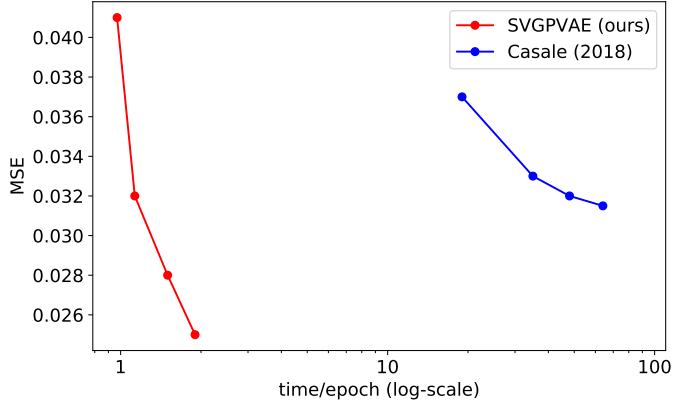


Figure 6: Performance of our proposed model with different numbers of inducing points and the Casale et al. (2018) model with different kernel dimensionalities as a function of runtime. For the SVGP-VAE, we consider four different configurations of inducing points, while for the Casale et al. (2018) model, we use four different dimensions of the linear kernel: $m, M \in \{8, 16, 24, 32\}$.

Casale et al. (2018) (in theory and practice).² This can also be seen visually in Figure 5 as our model produces the most faithful generations. For the SVGP-VAE, the number of inducing points was set to $m = 32$ and the batch size was set to $b = 256$. For the GP-VAE (Casale et al., 2018), the low-rank matrix factor H depends on the dimension of the linear kernel M used in their model ($M = 8$ and $H = 128$).

Tradeoff between runtime and performance. The performance of our sparse approximation can be increased by choosing a larger number of inducing points, at a quadratic cost in terms of runtime. The Casale et al. (2018) model, while being more restricted in its kernel choice, offers a similar tradeoff between runtime and performance by choosing a different dimensionality for the low-rank

²Note that in their paper, Casale et al. (2018) report a performance of 0.028 on this task. However, their code for the MNIST experiment is not openly available and we could not reproduce this result with our reimplementations (which is also available at <https://github.com/ratschlab/SVGP-VAE>).

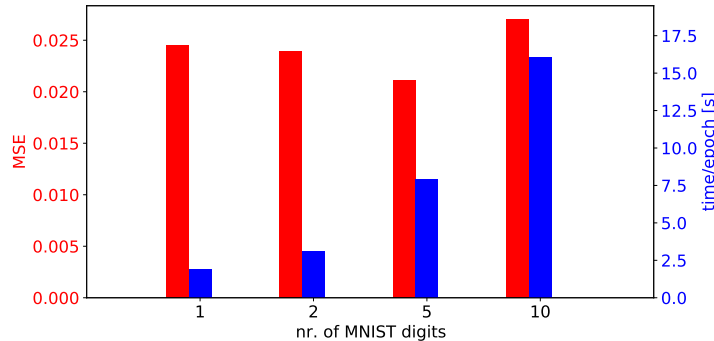


Figure 7: Performance and runtime of our proposed model on differently sized subsets of the MNIST dataset, including the full set. We see that the performance stays roughly the same, regardless of dataset size, while the runtime grows linearly as expected. The size of each dataset equals $4050 \times$ nr. of MNIST digits.

linear kernel used in their latent space (see Appendix B.3). In Figure 6 we depict performance for both models when varying the number of inducing points and the dimension of the linear kernel, respectively. We observe that SVGP-VAE, besides being much faster, exhibits a steeper decline in the MSE as the model’s capacity is increased.

Scaling to larger data. As mentioned above, Casale et al. (2018) restrict their experiment to a small subset of the MNIST dataset and indeed we did also not manage to scale their model to the whole dataset on our hardware (11 GB GPU memory). Our SVGP-VAE, however, is easily scalable to such dataset sizes. We report its performance on larger subsets of MNIST (including the full dataset) in Figure 7. We see that the performance of our proposed model does not deteriorate with increased dataset size, while the runtime grows linearly as expected. All in all, we thus see that our model is more flexible than the previous approaches, scales to larger datasets, and achieves a better performance at lower computational cost.

5 Conclusion

We have proposed a novel sparse inference method for GP-VAE models and have shown theoretically and empirically that it is more scalable than existing approaches, while achieving competitive performance. Our approach bridges the gap between sparse variational GP approximations and GP-VAE models, thus enabling the utilization of a large body of work in the sparse GP literature. As such, it represents an important step towards unlocking the possibility to perform GP regression on large datasets with complex likelihoods (e.g., natural images).

Fruitful avenues for future work include considering even more recently proposed sparse GP approaches (Cheng and Boots, 2017; Evans and Nair, 2020) and comparing our proposed scalable GP-VAE solution against other families of deep generative models (Mirza and Osindero, 2014; Es-lami et al., 2018). This would help identify real-world applications where GP-VAEs could be most impactful.

References

- M. Ashman, J. So, W. Tebbutt, V. Fortuin, M. Pearce, and R. E. Turner. Sparse gaussian process variational autoencoders. *arXiv preprint arXiv:2010.10177*, 2020.
- D. R. Burt, C. E. Rasmussen, and M. Van Der Wilk. Rates of convergence for sparse variational gaussian process regression. *arXiv preprint arXiv:1903.03571*, 2019.
- F. P. Casale, A. Dalca, L. Saglietti, J. Listgarten, and N. Fusi. Gaussian process prior variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 10369–10380, 2018.
- C.-A. Cheng and B. Boots. Variational inference for gaussian process models with linear complexity. In *Advances in Neural Information Processing Systems*, pages 5184–5194, 2017.

- T. R. Davidson, L. Falorsi, N. De Cao, T. Kipf, and J. M. Tomczak. Hyperspherical variational auto-encoders. *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, 2:856–865, 2018.
- N. Dilokthanakul, P. A. Mediano, M. Garnelo, M. C. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.
- S. M. A. Eslami, D. Jimenez Rezende, F. Besse, F. Viola, A. S. Morcos, M. Garnelo, A. Ruderman, A. A. Rusu, I. Danihelka, K. Gregor, D. P. Reichert, L. Buesing, T. Weber, O. Vinyals, D. Rosenbaum, N. Rabinowitz, H. King, C. Hillier, M. Botvinick, D. Wierstra, K. Kavukcuoglu, and D. Hassabis. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018.
- T. W. Evans and P. B. Nair. Quadruply stochastic gaussian processes. *arXiv preprint arXiv:2006.03015*, 2020.
- V. Fortuin, G. Dresdner, H. Strathmann, and G. Rätsch. Scalable gaussian processes on discrete domains. *arXiv preprint arXiv:1810.10368*, 2018.
- V. Fortuin, D. Baranchuk, G. Rätsch, and S. Mandt. Gp-vae: Deep probabilistic time series imputation. In *International Conference on Artificial Intelligence and Statistics*, pages 1651–1661. PMLR, 2020.
- D. A. Harville. Matrix algebra from a statistician’s perspective, 1998.
- H. V. Henderson and S. R. Searle. On deriving the inverse of a sum of matrices. *Siam Review*, 23(1):53–60, 1981.
- J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*, 2013.
- M. J. Johnson, D. K. Duvenaud, A. Wiltchko, R. P. Adams, and S. R. Datta. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in neural information processing systems*, pages 2946–2954, 2016.
- D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- D. P. Kingma and M. Welling. An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*, 2019.
- D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improved variational inference with inverse autoregressive flow. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4743–4751. Curran Associates, Inc., 2016.
- A. Kopf, V. Fortuin, V. R. Somnath, and M. Claassen. Mixture-of-experts variational auto-encoder for clustering and generating from similarity-based representations. *arXiv preprint arXiv:1910.07763*, 2019.
- N. D. Lawrence. Gaussian Process Latent Variable Models for Visualisation of High Dimensional Data. *Advances in Neural Information Processing Systems*, 2004. doi: 10.1115/OMAE2008-57170.
- Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- W. Lin, N. Hubacher, and M. E. Khan. Variational message passing with structured inference networks. *arXiv preprint arXiv:1803.05589*, 2018.
- L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.
- M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
- M. Pearce. The gaussian process prior vae for interpretable latent dynamics from pixels. In *Symposium on Advances in Approximate Bayesian Inference*, pages 1–12, 2020.

- J. Quiñonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6(Dec):1939–1959, 2005.
- D. J. Rezende and S. Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.
- D. J. Rezende and F. Viola. Taming vaes. *arXiv preprint arXiv:1810.00597*, 2018.
- D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.
- E. Snelson and Z. Ghahramani. Sparse gaussian processes using pseudo-inputs. In *Advances in neural information processing systems*, pages 1257–1264, 2006.
- K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pages 3483–3491, 2015.
- M. Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574, 2009.

A Experimental details

Here we report the details and parameter settings for our experiments to foster reproducibility.

A.1 Moving ball experiment

For the moving ball experiment described in Section 4.1, we use the same neural networks architectures and training setting as in Pearce (2020).

Table A.1: Parameter settings for the moving ball experiment.

Parameter	Value
Nr. of feedforward layers in inference network	2
Nr. of feedforward layers in generative network	2
Width of a hidden feedforward layer	500
Dimensionality of latent space (L)	2
Activation function	<i>tanh</i>
Learning rate	0.001
Optimizer	Adam
Nr. of epochs	25000
Nr. of frames in each video (N)	30
Dimension of each frame	32×32

A squared-exponential GP kernel with length scale $l = 2$ was used. For the exact data generation procedure, we refer to Pearce (2020). During training, 35 videos were generated in each epoch. The test MSE is reported on a held-out set of 350 videos. For the Adam optimizer (Kingma and Ba, 2014), the default Tensorflow parameters are used.

A.2 MNIST experiment

For the rotated MNIST experiment described in Section 4.2, we used the same neural networks architectures as in Casale et al. (2018): three convolutional layers followed by a fully connected layer in the inference network and vice-versa in the generative network.

Table A.2: Neural networks architectures for the MNIST experiment.

Parameter	Value
Nr. of CNN layers in inference network	3
Nr. of CNN layers in generative network	3
Nr. of filters per CNN layer	8
Filter size	3×3
Nr. of feedforward layers in inference network	1
Nr. of feedforward layers in generative network	1
Activation function in CNN layers	ELU
Dimensionality of latent space (L)	16

The SVGP-VAE model is trained for 1000 epochs with a batch size of 256. The Adam optimizer Kingma and Ba (2014) is used with its default parameters and a learning rate of 0.001. Moreover, the GECO algorithm (Rezende and Viola, 2018) was used for training our SVGP-VAE model in this experiment. The reconstruction parameter in GECO was set to $\kappa = 0.020$ in all reported experiments.

For the GP-VAE model from Casale et al. (2018), we used the same training procedure as reported in Casale et al. (2018). We have observed in our reimplementation that a joint optimization at the end does not improve performance. Hence, we report results for the regime where the VAE parameters are optimized for the first 100 epochs, followed by 100 epochs during which the GP parameters are optimized. Moreover, we could not get their proposed low-memory modified forward pass to work, so in our reimplementation the entire dataset is loaded into the memory at one point during

the forward pass. Our reimplementation of the GP-VAE model from Casale et al. (2018) is publicly available at <https://github.com/ratschlab/SVGP-VAE>.

For both models, the GP kernel proposed in Casale et al. (2018) is used. For more details on the kernel, we refer to Appendix B.3. Note that the auxiliary data \mathbf{X} is only partially observed in this experiment — for both models we use a GP-LVM to learn the missing parts of \mathbf{X} . For both models, we use Principal Component Analysis (PCA) to initialize the GP-LVM vectors, as it was observed to lead to a slight increase in performance. PCA is also used in SVGP-VAE to initialize the inducing points. For more details see Appendix C.1.

A.3 On the training of GP-VAE models (practitioner’s perspective)

While working on implementations of different GP-VAE models, we have noticed that balancing the absolute magnitudes of the reconstruction and the KL-term is critical for achieving optimal results, even more so than in standard VAE models. In Fortuin et al. (2020), this was tackled by introducing a weighting β parameter, whereas in Casale et al. (2018) a CV search on the noise parameter σ_y^2 of the likelihood $p_\psi(\mathbf{y}_i|\mathbf{z}_i)$ is performed. One downside of both solutions is that they introduce (yet) another training hyperparameter that needs to be manually tuned for every new dataset/model architecture considered.

To get around this, we instead used the GECO algorithm (Rezende and Viola, 2018) to train our SVGP-VAE in the rotated MNIST experiment. Compared to the original GECO algorithm in Rezende and Viola (2018), where the maximization objective is the KL divergence between a standard Gaussian prior and the variational distribution, the GECO maximization objective in the SVGP-VAE is composed of a cross-entropy term $\mathbb{E}_{q_S}[\log \tilde{q}_\phi(\cdot)]$ and a sparse GP ELBO $\mathcal{L}_H(\cdot)$. We have observed that GECO greatly simplifies training of GP-VAE models as it eliminates the need to manually tune the different magnitudes of the ELBO terms. Based on this, we would make a general recommendation for GECO to be used for training such models.

B Supporting derivations

B.1 Vanishing of the inference network in GP-VAE with ELBO from Hensman et al. (2013)

In this section we show that working with the sparse GP approach presented in Hensman et al. (2013) leads to vanishing of the inference network parameters ϕ in the GP-VAE model from Pearce (2020). Recall that the sparse GP posterior from Hensman et al. (2013) for the l -th latent channel has the form

$$q_S^l(\mathbf{z}_{1:N}^l|\cdot) = \mathcal{N}(\mathbf{z}_{1:N}^l | \mathbf{K}_{Nm} \mathbf{K}_{mm}^{-1} \boldsymbol{\mu}^l, \mathbf{K}_{NN} - \mathbf{K}_{Nm} \mathbf{K}_{mm}^{-1} \mathbf{K}_{mN} + \mathbf{K}_{Nm} \mathbf{K}_{mm}^{-1} \mathbf{A}^l \mathbf{K}_{mm}^{-1} \mathbf{K}_{mN})$$

with $\boldsymbol{\mu}^l \in \mathbb{R}^m$, $\mathbf{A}^l \in \mathbb{R}^{m \times m}$ as free variational parameters, while the sparse GP ELBO for the l -th latent channel is given as

$$\begin{aligned} \mathcal{L}_H^l(\mathbf{U}, \boldsymbol{\mu}^l, \mathbf{A}^l, \phi, \theta) = \sum_{i=1}^N \left\{ \log \mathcal{N}(\tilde{y}_{l,i} | \mathbf{k}_i^T \mathbf{K}_{mm}^{-1} \boldsymbol{\mu}^l, \tilde{\sigma}_{l,i}^{-2}) - \frac{1}{2\tilde{\sigma}_{l,i}^{-2}} (\tilde{k}_{ii} + \text{Tr}(\mathbf{A}^l \Lambda_i)) \right\} \\ - KL(q_S^l(\mathbf{f}_m|\cdot) || p_\theta(\mathbf{f}_m|\cdot)) \end{aligned}$$

with $q_S^l(\mathbf{f}_m|\cdot) = \mathcal{N}(\mathbf{f}_m | \boldsymbol{\mu}^l, \mathbf{A}^l)$ and $p_\theta(\mathbf{f}_m|\cdot) = \mathcal{N}(\mathbf{f}_m | \mathbf{0}, \mathbf{K}_{mm})$. \mathbf{k}_i represents the i -th column of \mathbf{K}_{mN} , $\Lambda_i := \mathbf{K}_{mm}^{-1} \mathbf{k}_i \mathbf{k}_i^T \mathbf{K}_{mm}^{-1}$ and \tilde{k}_{ii} is the i -th diagonal element of $\mathbf{K}_{NN} - \mathbf{K}_{Nm} \mathbf{K}_{mm}^{-1} \mathbf{K}_{mN}$. As mentioned in Section 3, \mathcal{L}_H^l depends on the inference network parameters ϕ through the (amortized) l -th latent dataset $\tilde{\mathbf{y}}_l = \mu_\phi^l(\mathbf{Y})$, $\tilde{\sigma}_l = \sigma_\phi^l(\mathbf{Y})$.

Note that the full sparse GP posterior equals $q_S(\mathbf{Z}) = \prod_{l=1}^L q_S^l(\mathbf{z}_{1:N}^l|\cdot)$. Similarly, the full sparse GP ELBO is $\mathcal{L}_H = \sum_{l=1}^L \mathcal{L}_H^l(\mathbf{U}, \boldsymbol{\mu}^l, \mathbf{A}^l, \phi, \theta)$.

Proposition B.1 *For the l -th latent channel in the GP-VAE model with the bound from Hensman et al. (2013), the following relation holds:*

$$\mathbb{E}_{q_S^l} [\log \tilde{q}_\phi(\mathbf{z}_{1:N}^l | \mathbf{Y})] = \sum_{i=1}^N \left\{ \log \mathcal{N}(\tilde{y}_{l,i} | \mathbf{k}_i^T \mathbf{K}_{mm}^{-1} \boldsymbol{\mu}^l, \tilde{\sigma}_{l,i}^{-2}) - \frac{1}{2\tilde{\sigma}_{l,i}^{-2}} (\tilde{k}_{ii} + \text{Tr}(\mathbf{A}^l \Lambda_i)) \right\}.$$

Proof For notational convenience, define $\tilde{D}_l := \text{diag}(\tilde{\sigma}_l^2)$ and $\mathbf{B} := \mathbf{K}_{Nm} \mathbf{K}_{mm}^{-1}$. Also recall that $\tilde{q}_\phi(\mathbf{z}_{1:N}^l | \mathbf{Y}) = \mathcal{N}(\mathbf{z}_{1:N}^l | \tilde{\mathbf{y}}_l, \tilde{D}_l)$. Using the formula for the cross-entropy between two multivariate Gaussian distributions, we proceed as

$$\begin{aligned} \mathbb{E}_{q_S^l} [\log \tilde{q}_\phi(\mathbf{z}_{1:N}^l | \mathbf{Y})] &= -\frac{N}{2} \log(2\pi) - \frac{1}{2} \log |\tilde{D}_l| - \frac{1}{2} (\tilde{\mathbf{y}}_l - \mathbf{B} \boldsymbol{\mu}^l)^T \tilde{D}_l^{-1} (\tilde{\mathbf{y}}_l - \mathbf{B} \boldsymbol{\mu}^l) - \frac{1}{2} \text{Tr}(\tilde{D}_l^{-1} (\tilde{\mathbf{K}} + \mathbf{B} \mathbf{A}^l \mathbf{B}^T)) \\ &= \log \mathcal{N}(\tilde{\mathbf{y}}_l | \mathbf{B} \boldsymbol{\mu}^l, \tilde{D}_l) - \frac{1}{2} \text{Tr}(\tilde{D}_l^{-1} \tilde{\mathbf{K}}) - \frac{1}{2} \text{Tr}(\tilde{D}_l^{-1} \mathbf{B} \mathbf{A}^l \mathbf{B}^T). \end{aligned}$$

It remains to show that the last trace term equals $\sum_{i=1}^N \tilde{\sigma}_{l,i}^{-2} \text{Tr}(\mathbf{A}^l \Lambda_i)$, which follows from

$$\text{Tr}(\tilde{D}_l^{-1} \mathbf{B} \mathbf{A}^l \mathbf{B}^T) = \text{Tr}(\mathbf{A}^l \mathbf{B}^T \tilde{D}_l^{-1} \mathbf{B}) = \text{Tr}(\mathbf{A}^l \mathbf{K}_{mm}^{-1} (\sum_{i=1}^N \tilde{\sigma}_{l,i}^{-2} \mathbf{k}_i \mathbf{k}_i^T) \mathbf{K}_{mm}^{-1}) = \sum_{i=1}^N \tilde{\sigma}_{l,i}^{-2} \text{Tr}(\mathbf{A}^l \Lambda_i).$$

Proposition B.2 *The GP-VAE ELBO with the bound from Hensman et al. (2013) reduces to*

$$\mathcal{L}_{PH}(\mathbf{U}, \psi, \theta, \boldsymbol{\mu}^{1:L}, \mathbf{A}^{1:L}) = \sum_{i=1}^N \mathbb{E}_{q_S} \left[\log p_\psi(\mathbf{y}_i | \mathbf{z}_i) \right] + \sum_{l=1}^L KL(q_S^l(\mathbf{f}_m|\cdot) || p_\theta(\mathbf{f}_m|\cdot))$$

Proof Using the above proposition, we have

$$\begin{aligned}
& \mathbb{E}_{q_S} \left[\sum_{i=1}^N \log p_\psi(\mathbf{y}_i | \mathbf{z}_i) - \log \tilde{q}_\phi(\mathbf{z}_i | \mathbf{y}_i) \right] + \sum_{l=1}^L \mathcal{L}_H^l \\
&= \mathbb{E}_{q_S} \left[\sum_{i=1}^N \log p_\psi(\mathbf{y}_i | \mathbf{z}_i) \right] - \mathbb{E}_{q_S} \left[\sum_{l=1}^L \log \tilde{q}_\phi(\mathbf{z}_{1:N}^l | \mathbf{Y}) \right] + \sum_{l=1}^L \mathcal{L}_H^l \\
&= \mathbb{E}_{q_S} \left[\sum_{i=1}^N \log p_\psi(\mathbf{y}_i | \mathbf{z}_i) \right] - \sum_{l=1}^L \left(\mathbb{E}_{q_S^l} [\log \tilde{q}_\phi(\mathbf{z}_{1:N}^l | \mathbf{Y})] - \mathcal{L}_H^l \right) \\
&= \mathbb{E}_{q_S} \left[\sum_{i=1}^N \log p_\psi(\mathbf{y}_i | \mathbf{z}_i) \right] + \sum_{l=1}^L KL(q_S^l(\mathbf{f}_m | \cdot) \parallel p_\theta(\mathbf{f}_m | \cdot)) . \quad \square
\end{aligned}$$

Observe that in $\mathcal{L}_{PH}(\cdot)$ all terms that include $\tilde{\mathbf{y}}_l$ or $\tilde{\sigma}_l$ cancel out, hence such ELBO is independent of the inference network parameters ϕ .

B.2 Monte Carlo estimators in the SVGP-VAE

The idea behind the estimators used in q_S in our SVGP-VAE is based on the work presented in Evans and Nair (2020). The main insight is to rewrite the matrix operations as expectations with respect to the empirical distribution of the training data. Those expectations are then approximated with Monte Carlo estimators.

Recall that the (amortized) latent dataset for the l -th channel is denoted by $\{\mathbf{X}, \tilde{\mathbf{y}}_l, \tilde{\sigma}_l\}$, with $\tilde{\mathbf{y}}_l := \mu_\phi^l(\mathbf{Y})$ and $\tilde{\sigma}_l := \sigma_\phi^l(\mathbf{Y})$. For notational convenience, additionally denote $\tilde{D}_l := \text{diag}(\tilde{\sigma}_l^2)$. First, observe that the matrix product $\mathbf{K}_{mN} \tilde{D}_l^{-1} \mathbf{K}_{Nm}$ in Σ^l can be rewritten as a sum over data points $\sum_{i=1}^N B_i(\mathbf{x}_i, \mathbf{y}_i)$ with

$$B_i(\mathbf{x}_i, \mathbf{y}_i) := \frac{1}{\tilde{\sigma}_{l,i}^2} \begin{bmatrix} k_\theta(\mathbf{u}_1, \mathbf{x}_i) k_\theta(\mathbf{u}_1, \mathbf{x}_i) & \dots & k_\theta(\mathbf{u}_1, \mathbf{x}_i) k_\theta(\mathbf{u}_m, \mathbf{x}_i) \\ \vdots & \ddots & \vdots \\ k_\theta(\mathbf{u}_m, \mathbf{x}_i) k_\theta(\mathbf{u}_1, \mathbf{x}_i) & \dots & k_\theta(\mathbf{u}_m, \mathbf{x}_i) k_\theta(\mathbf{u}_m, \mathbf{x}_i) \end{bmatrix} .$$

Let \bar{b} represent a set of indices of data points in the current batch with size b . Moreover, define $\mathbf{K}_{bm} \in \mathbb{R}^{b \times m}$, $\tilde{D}_{l,b} \in \mathbb{R}^{b \times b}$, $\tilde{\mathbf{y}}_b^l \in \mathbb{R}^b$ as the sub-sampled versions of $\mathbf{K}_{Nm} \in \mathbb{R}^{N \times m}$, $\tilde{D}_l \in \mathbb{R}^{N \times N}$ and $\tilde{\mathbf{y}}_l \in \mathbb{R}^N$, respectively, consisting only of data points in \bar{b} . An (unbiased) Monte Carlo estimator for Σ^l is then derived as follows

$$\begin{aligned}
\Sigma^l &= \mathbf{K}_{mm} + \mathbf{K}_{mN} \tilde{D}_l^{-1} \mathbf{K}_{Nm} = \mathbf{K}_{mm} + N \sum_{i=1}^N \frac{1}{N} B_i(\mathbf{x}_i, \mathbf{y}_i) = \mathbf{K}_{mm} + N \cdot \mathbb{E}_{i \sim \{1, \dots, N\}} [B_i(\mathbf{x}_i, \mathbf{y}_i)] \\
&\approx \mathbf{K}_{mm} + \frac{N}{b} \sum_{i \in \bar{b}} B_i(\mathbf{x}_i, \mathbf{y}_i) = \mathbf{K}_{mm} + \frac{N}{b} \mathbf{K}_{mb} \tilde{D}_{l,b}^{-1} \mathbf{K}_{bm} =: \Sigma_b^l .
\end{aligned}$$

Additionally, define $\mathbf{c}_l := \mathbf{K}_{mN} \tilde{D}_l^{-1} \tilde{\mathbf{y}}_l$ and proceed similarly as above

$$\mathbf{c}_l = \sum_{i=1}^n b_i(\mathbf{x}_i, \mathbf{y}_i) = N \cdot \mathbb{E}_{i \sim \{1, \dots, N\}} [b_i(\mathbf{x}_i, \mathbf{y}_i)] \approx \frac{N}{b} \sum_{i \in \bar{b}} b_i(\mathbf{x}_i, \mathbf{y}_i) = \frac{N}{b} \mathbf{K}_{mb} \tilde{D}_{l,b}^{-1} \tilde{\mathbf{y}}_b^l =: \mathbf{c}_b^l ,$$

where

$$b_i(\mathbf{x}_i, \mathbf{y}_i) := \frac{\tilde{y}_{l,i}}{\tilde{\sigma}_{l,i}^2} \begin{bmatrix} k_\theta(\mathbf{u}_1, \mathbf{x}_i) \\ \vdots \\ k_\theta(\mathbf{u}_m, \mathbf{x}_i) \end{bmatrix} .$$

The estimators for μ_T^l and A_T^l are then obtained using a plug-in approach,

$$\begin{aligned}\mu_T^l &= \mathbf{K}_{mm}(\Sigma^l)^{-1} \mathbf{c}_l \approx \mathbf{K}_{mm}(\Sigma_b^l)^{-1} \mathbf{c}_b^l =: \mu_b^l, \\ A_T^l &= \mathbf{K}_{mm}(\Sigma^l)^{-1} \mathbf{K}_{mm} \approx \mathbf{K}_{mm}(\Sigma_b^l)^{-1} \mathbf{K}_{mm} =: A_b^l.\end{aligned}$$

Note that neither of the above estimators is unbiased, since both depend on the inverse $(\Sigma_b^l)^{-1}$. For the empirical investigation of the magnitude of the bias, see Appendix C.4. However, A_b^l can be shown to be approximately (up to the first order Taylor approximation) unbiased.

Proposition B.3 *For the estimator A_b^l in SVGP-VAE, it holds that*

$$\mathbb{E}[A_b^l] - A_T^l \approx 0.$$

Proof Note that expectation here is taken with respect to the empirical distribution of the training data, that is, $\mathbb{E}_{i \sim \{1, \dots, N\}}$. Using the definitions of A_b^l and A_T^l , we get

$$\mathbb{E}[A_b^l] - A_T^l = \mathbf{K}_{mm}(\mathbb{E}[(\Sigma_b^l)^{-1}] - (\Sigma^l)^{-1})\mathbf{K}_{mm},$$

so it remains to show that $\mathbb{E}[(\Sigma_b^l)^{-1}] - (\Sigma^l)^{-1} \approx 0$. To this end, we exploit the positive definiteness of the kernel matrix \mathbf{K}_{mm} and we approximate both inverse terms with the first order Taylor expansion:

$$\begin{aligned}(\Sigma_b^l)^{-1} &= (\mathbf{K}_{mm} + \frac{N}{b} \mathbf{K}_{mb} \tilde{D}_{l,b}^{-1} \mathbf{K}_{bm})^{-1} = \mathbf{K}_{mm}^{-\frac{1}{2}} (\mathbf{I} + \frac{N}{b} \mathbf{K}_{mm}^{-\frac{1}{2}} \mathbf{K}_{mb} \tilde{D}_{l,b}^{-1} \mathbf{K}_{bm} \mathbf{K}_{mm}^{-\frac{1}{2}})^{-1} \mathbf{K}_{mm}^{-\frac{1}{2}} \\ &\approx \mathbf{K}_{mm}^{-\frac{1}{2}} (\mathbf{I} - \frac{N}{b} \mathbf{K}_{mm}^{-\frac{1}{2}} \mathbf{K}_{mb} \tilde{D}_{l,b}^{-1} \mathbf{K}_{bm} \mathbf{K}_{mm}^{-\frac{1}{2}}) \mathbf{K}_{mm}^{-\frac{1}{2}} = \mathbf{K}_{mm}^{-1} - \frac{N}{b} \mathbf{K}_{mm}^{-1} \mathbf{K}_{mb} \tilde{D}_{l,b}^{-1} \mathbf{K}_{bm} \mathbf{K}_{mm}^{-1}.\end{aligned}$$

Similarly, we have $(\Sigma^l)^{-1} \approx \mathbf{K}_{mm}^{-1} - \mathbf{K}_{mm}^{-1} \mathbf{K}_{mN} \tilde{D}_l^{-1} \mathbf{K}_{Nm} \mathbf{K}_{mm}^{-1}$. Using this, we proceed as

$$\begin{aligned}\mathbb{E}[(\Sigma_b^l)^{-1}] - (\Sigma^l)^{-1} &\approx -\frac{N}{b} \mathbf{K}_{mm}^{-1} \mathbb{E}[\mathbf{K}_{mb} \tilde{D}_{l,b}^{-1} \mathbf{K}_{bm}] \mathbf{K}_{mm}^{-1} + \mathbf{K}_{mm}^{-1} \mathbf{K}_{mN} \tilde{D}_l^{-1} \mathbf{K}_{Nm} \mathbf{K}_{mm}^{-1} \\ &= -\frac{N}{b} \mathbf{K}_{mm}^{-1} \mathbb{E}\left[\sum_{i \in b} B_i(\mathbf{x}_i, \mathbf{y}_i)\right] \mathbf{K}_{mm}^{-1} + \mathbf{K}_{mm}^{-1} \left(\sum_{i=1}^N B_i(\mathbf{x}_i, \mathbf{y}_i)\right) \mathbf{K}_{mm}^{-1} \\ &= -N \mathbf{K}_{mm}^{-1} \mathbb{E}[B_i(\mathbf{x}_i, \mathbf{y}_i)] \mathbf{K}_{mm}^{-1} + N \mathbf{K}_{mm}^{-1} \mathbb{E}[B_i(\mathbf{x}_i, \mathbf{y}_i)] \mathbf{K}_{mm}^{-1} = 0\end{aligned}$$

□

Note that a similar proof technique unfortunately cannot be used to show that μ_b^l is approximately unbiased for μ_T^l , due to the product of two plug-in estimators that both depend on the data in the same batch.

B.3 Low-rank kernel matrix in Casale et al. (2018)

In the following, we present an approach from Casale et al. (2018) to reduce the cubic GP complexity in their GP-VAE model. Note that the exact approach is not given in Casale et al. (2018) and the derivation shown here is our best attempt at recreating the results.

In Casale et al. (2018), datasets composed of P unique objects observed in Q unique views are considered, for instance, images of faces captured from different angles. In total, this amounts to $N = P \cdot Q$ images. The auxiliary data consist of two sets of features $\mathbf{X} = [\mathbf{X}_o \ \mathbf{X}_v]$, with $\mathbf{X}_o \in \mathbb{R}^{N \times p_1}$ containing information about objects (e.g., drawing style of the digit or characteristics of the face) and $\mathbf{X}_v \in \mathbb{R}^{N \times p_2}$ containing information about views (e.g., an angle or position in space). Let $\mathbf{x}_i = [\mathbf{x}_{o,i} \ \mathbf{x}_{v,i}]$ denote auxiliary data for the i -th image (corresponding to the i -th row of the \mathbf{X} matrix). Additionally, denote by $\mathbf{P} \in \mathbb{R}^{P \times p_1}$ and $\mathbf{Q} \in \mathbb{R}^{Q \times p_2}$ matrices consisting of all unique object and view representations, respectively. A product kernel between a linear kernel for object information and a periodic kernel for view information is used:

$$k_\theta(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \exp\left(-\frac{2 \sin^2(\|\mathbf{x}_{v,i} - \mathbf{x}_{v,j}\|)}{l^2}\right) \cdot \mathbf{x}_{o,i}^T \mathbf{x}_{o,j}, \quad \theta = \{\sigma^2, l\}.$$

Exploiting the product and (partial) linear structure of the kernel and using properties of the Kronecker product, \mathbf{K}_{NN} can be written in a low-rank form as

$$\mathbf{K}_{NN}(\mathbf{X}, \mathbf{X}) = \mathbf{P}\mathbf{P}^T \otimes \mathbf{K}(\mathbf{Q}) = \mathbf{P}\mathbf{P}^T \otimes \mathbf{L}\mathbf{L}^T = (\mathbf{P} \otimes \mathbf{L})(\mathbf{P}^T \otimes \mathbf{L}^T) = (\mathbf{P} \otimes \mathbf{L})(\mathbf{P} \otimes \mathbf{L})^T =: \mathbf{V}\mathbf{V}^T,$$

where $\mathbf{K}(\mathbf{Q}) \in \mathbb{R}^{Q \times Q}$ is a kernel matrix of all unique view vectors based on the periodic kernel, \mathbf{L} is its Cholesky decomposition and $\mathbf{V} \in \mathbb{R}^{N \times H}$, $H = Q \cdot p_1$, is the obtained low-rank matrix ($H \ll N$ due to the assumption that the number of unique views Q is not large). For such matrices, the inverse and log-determinant can be computed in $O(NH^2)$ using a matrix inversion lemma (Henderson and Searle, 1981) and a matrix determinant lemma (Harville, 1998), respectively.

While the above approach elegantly reduces the GP complexity for a given dataset (for auxiliary data \mathbf{X} with a product structure), it is not readily extensible for other types of datasets (e.g. time series). In contrast, our SVGP-VAE makes no assumptions on neither the data nor the GP kernel used. Therefore, it is a more general solution to scale GP-VAE models.

B.4 Sparse GP-VAE based on Titsias (2009)

Using the sparse GP posterior q_S (Equation 3) and ELBO \mathcal{L}_T (Equation 6) from Titsias (2009) gives rise to the following sparse GP-VAE ELBO:

$$\mathcal{L}_{PT}(\mathbf{U}, \psi, \phi, \theta) := \sum_{l=1}^L \mathcal{L}_T^l(\mathbf{U}, \phi, \theta) + \sum_{i=1}^N \mathbb{E}_{q_S} \left[\log p_\psi(\mathbf{y}_i | \mathbf{z}_i) - \log \tilde{q}_\phi(\mathbf{z}_i | \mathbf{y}_i) \right].$$

In Section 3.3, we have outlined how to obtain the above sparse ELBO from the GP-VAE ELBO proposed in Pearce (2020). Alternatively, \mathcal{L}_{PT} can be derived in the standard way by directly considering the KL divergence between the sparse GP posterior and the (intractable) true posterior for the latent variables $\text{KL}(q_S(\mathbf{Z}|\cdot) || p_{\psi, \theta}(\mathbf{Z}|\mathbf{Y}, \mathbf{X}))$.

Following Titsias (2009), we consider the joint distribution of observed and *augmented* latent variables $p_{\psi, \theta}(\mathbf{Z}, \mathbf{F}_m, \mathbf{Y}|\mathbf{X})$ where $\mathbf{F}_m := [\mathbf{f}^1, \dots, \mathbf{f}^L]$, $\mathbf{f}^l := f^l(\mathbf{U}) \in \mathbb{R}^m$. The sparse GP posterior decomposes as $q_S(\mathbf{Z}, \mathbf{F}_m|\cdot) = p_\theta(\mathbf{Z}|\mathbf{F}_m)p_S(\mathbf{F}_m)$, where $p_S(\mathbf{F}_m) := \prod_{l=1}^L \mathcal{N}(\mathbf{f}_m^l | \boldsymbol{\mu}^l, \mathbf{A}^l)$ is a free variational distribution and $p_\theta(\mathbf{Z}|\mathbf{F}_m)$ is a (standard) conditional GP prior. The problem of minimizing the KL divergence is then equivalently posed as a maximization of a lower bound of the model evidence as follows, where in the first steps we introduce $\tilde{q}_\phi(\mathbf{Z}|\mathbf{Y})$ and $q_S(\mathbf{Z}, \mathbf{F}_m|\cdot)$ and apply Jensen's inequality:

$$\begin{aligned} \log p(\mathbf{Y}|\mathbf{X}) &= \log \int p_{\psi, \theta}(\mathbf{Z}, \mathbf{F}_m, \mathbf{Y}|\mathbf{X}) \frac{q_S(\mathbf{Z}, \mathbf{F}_m|\cdot)}{q_S(\mathbf{Z}, \mathbf{F}_m|\cdot)} \frac{\tilde{q}_\phi(\mathbf{Z}|\mathbf{Y})}{\tilde{q}_\phi(\mathbf{Z}|\mathbf{Y})} d\mathbf{Z} d\mathbf{F}_m \\ &\geq \int q_S(\mathbf{Z}, \mathbf{F}_m|\cdot) \log \frac{p_{\psi, \theta}(\mathbf{Z}, \mathbf{F}_m, \mathbf{Y}|\mathbf{X})}{q_S(\mathbf{Z}, \mathbf{F}_m|\cdot)} \frac{\tilde{q}_\phi(\mathbf{Z}|\mathbf{Y})}{\tilde{q}_\phi(\mathbf{Z}|\mathbf{Y})} d\mathbf{Z} d\mathbf{F}_m \\ &= \int q_S(\mathbf{Z}, \mathbf{F}_m|\cdot) \log \frac{\tilde{q}_\phi(\mathbf{Z}|\mathbf{Y}) p_\psi(\mathbf{Y}|\mathbf{Z}) p_\theta(\mathbf{Z}|\mathbf{F}_m) p_\theta(\mathbf{F}_m|\mathbf{X})}{\tilde{q}_\phi(\mathbf{Z}|\mathbf{Y}) p_\theta(\mathbf{Z}|\mathbf{F}_m) p_S(\mathbf{F}_m)} d\mathbf{Z} d\mathbf{F}_m \\ &= \sum_{l=1}^L \int q_S(\mathbf{z}^l, \mathbf{f}_m^l|\cdot) \log \frac{\tilde{q}_\phi(\mathbf{z}^l|\mathbf{Y}) p_\theta(\mathbf{f}_m^l|\mathbf{X})}{p_S(\mathbf{f}_m^l)} d\mathbf{z}^l d\mathbf{f}_m^l + \sum_{i=1}^N \int q_S(\mathbf{z}_i|\cdot) \left(\log p_\psi(\mathbf{y}_i|\mathbf{z}_i) - \log \tilde{q}_\phi(\mathbf{z}_i|\mathbf{y}_i) \right) d\mathbf{z}_i \\ &= \sum_{l=1}^L \mathcal{L}_T(\mathbf{U}, \phi, \theta, \boldsymbol{\mu}^l, \mathbf{A}^l) + \sum_{i=1}^N \mathbb{E}_{q_S} [\log p_\psi(\mathbf{y}_i|\mathbf{z}_i) - \log \tilde{q}_\phi(\mathbf{z}_i|\mathbf{y}_i)] \end{aligned}$$

Recall the symmetry of the Gaussian distribution $\tilde{q}_\phi(\mathbf{z}_i^l|\mathbf{y}_i) = \mathcal{N}(\mathbf{z}_i^l | \boldsymbol{\mu}^l(\mathbf{y}_i), \sigma^l(\mathbf{y}_i)) = \mathcal{N}(\boldsymbol{\mu}^l(\mathbf{y}_i) | \mathbf{z}_i^l, \sigma^l(\mathbf{y}_i))$. Hence, the first term of the penultimate expression is a sum over sparse Gaussian processes, one for each latent channel, and each term is precisely Equation 8 of Titsias (2009) for sparse Gaussian process regression. Therefore we write \mathcal{L}_T^l and let $\boldsymbol{\mu}^l = \boldsymbol{\mu}_T^l$ and $\mathbf{A}^l = \mathbf{A}_T^l$. For further derivation steps see Titsias (2009).

C Additional experiments

C.1 PCA initialization of GP-LVM vectors and inducing points

In this section, we describe how Principal Component Analysis (PCA) is used to initialize the GP-LVM digit representations as well as the inducing points in the rotated MNIST experiment. Note that both the GP-VAE (Casale et al., 2018) and the SVGP-VAE depend on GP-LVM vectors, with the SVGP-VAE additionally relying on inducing points.

To obtain a continuous digit representations for each digit instance, we start with the data matrix $\mathbf{X} \in \mathbb{R}^{P \times K}$ that consists of unrotated MNIST images. PCA is then performed on \mathbf{X} , yielding a matrix $\mathbf{D} \in \mathbb{R}^{P \times M}$ whose rows \mathbf{d}_i are used as initial values for the GP-LVM vectors. M represents the number of principal components kept.

For initialization of the inducing points, we sample n GP-LVM vectors from the empirical distribution based on the PCA matrix \mathbf{D} for each of the Q angles. This results in a matrix $\mathbf{U}_{init} \in \mathbb{R}^{m \times (1+M)}$ with $m = n \cdot Q$ representing the number of inducing points. The exact procedure is given in Algorithm 1. Results from the ablation study on the PCA initialization described here are presented in Table C.1.

Algorithm 1: Initialization of inducing points in the SVGP-VAE (rotated MNIST experiment)

input : PCA matrix \mathbf{D} , number of inducing points per angle n , set of angles $\{\frac{2\pi k}{Q} \mid k = 1, \dots, Q\}$

$\mathbf{U}_{init} = []$
sample $m = n \cdot Q$ points from empirical distribution of each principle component
for $i = 1, \dots, M$ **do**
| $\mathbf{U}_{init} = [\mathbf{U}_{init}, \text{sample}(\mathbf{D}[:, i], \text{nr_samples} = n)]$
end
add column with angle information
 $\mathbf{a} = [\underbrace{2\pi/Q, \dots, 2\pi/Q}_{n \times}, \dots, \underbrace{2\pi, \dots, 2\pi}_{n \times}]^T \in \mathbb{R}^m$
 $\mathbf{U}_{init} = [\mathbf{a}, \mathbf{U}_{init}]$
return \mathbf{U}_{init}

	PCA init	random init
GP-VAE Casale et al. (2018)	0.0370 ± 0.0012	0.0374 ± 0.0009
SVGP-VAE	0.0251 ± 0.0005	0.0272 ± 0.0006

Table C.1: A comparison of different initialization regimes for GP-LVM vectors and inducing points in the rotated MNIST experiment. For random initialization, a Gaussian distribution with mean 0 and standard deviation 1.5 was used.

C.2 SVGP-VAE latent space visualization

In Figure C.1, we depict two-dimensional t-SNE (Maaten and Hinton, 2008) embeddings of SVGP-VAE latent vectors ($L = 16$). Visualized here are latent vectors for training data of the *five-digit* version of the rotated MNIST dataset ($N = 20250$). As expected, the model clusters images based on the digit identity. More interestingly, SVGP-VAE also seems to order images within each digit cluster with respect to angles. For example, looking at the cluster of the digit 3 (the blue cluster in the middle of the lower plot), we observe that embeddings of rotated images are ordered continuously from 0 to 2π as we move in clockwise direction around the circular shape of the cluster.

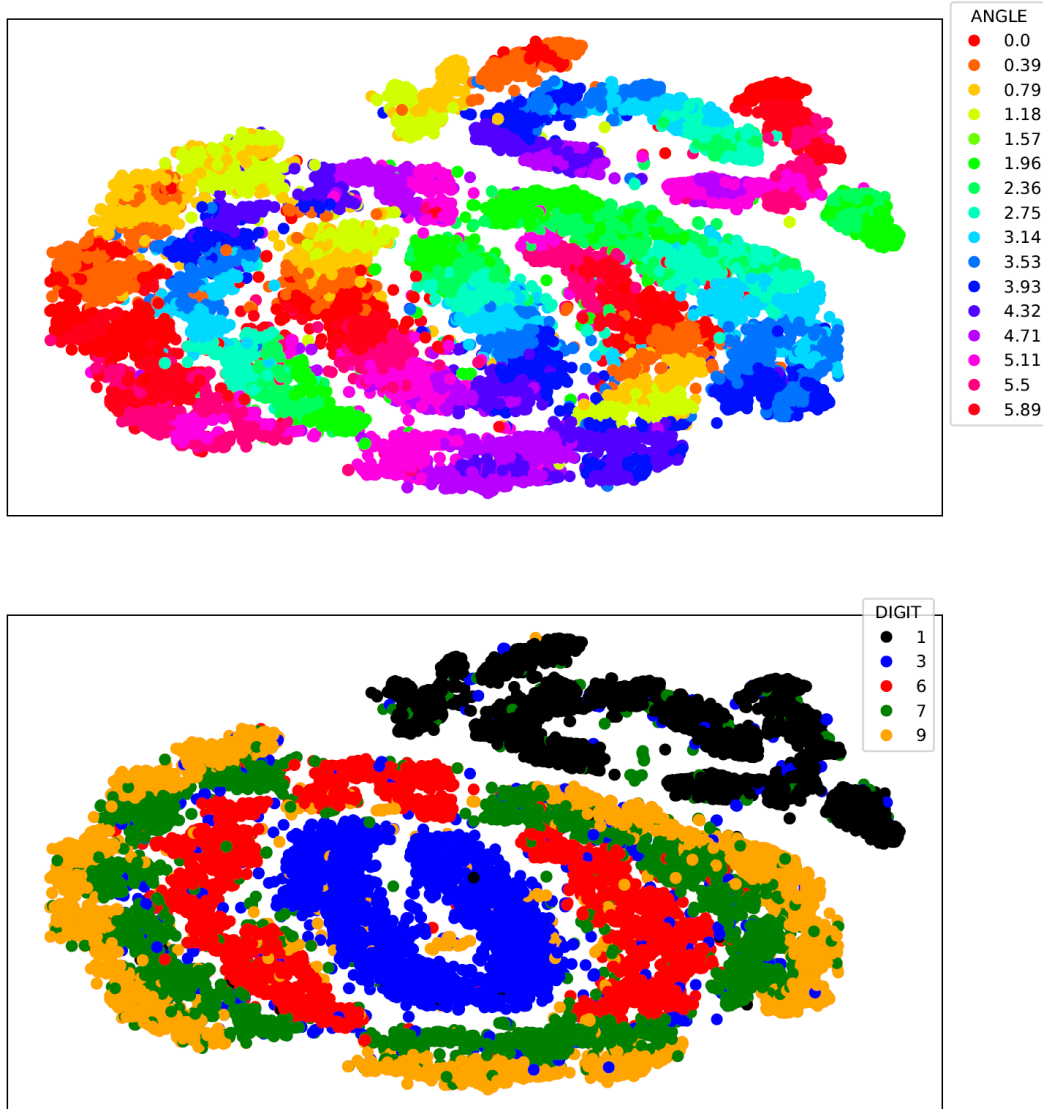


Figure C.1: t-SNE embeddings of SVGP-VAE latent vectors on the training data for rotated MNIST. On the upper scatter plot, each image embedding is colored with respect to its associated angle. On the lower scatter plot, each image embedding is colored with respect to its associated digit. The t-SNE perplexity parameter was set to 50.

C.3 Rotated MNIST: generated images

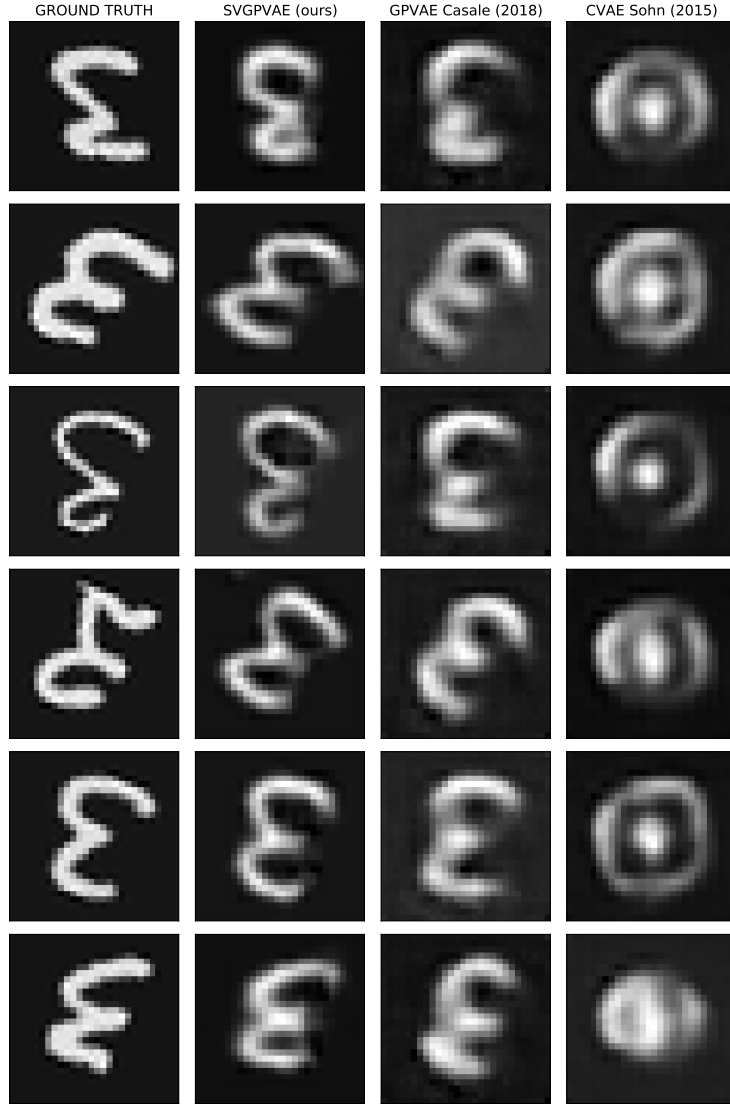


Figure C.2: Generated test images in the rotated MNIST experiment for all considered models.

C.4 Bias analysis of MC estimators in SVGP-VAE

Here we look at some additional experiments that were conducted to get a better understanding of the SVGP-VAE model. Depicted in Figure C.3 are the results when varying the batch size and the number of inducing points. We first notice that the SVGP-VAE performance improves as the batch size is increased. As pointed out in Section 3.4, this is a consequence of the Monte Carlo estimators from (9) used in q_S whose quality depends on the batch size. While the dependence on the batch size can surely be seen as one limitation of the model, it is encouraging to see that the model achieves good performance already for a reasonably small batch size (e.g., $b = 128$). Moreover, the batch size parameter in the SVGP-VAE offers a simple and intuitive way to navigate a trade-off between performance and computational demands. If one is more concerned regarding the performance, a higher batch size should be used. On the other hand, if one only has limited computational resources at disposal, a lower batch size can be utilized resulting in a faster and less memory-demanding model.

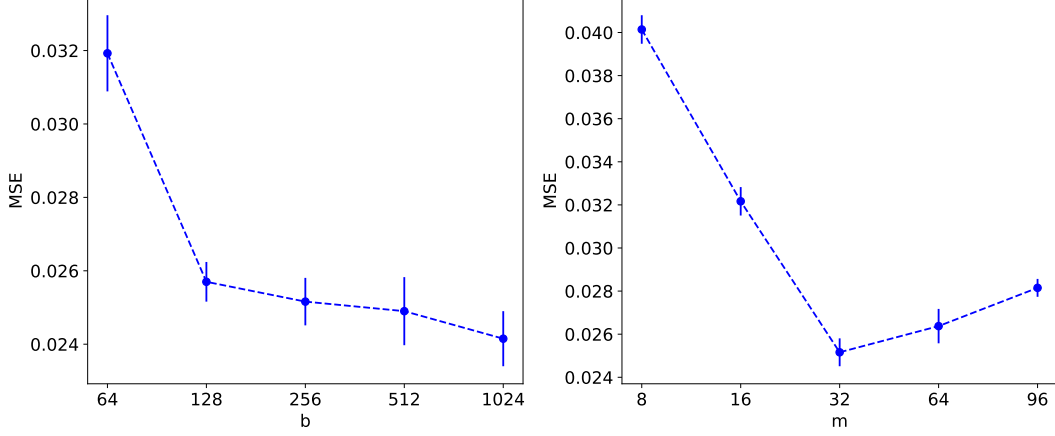


Figure C.3: SVGP-VAE results on the rotated MNIST dataset (digit 3) for varying batch size (left) and number of inducing points (right). For the batch size experiment, m was set to 32. For the inducing points experiment, b was set to 256. For each configuration, a mean MSE together with a standard deviation based on 5 runs is shown.

Looking at the plot with the varying number of inducing points next, we observe that the model achieves a solid performance with as little as 16 inducing points on the rotated MNIST data. However, increasing the number of inducing points m starts to have a negative impact on the performance after a certain point. This can be partly attributed to numerical issues that arise during training — the higher the m , the more numerically unstable the inducing point kernel matrix \mathbf{K}_{mm} becomes. Moreover, since the number of inducing points equals the dimension of the Monte Carlo estimators in (9), increasing m results in a larger dimension of the space, potentially increasing the complexity of the estimation problem.

To better understand the effect of the number of inducing points m on the quality of estimation in our proposed MC estimators, we investigate here the trajectory of the bias throughout training. To this end, for each epoch i an estimator $\mu_{j,i}^l$ is calculated for each latent channel l and for each batch j . Additionally, the true value $\mu_{T,i}^l$ is obtained (based on the entire dataset) for every epoch and every latent channel using model weights from the end of the epoch. The bias for the l -th latent channel and i -th epoch is then computed as

$$b_i^l := \frac{1}{B} \sum_{j=1}^B \mu_{j,i}^l - \mu_{T,i}^l$$

where $B := \lceil \frac{N}{b} \rceil$ represents the number of batches in a single epoch. Finally, for each epoch i the $L1$ norms of the bias vectors for each latent channel are averaged $\bar{b}_i = \frac{1}{L} \sum_{l=1}^L b_i^l$.

Moving averages of the resulting bias trajectories are depicted in Figure C.4. For comparison purposes, each trajectory is normalized by the number of inducing points used. Notice how for smaller m , the bias trajectories display the expected behavior and converge (or stay close) to 0. Conversely, for larger numbers of inducing points ($m = 64$ and $m = 96$), the bias is larger and does not decline as the training progresses. This suggests that the proposed estimation might get worse in larger dimensions.

However, despite seemingly deteriorating approximation in higher dimensions, it is also evident that the approximation does not completely break down — the model still achieves a solid performance even for a larger number of inducing points. Nevertheless, we note that getting a better theoretical grasp of the quality of estimation or reparameterizing the SVGP-VAE ELBO in a way such that these estimators are no longer needed could be a fruitful area of future work.

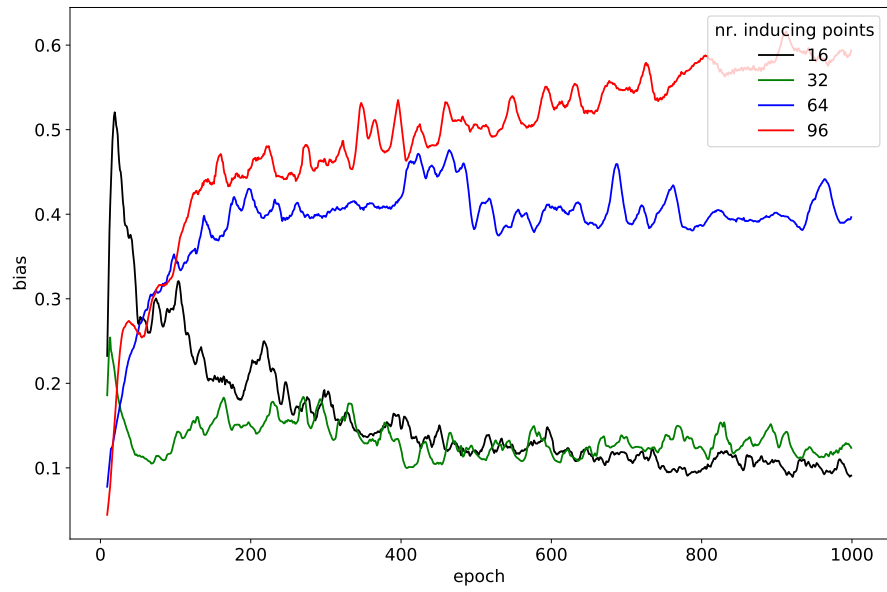


Figure C.4: Bias trajectories in the SVGP-VAE model for a varying number of inducing points. For all runs, the batch size was set to 256.