3rd Symposium on Advances in Approximate Bayesian Inference, 2020 1–11

Factorized Gaussian Process Variational Autoencoders

Anonymous Authors
Anonymous Institution

Abstract

Variational autoencoders often assume isotropic Gaussian priors and mean-field posteriors, hence do not exploit structure in scenarios where we may expect similarity or consistency across latent variables. Gaussian process variational autoencoders alleviate this problem through the use of a latent Gaussian process, but lead to a cubic inference time complexity. We propose a more scalable extension of these models by leveraging the independence of the auxiliary features, which is present in many datasets. Our model factorizes the latent kernel across these features in different dimensions, leading to a significant speed-up (in theory and practice), while empirically performing comparably to existing non-scalable approaches. Moreover, our approach allows for additional modeling of global latent information and for more general extrapolation to unseen input combinations.

1. Introduction

Variational autoencoders (VAEs) have achieved great success in many representation learning tasks (Kingma and Welling, 2013; Rezende et al., 2014). However, their isotropic Gaussian prior and variational posterior hinge on the strong assumption that all data points are independent. This can often lead to problems in real-world use cases, where the data exhibit significant correlations (Fraccaro et al., 2017; Krishnan et al., 2016).

Many alternative priors and posteriors have been proposed for VAEs (Tomczak and Welling, 2018; Fortuin et al., 2018; Kopf et al., 2019; Abiri and Ohlsson, 2020). Especially when each input comes with meta-data, or *auxiliary features*, such extra information can be used to construct Gaussian processes (GPs) and use them as priors in the latent space (Casale et al., 2018; Fortuin et al., 2020; Pearce, 2020). By choosing appropriate kernels, these resulting GP-VAE models allow to capture the structured correlations across latent variables of different data points (Rasmussen and Williams, 2006). However, they are also haunted by the computational cost of exact GP inference, which scales as $\mathcal{O}(N^3)$ for N elements in the dataset.

We consider the setting in which the inputs are images, and each image is associated with several features, some of which are unique to the image and some of which are shared with other images. For example in a set of MNIST digits rotated by multiple angles (Casale et al., 2018), the digit ID is shared by other images of the same digit while the angle is unique. Similarly, in a set of faces viewed from multiple perspectives, the person is common to multiple images while each image has a unique position (Casale et al., 2018). Further applications include a set of scenes viewed from multiple positions (Eslami et al., 2018), high dimensional spatio-temporal datasets (Ashman et al., 2020), or speech segments where one speaker's voice is shared by multiple segments (Li and Mandt, 2018).

For such settings, we propose a novel *factorized* GP-VAE model, FGP-VAE, with two desirable properties. Firstly, by carefully exploiting factorization, inference is significantly

© A. Authors.

reduced from $\mathcal{O}(N^3)$. Secondly, the representations are encouraged to be disentangled, which is a highly desirable property for VAE models (Locatello et al., 2019; van Steenkiste et al., 2019; Träuble et al., 2020).

We describe the problem setting in Section 2 and the proposed model in Section 3. We present experimental results in Section 4, and conclude in Section 5.

2. Problem Setting

Consider high-dimensional data of N elements $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_N]^\top$ where $\mathbf{y}_i \in \mathbb{R}^K$ and each data point has corresponding low-dimensional auxiliary data $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]^\top \in \mathcal{X}^N, \mathcal{X} \subseteq \mathbb{R}^D$. For ease of exposition, we will focus on the example of the rotated MNIST dataset (Casale et al., 2018). It consists of P digits, each observed at Q different angles, amounting to a total of $N = P \cdot Q$ images.[1] Each $\mathbf{x}_i = (d_i,\ w_i) \in \mathcal{X}$ is composed of a categorical digit instance d_i (integer index or one-hot encoding) and a continuous angle w_i. We wish to train a model that can (1) given new $\mathbf{x}_* \in \mathcal{X}$ generate $\mathbf{y}_* \in \mathbb{R}^K$, and (2) infer an interpretable and disentangled latent representation.

3. Method

Generative model: we follow a latent GP approach, first proposed in Casale et al. (2018) and later extended in Pearce (2020). As in a standard VAE, each image $\mathbf{y}_i$ is associated with a latent variable $\mathbf{z}_i \in \mathbb{R}^L$. Making use of the auxiliary data, we further expect that two images $\mathbf{y}_i$ and $\mathbf{y}_j$ with similar $\mathbf{x}_i$ and $\mathbf{x}_j$ should also have similar $\mathbf{z}_i$ and $\mathbf{z}_j$. To this end, a Gaussian process regression is used to model a joint distribution over all latent variables $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_N]^T \in \mathbb{R}^{N \times L}$. Given L latent dimensions, we assume L independent latent functions $f^l \sim GP(0, k^l_\theta), l = 1, \dots, L$ with kernels k^l_θ and therefore the latent variable for $\mathbf{y}_i$ may be written as $\mathbf{z}_i = [f^1(\mathbf{x}_i), \dots, f^L(\mathbf{x}_i)]^T$. Likewise, all latent variables of the lth channel $\mathbf{z}^l_{1:N} = [f^l(\mathbf{x}_1), \dots, f^l(\mathbf{x}_N)] \in \mathbb{R}^N$ are assumed to come from a single (unknown) function, specifically $\mathbf{z}^l_{1:N}$ has a correlated Gaussian prior with covariance $\mathbf{K}^l_{NN} = k^l_\theta(\mathbf{X}, \mathbf{X}) \in \mathbb{R}^{N \times N}$. The generative model, $p_\psi(\mathbf{Y}, \mathbf{Z}|\mathbf{X}) = p_\psi(\mathbf{Y}|\mathbf{Z}) p_\theta(\mathbf{Z}|\mathbf{X})$, is thus

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

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

where $\mu_\psi : \mathbb{R}^L \to \mathbb{R}^K$ is a (generative) network with parameters ψ. Note that if $\mathbf{K}_{NN} = \mathbf{I}$ is the identity matrix, the model recovers a standard VAE; using the $\mathbf{X}$ values enables the use of more sophisticated prior.

In prior work (Casale et al., 2018; Jazbec et al., 2020), a single GP prior is used in all L latent channels. Specifically, for rotated MNIST, a product kernel between a periodic and

1. The assumption that all digit instances are observed in the same number of angles Q is made to simplify notation; the presented approach does not rely on this assumption.

a linear kernel is considered

$$k_\theta(\mathbf{x}_i, \mathbf{x}_j) = \Sigma_{d_i,d_j} \cdot \exp\left(-\frac{2\sin^2\left(|w_i - w_j|\right)}{r^2} \right),$$

with parameters $\theta = \{\sigma, r, \Sigma\}$. $\Sigma = DD^T$ has a low-rank form, and $D \in \mathbb{R}^{P\times m}$ is a (learned) matrix that captures information common to all images of each digit such as written style. This approach has multiple drawbacks. Firstly, given a set of N images, the above kernel gives rise to a dense matrix $\mathbf{K}_{NN}$, where all latent variables are correlated with each other. This necessitates either $\mathcal{O}(N^3)$ cost or non-trivial approximations, such as sparse GPs (Jazbec et al., 2020) or assuming equally spaced x_i enabling specialized matrix decompositions (Casale et al., 2018). Statistically, the prior does not factorize across any subsets of the data. Secondly, if new digits are added to the dataset, the Σ matrix needs be augmented with a new row and column, the new hyperparamters of D must be learned from scratch, and it is not "amortized" over digits.

In this work, we propose two simple (yet still unexplored) changes that alleviate the aforementioned issues. We start by partitioning the dataset into digit specific subsets $\{\mathbf{X}, \mathbf{Y}, \mathbf{Z}\} = \cup_{p=1}^{P}\{\mathbf{X}_p, \mathbf{Y}_p, \mathbf{Z}_p\}$ and denote partitions by

$$\begin{aligned}
\mathbf{X}_p &= \{\mathbf{x}_i | d_i = p\} \in \mathbb{R}^{Q\times D}, \\
\mathbf{Y}_p &= \{\mathbf{y}_i | d_i = p\} \in \mathbb{R}^{Q\times K}, \\
\mathbf{Z}_p &= \{\mathbf{z}_i | d_i = p\} \in \mathbb{R}^{Q\times L}.
\end{aligned}$$

Instead of assuming a single correlated prior over all N latent variables, we assume P separate correlated priors over Q latent variables each, that is, one prior for each partition of the data. Secondly, we assume that each latent variable $\mathbf{z}_i \in \mathbb{R}^L$ is composed of two parts. Temporarily dropping i for clarity, we propose to have

$$\mathbf{z} = \big(\underbrace{z^1, ..., z^J}_{local}, \underbrace{z^{J+1}, ..., z^L}_{global} \big),$$

where the *local* variable $\mathbf{z}_i^{1:J}$ is unique to the given image and the *global* variable $\mathbf{z}_i^{J+1:L}$ is shared by all elements in the subset $\mathbf{Z}_p$. The global variable captures style information or other common features of digit d_i that are angle-agnostic. For example, for frames from videos of moving objects, local variables could capture object position and location while global variables could capture object color and shape.

We must construct a kernel for each latent dimension $l \in \{1, ..., L\}$ satisfying the above criteria. For the local latent channels $l \in \{1, ..., J\}$, we specify the following kernel[2]

$$k_\theta^1(\mathbf{x}_i, \mathbf{x}_j) = \delta_{d_i,d_j} \cdot \sigma^2 \exp\left(-\frac{2\sin^2\left(|w_i - w_j|\right)}{r^2} \right).$$

The first Kronecker delta term ensures that two images corresponding to different digits have zero assumed latent variable similarity, while for two images of the same digit the

2. $\delta_{d_i,d_j} = 1$ if $d_i = d_j$ and 0 else.

assumed similarity depends upon the difference in the rotation angles. This enforces that the GP priors for each subset of digits are independent.

For the global latent channels $l \in \{J+1, ..., L\}$, we wish to capture rotation-agnostic characteristics of each image, that is, its style. To this end, a simple binary kernel is used

$$k^2_\theta(\mathbf{x}_i, \mathbf{x}_j) = \delta_{d_i, d_j}.$$

Thus, among rotated images of the same digit, the global latents have perfect correlation; for each global channel there is a single univariate distribution shared by all images of a single digit. Simultaneously, the global latent variable of images in another subset is treated as independent (this may also be viewed as the local kernel with length scale $r \to \infty$). Due to the kernel structure, there is now a separate generative model for every digit instance $p_{\psi,\theta}(\mathbf{Y}, \mathbf{Z}|\mathbf{X}) = \prod_{p=1}^{P} p_\psi(\mathbf{Y}_p|\mathbf{Z}_p)p_\theta(\mathbf{Z}_p|\mathbf{X}_p)$. Within each of the digit-specific generative models, working with the GP prior is much less prohibitive as $Q \ll N$. Secondly, digit style is captured in the global latent variables $\mathbf{z}_i^{j+1:L}$ which can be estimated from images via amortization. Such global information is no longer encoded in the generative model hyperparameters Σ.

Approximate Posterior: exploiting the factorized structure of the generative model, we may consider the posterior of each subset $\mathbf{Z}_p$ independently. Since the true posterior for latent variables $p_{\psi,\theta}(\mathbf{Z}_p|\mathbf{Y}_p, \mathbf{X}_p)$ is intractable, approximate inference is required. In VAEs, an inference network (with parameters ϕ) takes $\mathbf{y}_q \in \mathbf{Y}_p$ as input to predict the mean and variance of a mean-field approximate posterior of each latent encoding which we denote as

$$\tilde{q}_\phi(\mathbf{z}_q|\mathbf{y}_q) = \prod_{l=1}^{L} \mathcal{N}\left(z_q^l | \mu_\phi^l(\mathbf{y}_q), \sigma_\phi^l(\mathbf{y}_q)^2\right) ,$$

and one possible approximate posterior is to use the product of the above factors over all N latent variables. Instead, closely following Pearce (2020), we use $\tilde{q}_\phi(\cdot)$ to replace only the intractable likelihood $p_\psi(\mathbf{y}_i|\mathbf{z}_i)$ in the exact posterior. This gives rise to the following approximate posterior

$$q(\mathbf{Z}_p|\mathbf{Y}_p, \mathbf{X}_p, \phi, \theta) := \frac{\prod_{q=1}^{Q} \tilde{q}_\phi(\mathbf{z}_q|\mathbf{y}_q) \cdot p_\theta(\mathbf{Z}_p|\mathbf{X}_p)}{Z_{\phi,\theta}(\mathbf{Y}_p, \mathbf{X}_p)}.$$

The conjugacy of the Gaussian prior and (approximate) Gaussian likelihoods yields a closed-form solution for the normalizing constant $Z_{\phi,\theta}(\mathbf{Y}_p, \mathbf{X}_p)$. Moreover, the approximate posterior $q(\mathbf{Z}_p|\cdot)$ is mathematically equivalent to a product of J exact GP posteriors of Q points (one GP for each angle latent channel) and $L - J$ univariate Gaussian distributions that are common to all Q elements in the subset. Exact derivations are given in Appendix B.

Finally, the FGP-VAE ELBO has the form:

$$\log p(\mathbf{Y}|\mathbf{X}) \geq \sum_{p=1}^{P} \mathbb{E}_q \left[\sum_{q=1}^{Q} \log p_\psi(\mathbf{y}_q|\mathbf{z}_q) - \log \tilde{q}_\phi(\mathbf{z}_q|\mathbf{y}_q) \right] + \log Z_{\phi,\theta}(\mathbf{Y}_p, \mathbf{X}_p).$$

Due to its factorization across digit subsets, and the assumption that $Q \ll N$, exact GP inference (with approximate likelihoods) is feasible, resulting in $\mathcal{O}(PQ^3)$ complexity for one

Factorized Gaussian Process Variational Autoencoders

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. P represents the number of unique digits, Q the number of rotations for each digit and m the dimension of the low-rank matrix in the GP kernel used in Casale et al. (2018).

	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	$\mathcal{O}(PQ^3m^2)$	19.10 ± 0.66
FGP-VAE (ours)	0.0284 ± 0.0004	$\mathcal{O}(PQ^3)$	1.41 ± 0.08

epoch. Additionally, training can be done in mini-batches of digit subsets $\{\mathbf{Y}_p, \mathbf{X}_p\}$, hence the ELBO does not require the whole dataset in memory. In cases where Q is large or $\mathbf{Y}_p$ do not fit into memory, the factorized kernel we propose may be combined with a sparse GP-VAE method (Ashman et al., 2020; Jazbec et al., 2020) to further reduce computational complexity.

4. Experiments

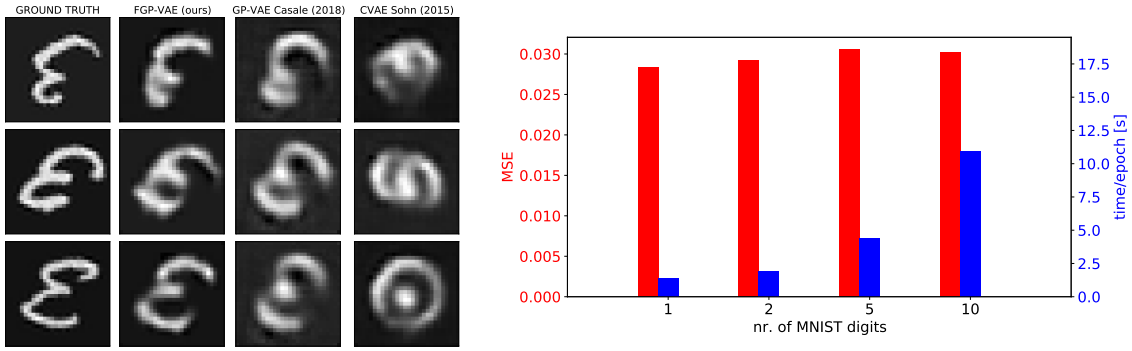
We follow the experimental setup from Casale et al. (2018) in conditionally generating rotated images of MNIST handwritten digits (LeCun et al., 1998). The dataset consists of $P = 400$ different instances of the digit 3 at $Q = 16$ different angles each, resulting in a total of $N = 6400$ possible combinations. From these combinations, $N_{train} = 4050$ images are used for training and $N_{test} = 270$ for testing. We choose the same network architecture as the one in Casale et al. (2018) for all models (see Appendix A for implementation details). For running the baselines, we used the code from Jazbec et al. (2020). Moreover, we make use of the GECO algorithm (Rezende and Viola, 2018) to train our FGP-VAE model, as it improves training stability. Our code is made available at `ANONYMOUS LINK`.

Reconstruction performance. We see qualitatively in Figure 1(a) and quanitatively in Table 1 that our proposed FGP-VAE clearly outperforms the non-correlated CVAE model (Sohn et al., 2015) and performs comparably to the non-factorized GP-VAE (Casale et al., 2018). However, our proposed model is an order of magnitude faster than the non-factorized GP-VAE and reaches runtimes that are almost as fast as the CVAE.

Scaling behavior. We additionally studied the scaling of our proposed model to differently sized subsets, including the full dataset. (We do not compare against the other GP-VAE model here, since we did not manage to scale it to this larger dataset). We can see in Figure 1(b) that our proposed model does not deteriorate in performance when scaling the dataset, while the runtime scales as gracefully as theoretically predicted (see Sec. 3).

Extrapolation in the digit space. An additional attractive feature of the factorized model is that it can (unlike past works) extrapolate beyond digits observed in the training

5



(a) Conditional generations

(b) Scaling properties

Figure 1: (a) Conditionally generated rotated MNIST images. The generations of our proposed model are qualitatively more faithful to the ground truth. (b) 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.

data. Past works only considered images seen during training and generated them at new angles. The FGP-VAE can also generate arbitrary rotations for previously unseen digits given either a single example (sample the posterior of global latents), or even generated randomly (sample the prior of global latents). This is a consequence of our disentangled (and arguably much simpler) GP kernel. The FGP-VAE achieved an MSE of 0.0316 ± 0.0005 for the extrapolation experiment, which is only slightly worse than the 0.0284 ± 0.0004 for the non-extrapolation version of the experiment. This demonstrates the FGP-VAE's strong extrapolation ability in the digit space. Generated images from the extrapolation experiment are shown in Appendix D. We further elaborate on the extrapolation properties of GP-VAE models in Appendix C.

5. Conclusion

We have proposed a novel method for improving the scalability of GP-VAE models in settings where the auxiliary data consists of several independent features. Our method factorizes the latent GP kernel across the different data features in different latent dimensions, leading to a large reduction in inference time complexity. We have shown that our model is faster than existing non-factorized approaches in practice, while yielding a comparable predictive performance and offering more general extrapolation properties. In future work, it would be interesting to study the combination of our method with the recently proposed sparse GP-VAE approaches (Jazbec et al., 2020; Ashman et al., 2020), to reduce the inference time even further.

References

Najmeh Abiri and Mattias Ohlsson. Variational auto-encoders with Student's t-prior. 4 2020. URL http://arxiv.org/abs/2004.02581.

Matthew Ashman, Jonathan So, Will Tebbutt, Vincent Fortuin, Michael Pearce, and Richard E. Turner. Sparse gaussian process variational autoencoders. *arXiv preprint arXiv:2010.10177*, 2020.

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

S. M. Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S. Morcos, Marta Garnelo, Avraham Ruderman, Andrei A. Rusu, Ivo Danihelka, Karol Gregor, David P. Reichert, Lars Buesing, Theophane Weber, Oriol Vinyals, Dan Rosenbaum, Neil Rabinowitz, Helen King, Chloe Hillier, Matt Botvinick, Daan Wierstra, Koray Kavukcuoglu, and Demis Hassabis. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018. ISSN 10959203. doi: 10.1126/science.aar6170.

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

Vincent Fortuin, Dmitry Baranchuk, Gunnar Rätsch, and Stephan Mandt. Gp-vae: Deep probabilistic time series imputation. In *International Conference on Artificial Intelligence and Statistics*, pages 1651–1661. PMLR, 2020.

Marco Fraccaro, Simon Kamronn, Ulrich Paquet, and Ole Winther. A disentangled recognition and nonlinear dynamics model for unsupervised learning. In *Advances in Neural Information Processing Systems*, pages 3601–3610, 2017.

Metod Jazbec, Vincent Fortuin, Michael Pearce, Stephan Mandt, and Gunnar Rätsch. Scalable gaussian process variational autoencoders. *arXiv preprint arXiv:2010.13472*, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Andreas Kopf, Vincent Fortuin, Vignesh Ram Somnath, and Manfred Claassen. Mixture-of-experts variational autoencoder for clustering and generating from similarity-based representations. *arXiv preprint arXiv:1910.07763*, 2019.

Rahul G Krishnan, Uri Shalit, and David Sontag. Structured inference networks for nonlinear state space models. *arXiv preprint arXiv:1609.09869*, 2016.

Neil 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.

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

Yingzhen Li and Stephan Mandt. Disentangled sequential autoencoder. *arXiv preprint arXiv:1803.02991*, 2018.

Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. In *Advances in Neural Information Processing Systems*, pages 14611–14624, 2019.

Michael Pearce. The gaussian process prior vae for interpretable latent dynamics from pixels. In *Symposium on Advances in Approximate Bayesian Inference*, pages 1–12, 2020.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. 2006. ISBN 026218253X.

Danilo Jimenez Rezende and Fabio Viola. Taming vaes. *arXiv preprint arXiv:1810.00597*, 2018.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pages 3483–3491, 2015.

Jakub M. Tomczak and Max Welling. VAE with a vampprior. *International Conference on Artificial Intelligence and Statistics, AISTATS 2018*, pages 1214–1223, 2018.

Frederik Träuble, Elliot Creager, Niki Kilbertus, Anirudh Goyal, Francesco Locatello, Bernhard Schölkopf, and Stefan Bauer. Is independence all you need? on the generalization of representations learned from correlated data. *arXiv preprint arXiv:2006.07886*, 2020.

Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are disentangled representations helpful for abstract visual reasoning? In *Advances in Neural Information Processing Systems*, pages 14245–14258, 2019.

Appendix A. Implementation details

For the rotated MNIST experiment described, 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. For more details, see Table A.1.

Factorized Gaussian Process Variational Autoencoders

Table A.1: 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
Number of latent channels for angle info (J)	8

The FGP-VAE model is trained for 1000 epochs with a batch size of 220 images (20 digits subsets, each with 11 rotations). 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) is used for training our FGP-VAE model in this experiment. The reconstruction parameter in GECO was set to $\kappa = 0.020$ in all reported experiments.

GP parameters are kept fixed throughout training for FGP-VAE. The amplitude is set to $\sigma = 1$ and the length scale to $r = 1$. For the baseline GP-VAE model (Casale et al., 2018), GP parameters are optimized during training as proposed in Casale et al. (2018).

Appendix B. Derivations

To simplify the notation, we use here one latent channel per feature set, i.e. $J = 1$ and $L = 2$. For $\mathbf{z}_q \in \mathbf{Z}_p$, we thus have that $\mathbf{z}_q = [z_q \;\; z_p]^T$, where we drop latent channel superscripts for clarity. With z_q we denote a local latent variable that is specific to the q-th rotation, while z_p represents a global latent variable that is shared among all rotations of the p-th digit. Further, let $\mathbf{z}_p^1 = [z_1 \dots z_Q]^T \in \mathbb{R}^Q$ contain all local latent variables in $\mathbf{Z}_p$, and $\mathbf{K}_p = k_\theta^1(\mathbf{X}_p, \mathbf{X}_p)$. For notational convenience, let $\mu_{q,l} := \mu_\phi^l(\mathbf{y}_q)$ and $\sigma_{q,l} := \sigma_\phi^l(\mathbf{y}_q)$. We proceed as

$$
\begin{aligned}
Z_{\phi,\theta}(\mathbf{Y}_p, \mathbf{X}_p) = \int \prod_{q=1}^{Q} \tilde{q}_\phi(\mathbf{z}_q|\mathbf{y}_q) \cdot p_\theta(\mathbf{Z}_p|\mathbf{X}_p)\, d\mathbf{Z}_p = \\
\int \prod_{q=1}^{Q} \mathcal{N}(z_q|\mu_{q,1},\, \sigma_{q,1}^2)\, \mathcal{N}(z_p|\mu_{q,2},\, \sigma_{q,2}^2) \cdot \mathcal{N}(\mathbf{z}_p^1|\mathbf{0}, \mathbf{K}_p)\, \mathcal{N}(z_p|0,1)\, d\mathbf{Z}_p = \\
\int \prod_{q=1}^{Q} \mathcal{N}(z_q|\mu_{q,1},\, \sigma_{q,1}^2)\, \mathcal{N}(\mathbf{z}_p^1|\mathbf{0}, \mathbf{K}_p)\, d\mathbf{z}_p^1 \cdot \int \prod_{q=1}^{Q} \mathcal{N}(z_p|\mu_{q,2},\, \sigma_{q,2}^2)\, \mathcal{N}(z_p|0,1)\, dz_p\,.
\end{aligned}
$$

By exploiting the symmetry of Gaussian distribution, $\mathcal{N}(z|\mu,\sigma) = \mathcal{N}(\mu|z,\sigma)$, the first integral equals a marginal GP likelihood in the standard GP regression with inputs $\mathbf{X}_p$ and

9

outputs $\boldsymbol{\mu}_p := [\mu_{1,1} \dots \mu_{Q,1}]^T \in \mathbb{R}^Q$ with (heteroscedastic) noise $\boldsymbol{\sigma}_p := [\sigma_{1,1} \dots \sigma_{Q,1}]^T \in \mathbb{R}^Q$. Combining the same symmetry property with a formula for conjugate posterior parameters for Gaussian likelihood with known heteroscedastic variance yields the following expression for the second integral

$$\int \prod_{q=1}^{Q} \mathcal{N}(z_p|\mu_{q,2},\ \sigma_{q,2}^2)\, \mathcal{N}(z_p|0,1)\, dz_p = \frac{\mathcal{N}(0|0,1) \prod_{q=1}^{Q} \mathcal{N}(0|\mu_{q,2},\ \sigma_{q,2}^2)}{\mathcal{N}(0|\bar{\mu}_2, \bar{\sigma}_2^2)}\ ,$$

where

$$\bar{\sigma}_2^2 = \big(1 + \sum_{q=1}^{Q} \frac{1}{\sigma_{q,2}^2}\big)^{-1}\ , \quad \bar{\mu}_2 = \bar{\sigma}_2^2 \sum_{q=1}^{Q} \frac{\mu_{q,2}}{\sigma_{q,2}^2}\ .$$

Similarly, a closed form for the approximate posterior can be obtained as

$$q(\mathbf{Z}_p|\mathbf{Y}_p, \mathbf{X}_p, \phi, \theta) = \frac{\prod_{q=1}^{Q} \tilde{q}_\phi(\mathbf{z}_q|\mathbf{y}_q) \cdot p_\theta(\mathbf{Z}_p|\mathbf{X}_p)}{Z_{\phi,\theta}(\mathbf{Y}_p, \mathbf{X}_p)} =$$

$$\underbrace{\frac{\prod_{q=1}^{Q} \mathcal{N}(z_q|\mu_{q,1},\ \sigma_{q,1}^2)\, \mathcal{N}(\mathbf{z}_p^1|\mathbf{0}, \mathbf{K}_p)}{\int \prod_{q=1}^{Q} \mathcal{N}(z_q|\mu_{q,1},\ \sigma_{q,1}^2)\, \mathcal{N}(\mathbf{z}_p^1|\mathbf{0}, \mathbf{K}_p)\, d\mathbf{z}_p^1}}_{\text{(exact) GP posterior for } \{\mathbf{X}_p, \boldsymbol{\mu}_p, \boldsymbol{\sigma}_p\}} \cdot \underbrace{\frac{\prod_{q=1}^{Q} \mathcal{N}(z_p|\mu_{q,2},\ \sigma_{q,2}^2)\, \mathcal{N}(z_p|0,1)}{\int \prod_{q=1}^{Q} \mathcal{N}(z_p|\mu_{q,2},\ \sigma_{q,2}^2)\, \mathcal{N}(z_p|0,1)\, dz_p}}_{= \mathcal{N}(z_p|\bar{\mu}_2,\ \bar{\sigma}_2^2),\ \text{Gaussian posterior}}\ .$$

Finally, an FGP-VAE ELBO can be derived as follows using the standard steps:

$$\log p(\mathbf{Y}|\mathbf{X}) \geq \sum_{p=1}^{P} \int \log \frac{p_{\psi,\theta}(\mathbf{Y}_p, \mathbf{Z}_p|\mathbf{X}_p)}{q(\mathbf{Z}_p|\cdot)} q(\mathbf{Z}_p|\cdot) d\mathbf{Z}_p =$$

$$\sum_{p=1}^{P} \int \log \left(\frac{p_\psi(\mathbf{Y}_p|\mathbf{Z}_p) \cdot p_\theta(\mathbf{Z}_p|\mathbf{X}_p) \cdot Z_{\phi,\theta}(\mathbf{Y}_p, \mathbf{X}_p)}{\prod_{q=1}^{Q} \tilde{q}_\phi(\mathbf{z}_q|\mathbf{y}_q) \cdot p_\theta(\mathbf{Z}_p|\mathbf{X}_p)} \right) q(\mathbf{Z}_p|\cdot) d\mathbf{Z}_p =$$

$$\sum_{p=1}^{P} \mathbb{E}_q \left[\sum_{q=1}^{Q} \log p_\psi(\mathbf{y}_q|\mathbf{z}_q) - \log \tilde{q}_\phi(\mathbf{z}_q|\mathbf{y}_q) \right] + \log Z_{\phi,\theta}(\mathbf{Y}_p, \mathbf{X}_p).$$

Appendix C. Amortization of auxiliary data in GP-VAE models

Auxiliary data $\mathbf{X}$ is crucial in applications of GP-VAE models as it represents the data over which a GP prior is placed.[3] While it is often fully observed, there are cases where auxiliary data is not given (or is only partially observed). In such instances, the authors in Casale et al. (2018) rely on the GP-LVM (Lawrence, 2004) to learn the missing parts of the auxiliary information. Such an approach solves the issue of (partly) unobserved $\mathbf{X}$ in an elegant way, however by doing so, the extrapolation ability of GP-VAE models is diminished. Suppose we want to generate new views or angles for previously unseen digits or objects. In that case, we need to re-run the training optimization so that the respective GP-LVM

3. Auxiliary data in a GP-VAE corresponds to independent variables in a GP regression.

vectors are obtained. Note that GP-LVM vectors correspond to rows in the low-rank matrix $\Sigma \in \mathbb{R}^{P \times m}$ that is part of the GP kernel proposed in Casale et al. (2018).

Another way of endowing GP-VAE models with the extrapolation ability, besides considering factorized (and simpler) GP priors as done in our FGP-VAE, would be to amortize the GP-LVM information using a *representation network* $r_\zeta : \mathbb{R}^K \to \mathbb{R}^m$, similar to what is done in Eslami et al. (2018). The representation for the p-th digit instance is then

$$\mathbf{d}_p = f\big(r_\zeta(\mathbf{y}_1),\, \ldots,\, r_\zeta(\mathbf{y}_Q)\big) \in \mathbb{R}^m \,,$$

where $\mathbf{Y}_p = [\mathbf{y}_1 \ldots \mathbf{y}_Q]^T$, and f is a chosen aggregation function, for instance, a sum or a mean. Instead of GP-LVM vectors, the parameters of the representation network ζ would be learned jointly with the rest of GP-VAE parameters.

Appendix D. Extrapolation in the digit space

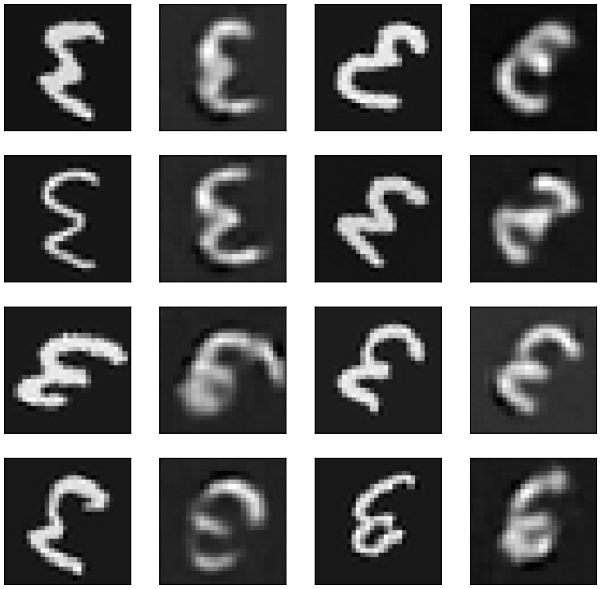


Figure D.1: Ground truths (columns 1 and 3) and generated images (columns 2 and 4) using FGP-VAE for new digit instances (not seen during training in any angle). To generate the rotations for each new digit in the test phase, 11 context images were given to the model.