4th Symposium on Advances in Approximate Bayesian Inference, 2022 1–19

On Disentanglement in Gaussian Process Variational Autoencoders

Simon Bing* BINGS@ETHZ.CH
Vincent Fortuin* FORTUIN@INF.ETHZ.CH
Gunnar Rätsch RAETSCH@INF.ETHZ.CH
Department of Computer Science, ETH Zürich, Zürich, Switzerland

1. Introduction

The success of machine learning applications is greatly influenced by the representation of the data. Bengio et al. (2013) state that if a representation caters to a certain learning task, better results and increased robustness may be expected. Data representations used in recent successful deep learning approaches for certain tasks (Silver et al., 2016; He et al., 2016; Mnih et al., 2015) overfit to the task at hand (Burgess et al., 2017), which partially explains why they still fall behind biological intelligence in terms of generality and knowledge transfer (Lake et al., 2017). Recent research suggests that disentangled representations could provide a solution (Bengio et al., 2013; Ridgeway, 2016; Tschannen et al., 2018).

While there is no single, widely accepted definition of disentanglement, the intuition of what constitutes such a representation is shared. Disentangled representations should separate the independent factors of variation that led to the generation of said data (Bengio et al., 2013). Such representations have the property that each latent factor is only sensitive to a change in a single underlying factor of variation. Disentangled representations have been argued to offer benefits in terms of interpretability (Adel et al., 2018; Bengio et al., 2013), predictive performance (Locatello et al., 2019b), fairness (Locatello et al., 2019a), and reducing the sample complexity for downstream tasks (van Steenkiste et al., 2019).

Sequential data appear in a wide variety of settings such as audio and video streams, communication signal processing, or longitudinal medical data, motivating the investigation of learning disentangled representations from such data. While there has been previous work on disentangling static from dynamic factors of sequential data (Li and Mandt, 2018; Hsu et al., 2017), none have attempted to disentangle the dynamic factors themselves. Previously introduced models belonging to the class of Gaussian process variational autoencoders (GP-VAEs) have very successfully leveraged the temporal correlations of such sequential data to tackle problems such as conditional generation (Casale et al., 2018) or missing value imputation (Fortuin et al., 2020), but have not explicitly looked at disentangling such data.

Recent work in the disentanglement literature has shown that learning disentangled representations in a fully unsupervised fashion is fundamentally impossible (Locatello et al., 2019b). Locatello et al. (2020a) show that inductive biases must be included to achieve this task and further provide the first model that makes explicit assumptions on the structure of the input data to improve the disentanglement of the learned representation. Inspired by the

* Equal contribution.

notion of weak supervision in this model as well as the the successful application of GP-VAE models to tasks involving sequential data, we investigate the disentanglement properties of GP-VAE type models. To this end, we study the Disentangled Gaussian Process Variational Autoencoder (DGP-VAE), an adaptation of Fortuin et al. (2020)'s GP-VAE model, that exploits the sequential structure of time series data to learn disentangled representations.

We make the following contributions:

- We study the disentanglement properties of the recently proposed GP-VAE model, by introducing a modification, the DGP-VAE, where latent GP priors with variable length scales are used to encourage disentanglement between latent dimensions.
- We demonstrate that these VAE models, by the pure virtue of their GP prior, already achieve disentanglement "for free", in contrast to conventional disentanglement models, which heavily rely on engineering tricks and specialized design choices.
- We compare against state-of-the-art disentanglement models and show that we outperform all considered baselines in terms of disentanglement on standard benchmark data sets.
- We perform a study on real-world medical time series data and demonstrate that our modeling assumptions are better suited to learning disentangled representations of real time series data compared to those of previously introduced weakly-supervised models.

2. Disentangled Gaussian Process VAE

We introduce the Disentangled Gaussian Process Variational Autoencoder (DGP-VAE), a GP-VAE-based model (Fortuin et al., 2020) for learning disentangled representations from time series data. The main idea of our model is to exploit the correlations in sequential data sets by the application of latent Gaussian process priors. We make certain smoothness assumptions about the input data and explicitly exploit this inductive bias for the benefit of disentanglement. We argue that while our assumptions on the dynamics of the sequential data are weaker than previous approaches, they are better aligned with real-world data, which we show on a data set consisting of real-world medical time series.

2.1. Generative model

Previous work has shown that learning fully disentangled representations is fundamentally impossible without inductive biases and it is beneficial to be explicit about these modeling assumptions (Locatello et al., 2019b). Following this impossibility result, there have been approaches that break with the paradigm of the fully unsupervised setting (Locatello et al., 2020b) and no longer assume the input data to be i.i.d. (Locatello et al., 2020a). While these models outperform fully unsupervised approaches in terms of disentanglement, we argue that their assumptions that some underlying generative factors must be shared from one time step to the next are too restrictive for realistically occurring time series data.

We provide a less restrictive set of assumptions on sequential data, namely that neighbouring time steps are correlated and that the changes over time are smooth. Therefore,

the time series of the latent representations is also assumed to be smooth and correlated in time. We model these assumptions in the form of a GP prior, $\mathbf{z}(\tau) \sim \mathcal{GP}(m_{\mathbf{z}}(\cdot), k_{\mathbf{z}}(\cdot,\cdot))$, for each latent channel. Gaussian processes have been previously shown to be suitable for time series modeling (Roberts et al., 2013), and we further justify this choice by their ability to model correlations of points in time while enforcing smoothness, which is congruent with the aforementioned assumptions we make.

This choice of latent prior distribution further grants our model the additional flexibility to capture a plethora of possible temporal coherence characteristics in the data, by employing different GP kernels. The Ada-GVAE model (Locatello et al., 2020a) can be seen as a special case of this architecture, albeit with a very rigid kernel that only captures pairwise correlations, as opposed to our more flexible architecture that can capture long-range dependencies. As multivariate time series often exhibit dynamics on a number of different timescales, we opt for the Cauchy kernel in practice which can be derived from an infinite mixture of radial basis function (RBF) kernels. The Cauchy kernel therefore naturally lends itself to modelling dynamics on a variety of time scales. Note that in our DGP-VAE model, in contrast to the original GP-VAE (Fortuin et al., 2020), the length scale is variable between different latent channels, in order to encourage the disentanglement of factors of variation that change with different frequencies over time.

2.2. Inference model

The inference model in our architecture yields the approximate posterior distribution $q_\psi(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})$, which is needed to infer the latent representation from the input data and to learn the parameters of the previously introduced generative model. Since inferring the exact posterior distribution is intractable, we employ a variational inference scheme (Jordan et al., 1999; Blei et al., 2017). Based on the architecture proposed by Fortuin et al. (2020), we use a structured variational distribution (Wainwright and Jordan, 2008) to capture the temporal correlation of the data, in conjunction with efficient amortized inference (Kingma and Welling, 2014).

We jointly train the generative network parameters θ and the inference network parameters ψ by optimizing the following objective:

$$\max_{\psi,\theta} \sum_{t=1}^{T} \mathbb{E}_{q_\psi(\mathbf{z}_t|\mathbf{x}_{1:T})}[\log p_\theta(\mathbf{x}_t|\mathbf{z}_t)] - \beta D_{\mathrm{KL}}(q_\psi(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})||p(\mathbf{z}_{1:T})). \tag{1}$$

In summary, our model can be seen as a modified version of the GP-VAE model from Fortuin et al. (2020) in which we have a higher fidelity of control over the independent GP priors of the different latent dimensions, given by the different length scales of the GP kernel. This encourages the disentanglement of underlying dynamic factors which vary with different rates. While this is not a major extension of the model, our goal is not to propose a radically new model, but to study the general disentanglement properties of GP-VAE models when imbued with reasonable assumptions about realistic time series data.

3. Experiments

We performed experiments on time series data synthesized from four different data sets commonly used in the disentanglement literature: dSprites (Matthey et al., 2017), SmallNORB

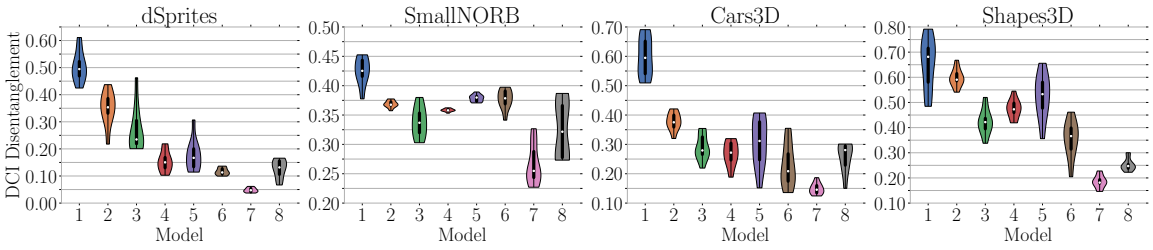

Figure 1: Results of the baseline experiments for all considered benchmark data sets. Models: 1) DGP-VAE (ours), 2) Ada-GVAE, 3) AnnealedVAE, 4) β-VAE, 5) β-TCVAE, 6) FactorVAE, 7) DIP-VAE-I, and 8) DIP-VAE-II.

(LeCun et al., 2004), Cars3D (Reed et al., 2015) and Shapes3D (Burgess and Kim, 2018). We provide an extensive comparison against state-of-the-art unsupervised approaches (Higgins et al., 2017; Burgess et al., 2017; Chen et al., 2018; Kim and Mnih, 2018; Kumar et al., 2018) and a recently proposed state-of-the-art weakly-supervised model (Locatello et al., 2020a). We present strong quantitative evidence that our model outperforms all of the competing approaches in learning disentangled representations from sequential data.

Additionally, we investigated the disentanglement properties of our model when applied to real-world medical time series data. To this end, we performed an experiment using the HiRID data set (Hyland et al., 2020) and compare against the Ada-GVAE model (Locatello et al., 2020a). This setting allows us to validate our model's applicability to real-world time series data, while demonstrating that our modeling assumptions are more closely aligned with such data than the more restrictive assumptions made by the Ada-GVAE.

Implementation details for all experiments can be found in Appendix A and additional results in Appendix B. All of our code is available online[1].

3.1. Standard benchmark data

Experimental setup We synthesize sequential data from four publicly available and commonly used data sets to investigate disentanglement. These synthetic data sets lend themselves to our experimental purposes for the practical reason that we have access to the underlying factors of variation for each observed sample. We need these underlying factors to synthesize time series of observations, as well as to qualitatively evaluate the disentanglement of the learned representations.

In contrast to Locatello et al. (2020a), we do not impose any restriction on the number of underlying factors that change from one time step to the next and in general this change can be dense, that is, all underlying factors may change. We believe that this reflects the nature of multivariate time series in the real world, such as medical time series, where the underlying causes of observed variables may exhibit dynamics on a wide range of time scales.

The considered baseline methods[2] and our model are all trained on the same data sets. The pairs required for the training of the Ada-GVAE model are taken as neighboring

1. https://github.com/ratschlab/dgp-vae
2. The baseline methods are implemented with the `disentanglement_lib` (Locatello et al., 2019b).

samples within a time series, as suggested in the original paper (Locatello et al., 2020a). Locatello et al. (2019b) show that all commonly used disentanglement metrics are correlated and we corroborate this result for our experiments by evaluating additional metrics (see Appendix B). In the following, we focus on the DCI Disentanglement (Eastwood and Williams, 2018) as a proxy for disentanglement in general.

Experimental results We see in Figure 1 that our approach outperforms all considered baseline methods on all considered data sets. Our model learns the most disentangled representations from the sequential data in all cases, which shows that GP-VAE models are well suited for the task of learning such representations when we consider data with a clear temporal structure. The additional metrics we evaluate confirm these results (see Appendix B.1). We observe that the weakly-supervised Ada-GVAE model is the runner-up in terms of its disentanglement in all of our experiments, probably due to the fact that it is the only other model that explicitly exploits potential correlations in time. While the Ada-GVAE approach outperforms the other (fully unsupervised) models, our model outperforms it on time series data with potentially dense changes of underlying factors. This is likely due to the fact that the Ada-GVAE assumes that changes in the underlying factors between consecutive time points are sparse, while we make the less restrictive assumption that the underlying factors change smoothly over time, as modeled by the chosen GP kernel. We argue that our more general assumptions about the structure of sequential data are better aligned with data found in the real world and test this hypothesis in the next experiment.

3.2. Real medical data

Real-world clinical time series data commonly consist of noisy, high-dimensional observations. Clinicians are trained to interpret these data and find patterns that help them identify the underlying causes for these observations. The observations give clues, but the real interest lies in the underlying health states that give rise to the data. Learning disentangled representations from high-dimensional medical time series could allow us to infer the state of independent clinical entities from these data, making the health states more salient.

Experimental setup To investigate the ability of our model to disentangle real-world medical time series data, we consider the HiRID data set (Hyland et al., 2020), consisting of 18 clinical variables for over 33,000 patients. We chose to compare our model with the Ada-GVAE since it is the only baseline model for disentanglement that explicitly takes the time series aspect into account. Moreover, it was the second-best model in our previous experiment and therefore the most serious competitor.

Measuring the disentanglement of learned representations on the HiRID data is not as straightforward as for the benchmark experiments since we do not have access to the underlying factors of generation. First, we group the different clinical variables into independent clinical concepts, with the help of a medical expert (for details, see Appendix C). Then, we train a classifier to predict the observed variables from the learned latent representation. The classifier returns the importance of each latent factor for predicting a given input variable and, by aggregating the variables into the aforementioned groups, we get a distribution over which latent variable each clinical concept is mapped to. We calculate the Disentanglement and Completeness of these distributions according to the DCI Disentanglement metric (Eastwood and Williams, 2018) in order to enable a quantitative comparison.

Additionally, we define a downstream mortality prediction task as a proxy measure for the informativeness of the representations. We train a linear classifier to predict the mortality label of each sample using the learned representation and report the performance in terms of the area under the receiver-operator-chracteristic curve (AUROC). Intuitively, a more disentangled representation should improve the test performance of a downstream classifier trained on this representation, as has indeed been shown by Locatello et al. (2020a).

Metric	DGP-VAE	Ada-GVAE
Dis	**0.325± 0.013**	0.133± 0.005
Com	**0.358± 0.013**	0.185± 0.007
AUROC	**0.769± 0.010**	0.684± 0.019

Table 1: DCI Disentanglement (Dis), DCI Completeness (Com), and downstream performance (AUROC) on a mortality prediction task.

Experimental results We observe that our model outperforms the Ada-GVAE approach in terms of Disentanglement and Completeness of the learned representations of the HiRID data set (Table 1). The additional metrics we report in Appendix B.2 reflect this as well. This indicates that we are able to more successfully learn which observed features arise from a shared clinical concept and to separate these independent concepts in the latent space. Qualitatively, this can also be observed in the comparison of the resulting feature mappings to the latent space, presented in Appendix B.2. Our representations also lead to a higher downstream performance on the mortality prediction task (Table 1), further highlighting their usefulness and confirming the superiority of our model.

The comparison of our model to the Ada-GVAE demonstrates that our assumptions are better aligned with realistically occurring time series data, at least in the medical setting. The experimental results can furthermore be explained by our model's ability to capture long-range dependencies in sequential data, while the Ada-GVAE can only hope to exploit correlations in directly neighbouring samples. Since we use GP priors to model the latent space, our model also automatically yields smooth and denoised latent time series, providing a more interpretable representation of a patient's physiological state than the original noisy and high-dimensional observations, which we visualize in Appendix B.2.

4. Conclusion

In this paper, we investigated the properties of a GP-VAE model to learn disentangled representations from time series data. Our model uses Gaussian process priors to model the latent space together with a structured variational distribution to capture dependencies in time. We showed that, in contrast to previous disentanglement methods, our approach yields disentangled representations "for free" by virtue of its prior, without relying on additional engineering tricks. We also demonstrated that it outperforms state-of-the-art models on benchmark disentanglement tasks involving sequential image data. Our modeling assumptions for the structure of time series data are more permissive than those of previous methods (Locatello et al., 2020a), and we provide evidence that they are better aligned with realistically occurring data by showing our model's favorable performance on real-world medical time series data in terms of disentanglement and downstream classification performance.

References

Tameem Adel, Zoubin Ghahramani, and Adrian Weller. Discovering interpretable representations for both deep generative and discriminative models. In *International Conference on Machine Learning*, pages 50–59, 2018.

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.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Sarthak Bhagat, Shagun Uppal, Zhuyun Yin, and Nengli Lim. Disentangling multiple features in video sequences using gaussian processes in variational autoencoders. In *European Conference on Computer Vision (ECCV)*, pages 102–117, 2020.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773.

Chris Burgess and Hyunjik Kim. 3D Shapes dataset, 2018. URL https://github.com/deepmind/3dshapes-dataset/.

Christopher P. Burgess, Irina Higgins, Arka Pal, Loïc Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. In *Workshop on Learning Disentangled Representations at the 31st Conference on Neural Information Processing Systems*, 2017.

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.

Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 2610–2620, 2018.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016.

Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

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

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. β-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

Wei-Ning Hsu, Yu Zhang, and James Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. In *Advances in Neural Information Processing Systems*, pages 1878–1889, 2017.

Stephanie L. Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, Marc Zimmermann, Dean Bodenham, Karsten Borgwardt, Gunnar Rätsch, and Tobias M. Merz. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine*, 26(3):364–373, 2020. doi: 10.1038/s41591-020-0789-4.

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

Metod Jazbec, Michael Arthur Leopold Pearce, and Vincent Fortuin. Factorized gaussian process variational autoencoders. In *Third Symposium on Advances in Approximate Bayesian Inference*, 2021.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999. doi: 10.1023/A:1007665907178.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658, 2018.

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

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

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40: e253, 2017. doi: 10.1017/S0140525X16001837.

Y. LeCun, Fu Jie Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–104 Vol.2, 2004. doi: 10.1109/CVPR.2004.1315150.

Yingzhen Li and Stephan Mandt. Disentangled sequential autoencoder. In *International Conference on Machine Learning*, pages 5670–5679, 2018.

F. Locatello, B. Poole, G. Rätsch, B. Schölkopf, O. Bachem, and M. Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pages 7753–7764, 2020a.

F. Locatello, M. Tschannen, S. Bauer, G. Rätsch, B. Schölkopf, and O. Bachem. Disentangling factors of variations using few labels. In *International Conference on Learning Representations*, 2020b.

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, 2019a.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pages 4114–4124, 2019b.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dSprites: Disentanglement testing sprites dataset, 2017. URL https://github.com/deepmind/dsprites-dataset/.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

Scott E Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. In *Advances in Neural Information Processing Systems*, pages 1252–1260, 2015.

Karl Ridgeway. A survey of inductive biases for factorial representation-learning. *arXiv preprint arXiv:1612.05299*, 2016.

Karl Ridgeway and Michael C Mozer. Learning deep disentangled embeddings with the f-statistic loss. In *Advances in Neural Information Processing Systems*, pages 185–194, 2018.

S. Roberts, M. Osborne, M. Ebden, S. Reece, N. Gibson, and S. Aigrain. Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371, 2013. doi: 10.1098/rsta.2011.0550.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Vedavyas Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy P. Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489, 2016.

Michael Tschannen, Olivier Bachem, and Mario Lucic. Recent advances in autoencoder-based representation learning. In *Workshop on Bayesian Deep Learning at the 32nd Conference on Neural Information Processing Systems*, 2018.

S. van Steenkiste, F. Locatello, J. Schmidhuber, and O. Bachem. Are disentangled representations helpful for abstract visual reasoning? In *Advances in Neural Information Processing Systems 32*, pages 14222–14235, 2019.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008. doi: 10.1561/2200000001.

Appendix A. Implementation details

The inference network in our model is implemented with a convolutional neural network (CNN) and for the generative network we use a multilayer perceptron (MLP). Details of the network hyperparameters used in the respective experiments are provided in the following. Our experiments took a combined 850 GPU hours on our internal cluster of NVIDIA GeForce GTX 1080 Ti GPUs.

A.1. Benchmark experiment

Since all data sets we consider in our benchmark experiment consist of images, we preprocess them using 2D convolutional layers. This intermediate representation at each time step is then flattened before being used as the input to a 1D convolution over time. Since the input to our model is the entire time series, for practical reasons we consider subsections of the original sequence for the temporal convolution step. This may be viewed as a limitation of our model compared to architectures that use RNNs, but we our results indicate that it did not prove to cause problems in practice. The hyperparameters of our model for this experiment are provided in Table A.1. It is worth noting that we found setting the β hyperparameter to 1.0 yielded the best results in terms of disentanglement, although one would expect higher values to encourage disentanglement more.

All baseline methods that we compare against are implemented with the `disentanglement_lib` (Locatello et al., 2019b) and their hyperparameters are given in Table A.2. The evaluation of the DCI metric is also implemented with the `disentanglement_lib`. We consider a train/test split of 8000/2000 data points for the classifier used in the DCI score calculation.

Hyperparameter	Value
Number of CNN layers in inference network	1
Number of filters per CNN layer	32
Filter size	3
Number of feedforward layers in inference network	2
Width of feedforward layers	256
Dimensionality of latent space	64
Length scale of Cauchy kernel	2.0
Number of feedforward layers in generative network	3
Width of feedforward layers	256
Activation function	ReLU
Optimizer	Adam (Kingma and Ba, 2015)
Learning rate	0.001
Training epochs	1
Train/test split (dSprites, SmallNORB)	10000/500
Train/test split (Cars3D, Shapes3D)	6190/310
Dimensionality if time points (dSprites, SmallNORB)	4096
Dimensionality if time points (Cars3D, Shapes3D)	12288
Original time series length	100
Training time series subsection length	5
Tradeoff parameter β	1.0

Table A.1: Hyperparameters used in the DGP-VAE model for the benchmark experiments.

Model	Hyperparameter	Value
Ada-GVAE	β	1.0
AnnealedVAE	c_{max}	25
	Iteration threshold	100000
	γ	100
β-VAE	β	4.0
β-TCVAE	β	4.0
FactorVAE	γ	30
DIP-VAE-I	λ_{od}	5
	λ_d	$10\lambda_{od}$
DIP-VAE-II	λ_{od}	5
	λ_d	λ_{od}

Table A.2: Hyperparameters used for the baseline models.

A.2. HiRID experiment

Since the HiRID data does not consist of images, we omit the convolutional preprocessing and perform the 1D convolution time directly on the input data. The details of the utilized hyperparameters for this experiment are provided in Table A.3. For the calculation of the DCI score we use a train/test split of 20000/5000.

Hyperparameter	Value
Number of CNN layers in inference network	1
Number of filters per CNN layer	128
Filter size	12
Number of feedforward layers in inference network	1
Width of feedforward layers	128
Dimensionanlity of latent space	8
Length scales of Cauchy kernel	[20.0, 10.0, 5.0, 2.5]
Number of feedforward layers in generative network	2
Width of feedforward layers	256
Activation function	ReLU
Optimizer	Adam (Kingma and Ba, 2015)
Learning rate	0.001
Training epochs	1
Train/test split	517995/25900
Dimensionality if time points	18
Original time series length	100
Training time series subsection length	25
Tradeoff parameter β	1.0

Table A.3: Hyperparameters used in the DGP-VAE model for the HiRID real-world medical data set experiment.

Appendix B. Additional results

B.1. Benchmark experiment

While we focus on the disentanglement component of the the DCI metric (Eastwood and Williams, 2018), we also evaluate the DCI Completeness, DCI Informativeness, Mutual Information Gap (MIG) (Chen et al., 2018), Modularity (Ridgeway and Mozer, 2018), and Seperated Attribute Predictability (SAP) (Kumar et al., 2018) scores for the benchmark experiment (see Figure B.1). Overall, we observe that these additional scores reflect the results reported in Section 3.1 and that our model outperforms the considered baselines, with some minor exceptions. This confirms the findings of Locatello et al. (2019b) that all common disentanglement metrics are positively correlated, and further justifies our focus on the DCI Disentanglement in this work.

B.2. HiRID experiment

We also report the Modularity and the SAP score for the experiment with real-world medical time series data in Table B.1. Both of these metrics are calculated in a similar spirit to the modified DCI Disentanglement and Completeness scores, i.e. treating the observed input features as the ground truth for the respective score calculation and aggregating them in an intermediate step according the their mapping to independent clinical concepts. This facilitates the calculation of the respective metrics even without having access to ground

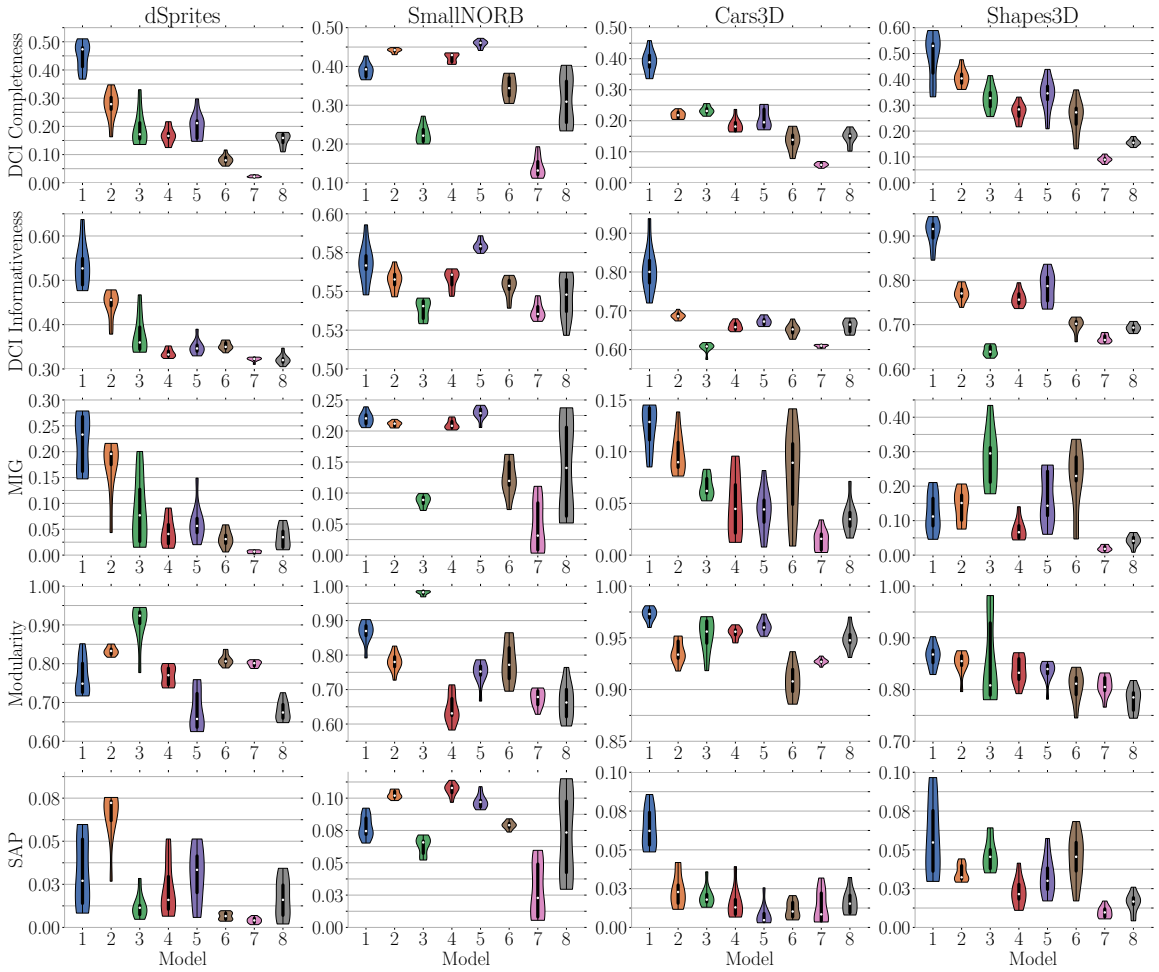

Figure B.1: Results for additional metrics on the benchmark experiment described in Sec. 3.1. The models we compare are: 1) DGP-VAE (ours), 2) Ada-GVAE, 3) AnnealedVAE, 4) β-VAE, 5) β-TCVAE, 6) FactorVAE, 7) DIP-VAE-I, and 8) DIP-VAE-II.

truth factors of variation. We omit the MIG score for this experiment, as its calculation does not lend itself to this reformulation by mapping features into independent clinical concepts.

Additionally, we visualize the latent time series obtained by our model and compare them to those given by the Ada-GVAE on Figure B.2. In Figure B.3, we compare the mapping of concepts to latent dimensions obtained by our model to the mappings of the Ada-GVAE. While our model learns to cluster related clinical concepts, the Ada-GVAE does not display such disentangled mappings.

Metric	DGP-VAE	Ada-GVAE
Modularity	**0.873± 0.005**	0.790± 0.010
SAP	**0.066± 0.007**	0.022± 0.002

Table B.1: Comparison of the Modularity and SAP scores between our model and the Ada-GVAE trained on the HiRID data set. Our model also outperforms the Ada-GVAE in terms of these additional metrics.

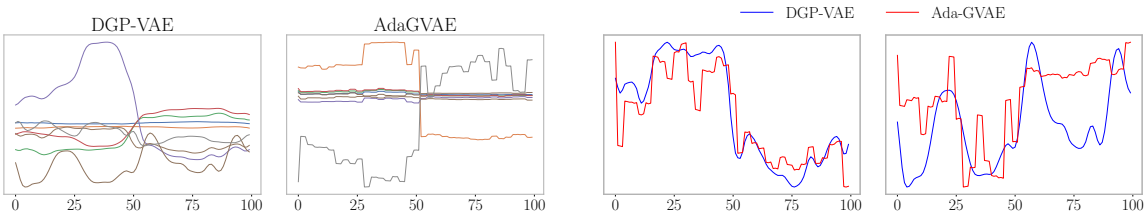

Figure B.2: (**left**) The latent time series in our model display dynamics on multiple time scales and multiple channels may change at once, while the latent series of the Ada-GVAE model does not display dense changes of latent factors. (**right**) Detailed time series of two latent channels. Our model learns smooth latent time series, while the Ada-GVAE exhibits more noisy dynamics.

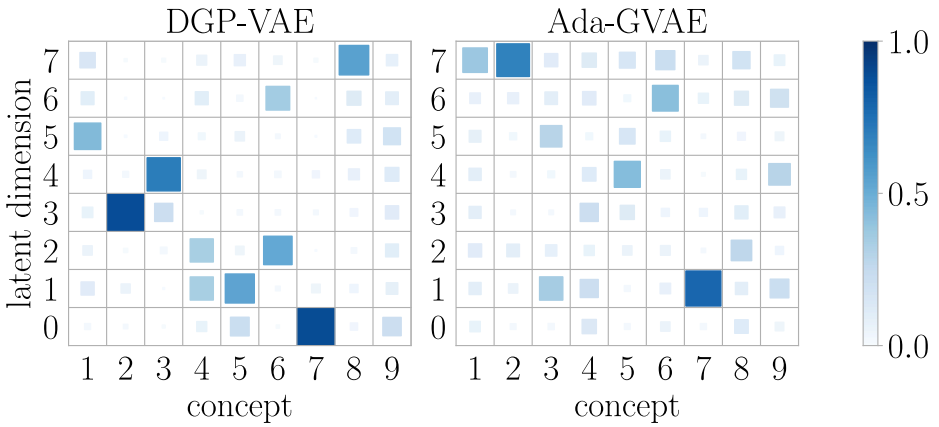

Figure B.3: Learned HiRID concept mappings. Our model learns to cluster related clinical concepts, indicated by features being mapped to a single latent, while disentangling unrelated concepts from each other, i.e. mapping independent concepts to different latents. The Ada-GVAE is not as successful in learning such mappings and also finds spurious relationships in the data. The mapping of HiRID features to clinical concepts is presented in Appendix C.

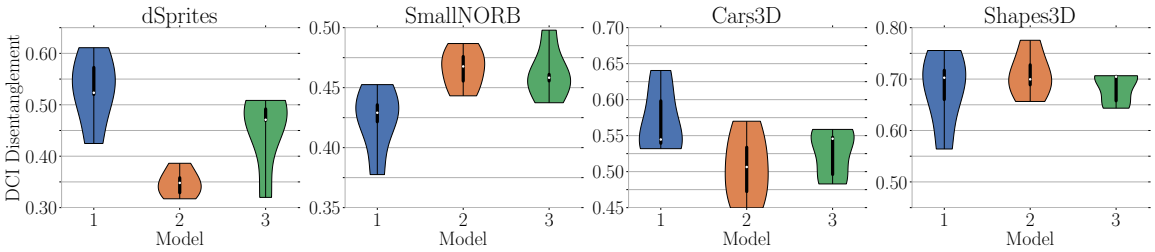

Figure B.4: Comparison of the Disentanglement score of our model (1), which uses a Cauchy kernel, to the MGP-VAE model with a Brownian bridge kernel (2) and a fractional Brownian motion kernel (3). For details on these additional kernels, see the original MGP-VAE paper (Bhagat et al., 2020).

B.3. Kernel comparison

We also investigated the kernels that the authors of the MGP-VAE model propose (Bhagat et al., 2020), namely the Brownian bridge kernel (BB) and the fractional Brownian motion kernel (fBM). The results presented in Figure B.4 show that these kernels do not provide any notable benefits in terms of disentanglement over the Cauchy kernel in our model, which is also computationally simpler.

Appendix C. Ground truth HiRID feature mapping

To evaluate how well our model can learn disentangled representations of medical time series data we require a ground truth mapping of the occuring medical variables to independent clinical concepts. This mapping, which was provided by a medical expert, is given in Table C.1.

Appendix D. Detailed related work

D.1. Sequential data in VAEs

Models based on variational autoencoders (VAEs) have succesfully been applied to tasks involving—to some extent—the disentanglement of these data. The factorized hierarchical variational autoencoder (FHVAE) introduced by Hsu et al. (2017) aims to exploit the correlations of sequential data by introducing two different hierarchical priors to the latent representation. The authors argue that this captures the multi-scale nature of sequential data and disentangles features that are shared across a sequence from those that vary from one sequence segment to another.

In a similar spirit, Li and Mandt (2018) introduce the Disentangled Sequential Autoencoder, a model that learns to disentangle static from dynamic parts of the data's representation. This is achieved by means of a factorized graphical model that encodes sequence-invariant information into one latent variable and all dynamic information into a separate set of latent variables. While both of the aforementioned models successfully dis-

Concept	Variable	Unit
1) Cardiovascular	Heart rate	$[bpm]$
	Systolic blood pressure (invasive)	$[mmHg]$
	Diastolic blood pressure (invasive)	$[mmHg]$
	MAP	$[mmHg]$
	Cardiac output	$[l/min]$
2) Lungs	SpO2	$[\%]$
	Peak inspiratory pressure (ventilator)	$[cmH_2O]$
3) Sedation level	RASS	[-]
4) Glucose	Serum glucose	$[mmol/l]$
5) Blood clotting	INR	[-]
6) Metabolic	Lactate arterial	$[mmol/l]$
	Lactate venous	$[mmol/l]$
7) Inflammation	C-reactive protein	$[mg/l]$
8) Heart medication	Dobutamine	Flow $[mg/min]$
	Milrinone	Flow $[mg/min]$
	Levosimendan	Flow $[mg/min]$
	Theophyllin	Flow $[mg/min]$
9) Pain medication	Non-opiod analgesics	[-]

Table C.1: Mapping of HiRID variables into independent clinical concepts.

entangle static from dynamic features, they do not disentangle the underlying, *individual* dynamic factors.

The class of models that employ Gaussian process priors for the latent variables of a variational autoencoder and thereby exploit the correlation of sequential data it the latent space have been successfully applied to a wide range of tasks. The Gaussian Process Prior Variational Autoencoder (GPPVAE) (Casale et al., 2018) was the first model to introduce this family of priors in the context of VAEs and break with the assumption that samples of the latent distribution must be independent and identically distributed (i.i.d.), thereby better modeling the specifications of sequential data. While the basis of this model is shared with ours, we use one GP per latent channel, as opposed to a joint GP prior over the whole data. This allows us to rely on standard inference techniques as opposed to the specialized inference method of the GPPVAE model, while also encouraging disentanglement between the latent channels.

The Multi-disentangled-features Gaussian Process Variational Autoencoder (MGP-VAE) (Bhagat et al., 2020) extends the GPPVAE model by using fractional Brownian motion and Brownian bridge kernels for the latent GP prior of each channel. The authors argue that this setting allows for the disentanglement of static as well as dynamic features, but only show qualitative results for sparsely changing input time series. In contrast, we show the efficacy of our approach to dynamic sequential data with dense changes of the factors in time.

The Gaussian Process Variational Autoencoder (GP-VAE) model introduced by Fortuin et al. (2020) can be viewed as a further extension to Casale et al. (2018)'s GPPVAE model. This model introduces a GP prior with a Cauchy kernel to each latent channel in

combination with a structured variational inference technique to impute missing values in time series data. While the GP-VAE was designed and used for missing data imputation we show that a very similar architecture lends itself to learning disentangled representations from dynamic sequential data.

The Sparse GP-VAE (SGP-VAE) (Ashman et al., 2020) and Scalable GP-VAE (SVGP-VAE) (Jazbec et al., 2020) extend the class the GP-VAE models with a sparse GP approximation, parameterized by a partial inference network. Moreover, the Factorized GP-VAE (FGP-VAE) (Jazbec et al., 2021) further improves the inference speed by using Kronecker-factorized kernels. These extensions could also be readily applied to our model, which we have not done in this study, since exact inference was still feasible in our experiments.

D.2. Disentangled representation learning

All state-of-the-art approaches to disentangled representation learning rely on variational autoencoders (Kingma and Welling, 2014) as their architectural backbone. Alternatives based on generative adversarial networks (GANs) (Goodfellow et al., 2014) such as variants of InfoGAN (Chen et al., 2016) have also been proposed, but previous work has found their performance to not be competitive when compared to VAE-based approaches (Kim and Mnih, 2018).

In the VAE setting, the representation $r(\mathbf{x})$ of a sample $\mathbf{x}$ is taken as the mean of the approximate posterior distribution $q(\mathbf{z}|\mathbf{x})$, that is, the encoding of a sample from the feature space to the latent space. The following approaches share the common theme of disentangling this approximate posterior, while their main differences arise from how this disentanglement is enforced.

Unsupervised models The β-VAE model (Higgins et al., 2017) adds a simple hyperparameter to the KL term of the standard ELBO. This β hyperparameter balances reconstruction quality with latent channel capacity and setting it greater than unity enforces the encoder distribution to better match the factorized Gaussian prior.

The AnnealedVAE model (Burgess et al., 2017) also focuses on latent bottleneck capacity. The authors argue that limiting the latent channel capacity forces the model to learn a single factor of variation at a time. Therefore, the bottleneck capacity is gradually increased during training to enforce the sequential learning of separate underlying factors of variation.

By further decomposing the vanilla VAE objective, the authors of the β-TCVAE model (Chen et al., 2018) identify a term which measures the total correlation between latent variables. By specifically penalizing this total correlation term, the model enforces disentanglement without adding any additional hyperparameters compared to β-VAE.

The FactorVAE model (Kim and Mnih, 2018) also penalizes the total correlation term, but differs in implementation compared to β-TCVAE. The idea however remains to push the aggregated posterior $q(\mathbf{z})$ towards a factorized form, thus enforcing independence across latent dimensions, without having to sacrifice reconstruction quality for disentanglement.

The authors of the DIP-VAE model (Kumar et al., 2018) propose to enforce disentanglement by means of disentangled priors. The aggregated posterior is pushed towards these disentangled priors by means of penalizing an arbitrary convergence between the two. Subtle implementation differences result in two models: DIP-VAE-I and DIP-VAE-II.

Weakly-supervised model The Ada-GVAE model (Locatello et al., 2020a) differs from the previously introduced approaches in that it is not a fully unsupervised approach, but utilizes a form of weak supervision to improve disentanglement. The authors acknowledge Locatello et al. (2019b)'s proof that learning disentangled representations is impossible without inductive biases. They therefore attempt to explicitly include such an inductive bias in their modeling assumptions and exploit this for the purposes of disentanglement. They make the assumption that some underlying factors of variation may be shared across pairs of input data. They then go on to prove that knowing the number of shared factors across individual pairs is sufficient to learn a fully disentangled representation. The introduced model provides an algorithm that is an extension of β-VAE (Higgins et al., 2017), which estimates the number of shared factors in a pair of samples and enforces the sharing of a latent representation for these estimated shared factors. The authors argue that a source of data where this assumption could be justified is sequential data, which inspired us to be explicit about the assumptions we make on sequential data and include these in the form of the inductive bias of smoothly varying GP priors.