# Nonparametric Inference for Auto-Encoding Variational Bayes

**Erik Bodin** [*] **Iman Malik** [*] **Carl Henrik Ek** [*] **Neill D. F. Campbell** [†]

[*] **University of Bristol** [†] **University of Bath**

Variational approximations are an attractive approach for inference of latent variables in unsupervised learning. However, they are often computationally intractable when faced with large datasets. Recently, Variational Autoencoders (VAEs) Kingma and Welling [2014] have been proposed as a method to tackle this limitation. Their methodology is based on formulating the approximating posterior distributions in terms of a deterministic relationship to observed data consequently the title "Auto-Encoding Variational Bayes". Importantly, this is a decision regarding an approximate inference scheme that should not be confused with an auto-encoder as a model.

Unsupervised learning is a ill-conditioned problem that requires prior knowledge to reach a solution. We would like to learn latent representations that are low-dimensional and highly interpretable. A model that has these characteristics is the Gaussian Process Latent Variable Model (GP-LVM) Lawrence [2005]. The benefits and negative of the GP-LVM are complementary to the VAE, the former provides useful low-dimensional latent representations while the latter is able to handle large amounts of data and can use non-Gaussian likelihoods. Our inspiration for this paper is to marry these two approaches and reap the benefits of both. In order to do so we will introduce a novel approximate inference scheme inspired by the GP-LVM to the VAE.

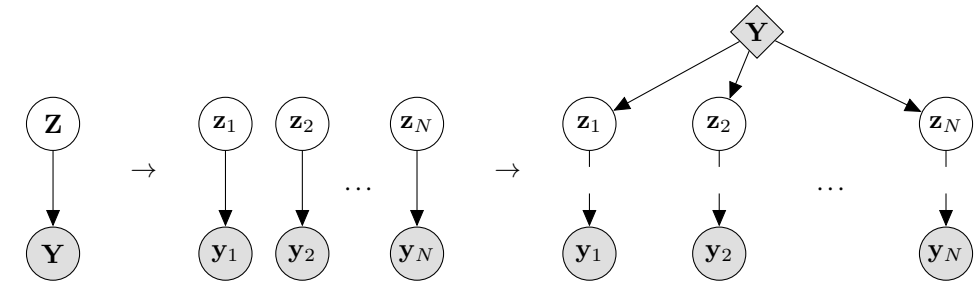
The standard VAE formulation Kingma and Welling [2014] adopts a unit Gaussian prior, creating a trade-off between the embedded data residing at the same location in the latent space and the ability to reconstruct the data in the observed space. This encourages a "tight packing" of the data around a shared origin, with hope that similar data in the observed space have overlapping probability mass in the latent space i.e. mutual information. It has been shown that a simple prior over-regularizes the latent space leading to poor reconstructions Hoffman and Johnson [2016]. Other more flexible priors can be used to change the dynamic between reconstruction and mutual information, such as in recent work of using a mixture Tomczak and Welling [2017] or let the prior be autoregressive Chen et al. [2016]. The quality of the output from a VAE can be improved by more expressive generative models, but this has been shown to lead to a tendency of ignoring the latent space, defeating the purpose of unsupervised learning Zhao et al. [2017]. In summary, reconstruction quality and mutual information in the latent space is traded against each other.

In this paper we address this limiting trade-off by escaping it; we let the space where we encourage sharing be separated from the space where the generative capacity is set. We do this by an approximation of a model where the observations can be generated from either space. We show experimentally that the approximation allows the capacity of the generative bottle-neck ($Z$) to be arbitrarily large without losing sharing and the beneficial properties of the sharing space ($X$), allowing reconstruction quality to be unlimited by $Z$ at the same time as a low-dimensional space can be used to perform ancestral sampling from as well as a means to reason about the embedded data.

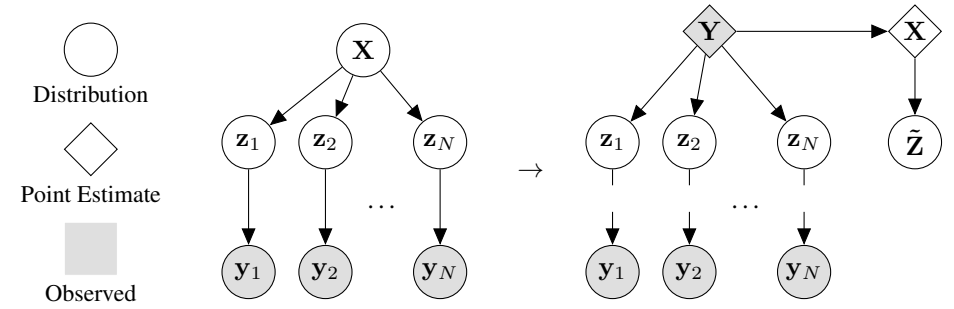
## 1 Method

The VAE Kingma and Welling [2014] inference scheme optimises a traditional evidence lower bound where the latent space posterior is approximated as a deterministic relationship from the observed data as $q(\mathbf{Z}|\mathbf{Y}) = \prod_{i=1}^{N} q(\mathbf{z}_i|\mathbf{y}_i)$ where each latent variable is conditionally independent given the

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.




**(a)** Model **(b)** VAE Approximation **(c)** VAE Inference Scheme




**(d)** Our Nonparametric VAE **(e)** Our Inference Scheme

Figure 1: Contrasting variational approximation schemes for unsupervised learning. (a) We specify an unsupervised generative model from latent $Z$ to observed $Y$. (b) The VAE proposes a fully factored variational approximation. (c) Inference proceeds by conditioning the variational latent parameters on observed data through an explicit deterministic function (e.g. MLP network). (d) Our model proposes an additional latent space $X$ that ties together the factored $Z$ space. (e) Inference then proceeds with $X$ also conditioned on observed data through an additional deterministic function. For tractable inference we match moments between $Z$ and $\tilde{Z}$.

observed data. In this paper we introduce an additional latent variable $\mathbf{X}$ that model the interaction between the latent variables $\mathbf{Z}$.

Our approach means that $\mathbf{Z}$ are no longer independent but conditionally independent given $\mathbf{X}$. This leads to the following updated evidence lower bound with an additional divergence term as,

$$\mathcal{L}_s = \mathcal{L}_g - \text{KL}(q_g(\mathbf{Z}|\mathbf{Y})||p_s(\mathbf{Z}|\mathbf{X})), \tag{1}$$

where $\mathcal{L}_g$ and $q_g(\mathbf{z}_i|\mathbf{y}_i)$ is the standard VAE lower bound and approximative posterior respectively. To facilitate the use of batch processing we will rather than matching the joint distribution of the latent space match the predictive posteriors. As each $\mathbf{Z}$ is conditionally independent given the observed data $\mathbf{Y}$ this leads to the following updated objective function,

$$\mathcal{L}'_s = \mathcal{L}_g - \text{KL}(\prod_i q_g(\mathbf{z}_i|\mathbf{y}_i)||\prod_i p_s(\mathbf{z}_i|\mathbf{X}, \mathbf{z}_{\neg i})). \tag{2}$$

The predictive posterior of the Gaussian process is,

$$\begin{aligned} p(\mathbf{z}_i|\mathbf{X}, \mathbf{z}_{\neg i}) = \mathcal{N}(&k(\mathbf{x}_i, \mathbf{X}_{\neg i})k(\mathbf{X}_{\neg i}, \mathbf{X}_{\neg i})^{-1}\mathbf{z}_{\neg i}, \\ &k(\mathbf{x}_i, \mathbf{x}_i) - k(\mathbf{x}_i, \mathbf{X}_{\neg i})k(\mathbf{X}_{\neg i}, \mathbf{X}_{\neg i})k(\mathbf{X}_{\neg i}, \mathbf{x}_i)) \end{aligned} \tag{3}$$

where $k(\cdot, \cdot)$ is the covariance function. Evaluating this posterior for $N$ datapoints is computationally expensive due to the inverse of the covariance function. To proceed we will introduce two additional approximations, first we will approximate the mean of the predictive posterior of the GP by directly parametrising it as,

$$\mu(\mathbf{z}_i) = k(\mathbf{x}_i, \mathbf{X}_{\neg i})k(\mathbf{X}_{\neg i}, \mathbf{X}_{\neg i})^{-1}\mathbf{z}_{\neg i} \approx \mathbf{W}_i\mathbf{z}_{\neg i}. \tag{4}$$

In specific we will parametrise the weight matrix $\mathbf{W}_i \in \mathbb{R}^{1\times N}$ such that the approximative predictive mean for $\mathbf{z}_i$ is a convex combination of $\mathbf{z}_{\neg i}$, i.e. that $\sum_j W_{ij} = 1, \forall i$ and $W_{ij} \geq 0, \forall i, j$. The

intuition behind this is that we want to encourage sharing in the latent space $X$ such that the space $Z$ is represented in a distributed fashion. Secondly rather than minimising the KL-divergence we will match the first mode of the two distributions. This leads to our final objective function,

$$\tilde{\mathcal{L}}_s = \mathcal{L}_g - \sum_i (\mathbb{E}[\mathbf{z}_i] - \mathbb{E}[\tilde{\mathbf{z}}_i])^2, \tag{5}$$

where $\tilde{\mathbf{z}}$ is the prediction of the latent space from $X$ while $\mathbf{z}$ is the prediction using the approximative posterior from the VAE. In effect we have separated the two models while retaining a connection by matching their first modes when predicting the latent space $Z$. We will now proceed to show the experimental evaluation of the model showing that we are capable of using the additional low-dimensional latent space $X$ as a proxy for the VAE latent space $Z$.

## 2 Experiments

In this paper we choose to model $W_{i,j}$ as a squared exponential covariance kernel between input locations, $\mathbf{x}_i$ and $\mathbf{x}_j$,

$$W_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2l^2}\right) \qquad i \neq j \tag{6}$$

and infer its parameters. We ensure $W_{i,i} = 0$ by subtracting the diagonal from the computed $\mathbf{W}$ and normalise by rows to create a convex combination. The latent locations $X$ are represented implicitly as a function of the observed data as in Lawrence and Quiñonero-Candela [2006].

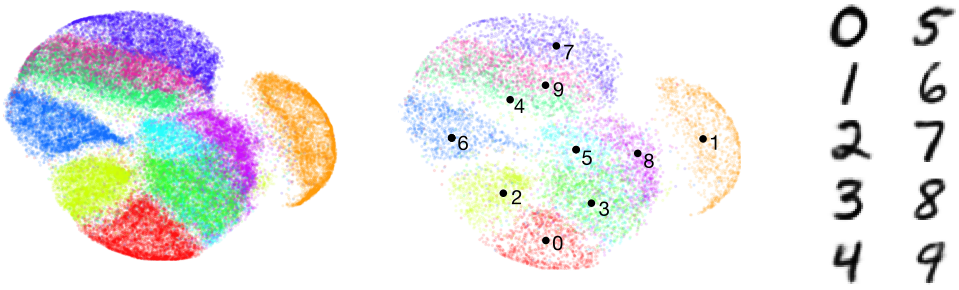
We trained models to illustrate that our extension can be used to obtain a low-dimensional $X$ space that is highly interpretable while permitting the use of a high-dimensional $Z$ space to provide high quality data generation. We validate our approach, with comparison to a standard VAE, by showing data embeddings in the $X$ space and generation of new data.

All experiments were performed with the decoder and both encoders as Multilayer Perceptrons (MLP) with the same architecture as in the original VAE Kingma and Welling [2014]; we used two hidden layers of $500$ units each, mini-batch sizes of $128$ and a drop-out probability of $0.9$ throughout training. The decoder used was the Bernoulli MLP variant. Furthermore, the ADAM Kingma and Ba [2014] optimiser was used with a learning rate of $10^{-3}$. We varied the dimensionality of the inner most layer of the autoencoder (the $Z$ space) for the experiments. We used the MNIST data set from LeCun et al. [1998] comprised of $55\,000$ training examples and $10\,000$ test examples of $28 \times 28$ pixel greyscale images, corresponding to $784$ data dimensions.

In Fig. 2 we show a 2-dimensional $X$ space corresponding to a $500$-dimensional $Z$ space with both the training and test data embedded as well as examples of data generation. Despite using a high capacity $Z$ space, the nonparametric VAE can still sample from a low dimension (in our example 2D) using ancestral sampling from the $X$ space. This ensures that test samples maintain a high fidelity from a space that is highly interpretable, easy to visualise and easy to sample from. In Fig. 3 we show corresponding $X$ space embeddings for different dimensionalities of $Z$; this demonstrates that the $X$ space maintains its virtues independent of the $Z$ dimension. Finally, in Fig. 4, we show sample interpolations in the latent space for the standard VAE (directly in $Z$) and with our extension (in $X$) illustrating that not only is the reconstruction quality preserved but interpolations are more meaningful.

## 3 Conclusions

We have presented a hierarchical model for unsupervised learning and an associated efficient approximative inference scheme. The inference takes inspiration from amortised inference and use a recognition model to parameterise the approximate posterior using a deterministic relationship from the observed data. Rather than using a traditional mean-field approximation which forces the latent representation to be independent we introduce an additional latent representation that models their dependence. Our model results in a significantly lower dimensional latent representation allowing us to visualise and generate data in a intuitive manner without sacrificing the quality of the reconstruction. We have shown experimental results on how we can retain the representative power of a 500 dimensional model with just a 2 dimensional latent space.




**(a)** Training $X$ **(b)** Test $X$ **(c)** Samples

Figure 2: Learned $X$ space embeddings from the nonparametric VAE. Inferred $X$ locations for (a) the training data and (b) the test data with colors encoding the MNIST digit classes. (c) Generated samples from the corresponding locations in (b) using a $Z$ space with 500 dimensions.

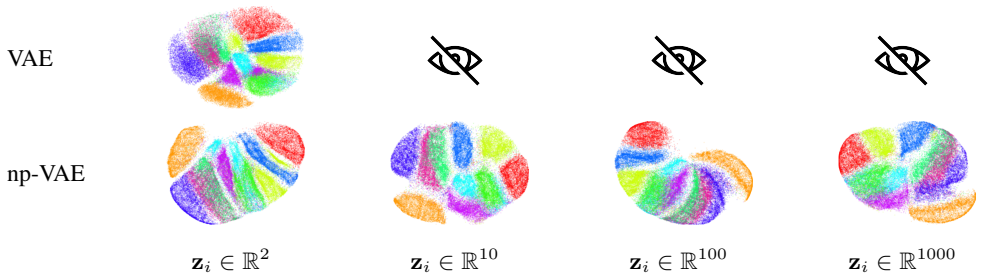



Figure 3: Latent space visualisation. Upper row: the $Z$ space embedding is visualised for the standard VAE where possible (we are unable to do this for high-dimensional $Z$). Bottom row: the same $Z$ space dimensionalities are used but the nonparametric VAE allows the $X$ space to be visualised and sampled (set to be 2-dimensional). $Z$ spaces of higher dimension become impractical to visualise and interpret whereas the $X$ space provides an embedding for easy display and interpretation.

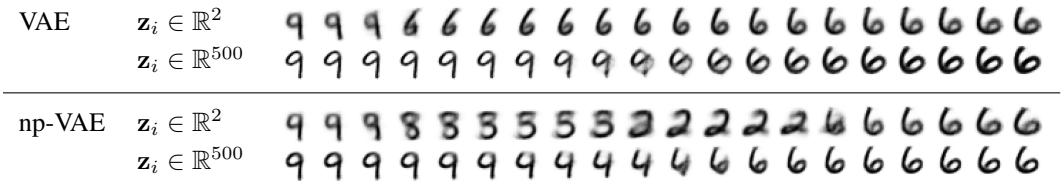



Figure 4: Latent space interpolation. The upper two rows show interpolants between two MNIST training examples for a standard VAE with a $Z$ latent dimensionality of 2 and 500. The bottom two rows show interpolants between the same training examples for our nonparametric VAE with the same respective dimensionalities for $Z$ but where the interpolation is performed in the inferred latent space $X$ of dimension 2. We observe that a similar reconstruction quality is obtained by corresponding $Z$-dimensionalities, however, the interpolants from the $X$ space of the nonparametric VAE are more meaningful with credible intermediate states between digits. Thus we can obtain a low dimensional latent space that provides interpretability without sacrificing reconstruction quality.

## References


Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Neil D Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783–1816, 2005.

Matthew D Hoffman and Matthew J Johnson. ELBO Surgery: Yet Another Way to Carve up the Variational Evidence Lower Bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, 2016.

Jakub M Tomczak and Max Welling. VAE with a VampPrior. *arXiv preprint arXiv:1705.07120*, 2017.

Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational Lossy Autoencoder. *arXiv preprint arXiv:1611.02731*, 2016.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Towards a Deeper Understanding of Variational Autoencoding Models. *arXiv preprint arXiv:1702.08658*, 2017.

Neil D Lawrence and J Quiñonero-Candela. Local distance preservation in the GP-LVM through back constraints. *Proceedings of the 23rd international conference on Machine learning*, pages 513–520, 2006.

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

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2323, 1998.