Learning Sparse Latent Representations with the Deep Copula Information Bottleneck

Anonymous authors
Paper under double-blind review

Abstract

Deep latent variable models are powerful tools for representation learning. In this paper, we adopt the deep variational information bottleneck model, identify its shortcomings and propose a model that circumvents them. To this end, we apply a copula transformation which, by restoring the invariance properties of the information bottleneck method, leads to disentanglement of the features in the latent space. Building on that, we show how to force sparsity of the latent space in the new model. We evaluate our method on artificial and real data.

1 Introduction

In recent years, deep latent variable models (Kingma & Welling, 2013; Rezende et al., 2014; Goodfellow et al., 2014) have become a popular toolbox in the machine learning community for a wide range of applications (Ledig et al., 2016; Reed et al., 2016; Isola et al., 2016). At the same time, the compact representation, sparsity and interpretability of the latent feature space have been identified as crucial elements of such models. In this context, multiple contributions have been made in the field of relevant feature extraction (Chalk et al., 2016; Alemi et al., 2016) and learning of disentangled representations of the latent space (Chen et al., 2016; Bouchacourt et al., 2017; Higgins et al., 2017).

In this paper, we consider latent space representation learning. We focus on disentangling features with the copula transformation and, building on that, on forcing an compact low-dimensional representation with a sparsity-inducing model formulation. To this end, we adopt the *deep variational information bottleneck (DVIB)* model (Alemi et al., 2016) which combines the information bottleneck and and variational autoencoder methods. The *information bottleneck (IB)* principle (Tishby et al., 2000) identifies relevant features with respect to a target variable. It takes two random vectors x and y and searches for a third random vector t which, while compressing x, preserves information contained in y. A *variational autoencoder (VAE)* (Kingma & Welling, 2013; Rezende et al., 2014) is a generative model which learns a latent representation t of x by using the variational approach.

Although DVIB produces good results in terms of image classification and adversarial attacks, it suffers from two major shortcomings. First, the IB solution only depends on the copula of x and y and is thus invariant to strictly monotone transformations of the marginal distributions. DVIB does not preserve this invariance, which means that it is unnecessarily complex by also implicitly modelling the marginal distributions. We elaborate on the fundamental issues arising from this lack of invariance in Section 3. Second, the latent space of the IB is not sparse which results in the fact that a compact feature representation is not feasible.

Our contribution is two-fold: In the first step, we restore the invariance properties of the information bottleneck solution in the DVIB. We achieve this by applying a transformation of x and y which makes the latent space only depend on the copula. This is a way to fully represent all the desirable features inherent to the IB formulation. The model is also simplified by ensuring robust and fully non-parametric treatment of the marginal distributions. In addition, the problems arising from the lack of invariance to monotone transformations of the marginals are solved. In the second step, once the invariance properties are restored, we

exploit the sparse structure of the latent space of DVIB. This is possible thanks to the copula transformation in conjunction with using the sparse parametrisation of the information bottleneck, proposed by (Rey et al., 2014). It translates to a more compact latent space that results in a better interpretability of the model.

The remainder of this paper is structured as follows: In Section 2, we review publications on related models. Subsequently, in Section 3, we describe the proposed copula transformation and show how it fixes the shortcomings of DVIB, as well as elaborate on the sparsity induced in the latent space. In Section 4, we present results of both synthetic and real data experiments. We conclude our paper in Section 5.

2 Related Work

The IB principle was introduced by (Tishby et al., 2000). The idea is to compress the random vector x while retaining the information of the random vector y. This is achieved by solving the following variational problem: $\min_{p(t|x)} I(x;t) - \lambda I(t;y)$, with the assumption that y is conditionally independent of t given x. In recent years, copula models were combined with the IB principle in (Rey & Roth, 2012) and extended to the sparse meta-Gaussian IB (Rey et al., 2014) to become invariant against strictly monotone transformations. Moreover, the IB method has been applied to the analysis of deep neural networks in (Tishby & Zaslavsky, 2015), by quantifying mutual information between the network layers and deriving an information theoretic limit on DNN efficiency.

The variational bound and reparametrisation trick for autoencoders were introduced in (Kingma & Welling, 2013; Rezende et al., 2014). The variational autoencoder aims to learn the posterior distribution of the latent space $p(t|x)$ and the *decoder* $p(x|t)$. The general idea of combining the two approaches is to identify the solution t of the information bottleneck with the latent space t of the variational autoencoder. Consequently, the terms $I(x;t)$ and $I(t;y)$ in the IB problem can be expressed in terms of the parametrised conditionals $p(t|x)$, $p(x|t)$.

Variational lower bounds on the information bottleneck optimisation problem have been considered in (Chalk et al., 2016) and (Alemi et al., 2016). Both approaches, however, treat the differential entropy of the marginal distribution as a positive constant, which is not always justified (see Section 3). A related model is introduced in (Pereyra et al., 2017), where a penalty on the entropy of output distributions of neural networks is imposed. These approaches do not introduce the invariance against strictly monotone transformations and thus do not address the issues we identify in Section 3.

A sizeable amount of work on modelling the latent space of deep neural networks has been done. The authors of (Alvarez & Salzmann, 2016) propose the use of a group sparsity regulariser. Other techniques, e.g. in (Mozer & Smolensky, 1989) are based on removing neurons which have a limited impact on the output layer, but they frequently do not scale well with the overall network size. More recent approaches include training neural networks of smaller size to mimic a deep network (Hinton et al., 2015; Romero et al., 2014). In addition, multiple contributions have been proposed in the area of latent space disentanglement (Chen et al., 2016; Bouchacourt et al., 2017; Higgins et al., 2017; Denton & Birodkar, 2017). None of the approaches consider the influence of the copula on the modelled latent space.

Copula models have been proposed in the context of Bayesian variational methods in (Suh & Choi, 2016), (Tran et al., 2015) and (Han et al., 2016). The former approaches focus on treating the latent space variables as indicators of local approximations of the original space. None of the three approaches relate to the information bottleneck framework.

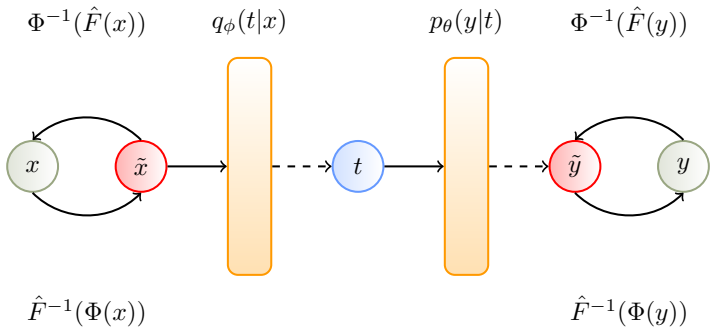


Figure 1: Deep variational information bottleneck with the copula augmentation. Green circles describe random variables and orange rectangles denote deep nets parametrising the random variables. The blue circle stands for latent random variables whereas the red circle denotes the copula transformed random variables.

3 Model

3.1 Formulation

In order to specify our model, we start with a parametric formulation of the information bottleneck:

$$\max_{\phi,\theta} -I_\phi(t;x) + \lambda I_{\phi,\theta}(t;y). \tag{1}$$

A parametric form of the conditionals $p_\phi(t|x)$ and $p_\theta(y|t)$ as well as the information bottleneck Markov chain $t - x - y$ are assumed. A graphical illustration of the proposed model is depicted in Figure 1.

The two terms in Eq. (1) have the following forms:

$$I_\phi(T;X) = D_{KL}\left(p(t|x)p(x) \| p(t)p(x)\right) = \mathbb{E}_{p(x)} D_{KL}\left(p_\phi(t|x) \| p(t)\right), \tag{2}$$

and

$$\begin{aligned} I_{\phi,\theta}(T;Y) &= D_{KL}\left(\left[\int p(t|y,x)p(y,x)\,\mathrm{d}x\right] \| p(t)p(y)\right) \\ &= \mathbb{E}_{p(x,y)}\mathbb{E}_{p_\phi(t|x)} \log p_\theta(y|t) + h(Y), \end{aligned} \tag{3}$$

because of the Markov assumption in the information bottleneck model $p_\phi(t|x,y) = p_\phi(t|x)$. We denote with $h(y) = -\mathbb{E}_{p(y)}[\log p(y)]$ the *entropy* for discrete y and the *differential entropy* for continuous y. We then assume a conditional independence copula and Gaussian margins:

$$p_\phi(t|x) = c_{T|X}(u(t|x)|x) \prod_j p_{\phi_j}(t_j|x) = \prod_j N(T_j|\mu_j(x), \sigma_j^2(x))$$

where t_j are the marginals of $t = (t_1, \ldots, t_d)$, $c_{t|x}$ is the copula density of $t|x$, $u(t|x) := F_{t|x}(t|x)$ is the uniform density indexed by $t|x$, and the functions $\mu_j(x), \sigma_j^2(x)$ are implemented by a deep network. We make the same assumption about $p_\theta(y|t)$.

3.2 Motivation

As we stated in Section 1, the deep variational information bottleneck model derived in Section 3.1 is not invariant to strictly increasing transformations of the marginal distributions. The IB method is formulated in terms of mutual information $I(x,y)$, which depends only on the copula and therefore does not depend on monotone transformations of the marginals:

$I(x,y) = MI(x,y) - MI(x) - MI(y)$, where $MI(x)$, for $x = (x_1, \dots, x_d)$, denotes the multi-information, which is equal to the negative copula entropy, as shown by Ma & Sun (2011):

$$MI(X) := D_{KL}(p(x) \| \prod_j p_j(x_j)) = \int c_X(u(x)) \log c_X(u(x)) \mathrm{d}u = -h(c_X(u(x))). \tag{4}$$

Issues with lack of invariance to marginal transformations.

1. On the encoder side (Eq. (2)), the optimisation is performed over the parametric conditional margins $p_\phi(t_j|x)$ in $I_\phi(t;x) = \mathbb{E}_{p(x)} D_{KL}\left(p_\phi(t|x) \| p(t)\right)$. When a strictly increasing transformation $x_j \to \tilde{x}_j$ is applied, the required invariance property can only be guaranteed if the model for ϕ (in our case a deep network) is flexible enough to compensate for this transformation, which can be a severe problem in practice (see example in Section 4.1).

2. On the decoder side, assuming Gaussian margins in $p_\theta(y_j|t)$ might be inappropriate for modeling y if the domain of y is not equal to the reals, e.g. when y is defined only on a finite interval. If used in a generative way, the model might produce samples outside the domain of t. Even if other distributions than Gaussian are considered, such as truncated Gaussian, one still needs to make assumptions concerning the marginals. According to the IB formulation, such assumptions are spurious.

3. Also on the decoder side, we have: $I_\phi(t;y) = \mathbb{E}_{p(x,y)} \mathbb{E}_{p_\phi(t|x)} \log p_\theta(y|t) + h(y)$. The authors of Alemi et al. (2016) argue that since $h(y)$ is constant, it can be ignored in computing $I_\phi(t;y)$ and the variational bound thereon. This is true for a fixed discrete y, but not for the class of strictly increasing transformations of y, which should be the case for a model specified with mutual informations only. Since the left hand side of this equation is invariant against strictly increasing transformations, the first term on the right hand side cannot share this invariance property, because $h(y)$ also depends on strictly increasing transformations. In fact, under such transformations, the differential entropy $h(y)$ can take any value from $-\infty$ to $+\infty$, which can be seen easily by decomposing the entropy into the copula entropy and the sum of marginal entropies:

$$h(y) = h(c_y(u(y))) + \sum_j h(y_j) = -MI(y) + \sum_j h(y_j).$$

The first term (i.e. the copula entropy which is equal to the negative multi-information, as in Eq. (4)) is a non-positive number, and the marginal entropies can take any value when using strictly increasing transformations (for instance, the marginal entropy of a uniform variable on $[a,b]$ is $\log(b-a)$). As a consequence, the left-out term $h(y)$ in Eq. (3) can be treated as a constant only for one specific y, but not for all elements of the equivalence class containing all strictly increasing transformations of y, and every such transformation would lead to different $(I(x,t), I(y,t))$ pairs in the information curve, which basically makes this curve arbitrary. It should also be noted that even if one does not allow for increasing transformations of the marginals, the incorporation of the differential entropy term $h(y)$ to the variational bound is not justified beyond the discrete case.

3.3 Proposed Solution

The issues described in Section 3.2 can be fixed by using transformed variables ($x = (x_1, \dots, x_d)$):

$$\tilde{x}_j = \Phi^{-1}(\hat{F}(x_j)), \quad t_j = \hat{F}^{-1}(\Phi(\tilde{x}_j)), \tag{5}$$

where Φ is the Gaussian cdf and $\hat{F}$ is the empirical cdf. In the copula literature, these transformed variables are sometimes called *normal scores*. Note that the mapping is (approximately) invertible: $x_j = \hat{F}^{-1}(\Phi(\tilde{x}_j))$, with $\hat{F}^{-1}$ being the empirical quantiles treated as a function (e.g. by linear interpolation). This transformation fixes the invariance problem on the encoding side (issue 1), as well as the problems on the decoding side: problem 2 disappeared because the transformed variables $\tilde{x}_j$ are standard normal distributed, and problem 3 disappeared because the decoder part (Eq. (3)) now has the form:

$$\mathbb{E}_{p(x,y)} \mathbb{E}_{p_\phi(t|x)} \log p_\theta(\tilde{y}|t) = I_\phi(T;Y) + MI(Y) - \sum_j h(\tilde{Y}_j) = I_\phi(T;\tilde{Y}) - h(c_{\text{inv}}(u(\tilde{y})))$$

where $c_{\text{inv}}(u(\tilde{y}))$ is indeed constant for all strictly increasing transformations applied to y. Note that by applying a variational approximation $q_\theta(\tilde{y}|t)$ of $p_\theta(\tilde{y}|t)$, a variational bound on $I_\phi(t;\tilde{y})$ equal to $\mathbb{E}_{p(x,y)}\mathbb{E}_{p_\phi(t|x)} \log p_\theta(\tilde{y}|t) + h(c_{\text{inv}}(u(\tilde{y}))$ can be obtained.

Having solved the IB problem in the transformed space, we can go back to the original space by using the inverse transformation according to Eq. (5) $t_j = \hat{F}^{-1}(\Phi(\tilde{t}_j))$. The resulting model is thus a variational autoencoder with x replaced by $\tilde{x}$ in the first term and x replaced by $\tilde{y}$ in the second term.

Technical details. We assume a simple prior $p(t) = \mathcal{N}(t; 0, I)$. Therefore, the KL divergence $D_{KL}\left(p_\phi(t|x) \| p(t)\right)$ is a divergence between two Gaussian distributions and admits an analytical form. We then estimate

$$I(T;X) = \mathbb{E}_{p(x)} D_{KL}\left(p_\phi(t|x) \| p(t)\right) \approx \frac{1}{n} \sum_i D_{KL}\left(p_\phi(t|x_i) \| p(t)\right)$$

and all the gradients on (mini-)batches.

For the decoder side, $\mathbb{E}_{p(x,y)}\mathbb{E}_{p_\phi(t|x)} \log p_\theta(y|t)$ is needed. We train our model using the backpropagation algorithm. However, this algorithm can only handle deterministic nodes. In order to overcome this problem, we make use of the reparametrisation trick (Kingma & Welling, 2013; Rezende et al., 2014):

$$I(T;Y) = \mathbb{E}_{p(x,y)}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)} \sum_j \log p_\theta(y_j | t = \vec{\mu}_j(x) + diag(\sigma_j(x))\epsilon) + \text{const.},$$

with $\tilde{y}_j = \Phi^{-1}(\hat{F}(y_j))$.

Learning the copula transformation with neural networks. Note that the copula transformation defined by Eq. (5) requires ranking the data or learning the empirical cdf and then applying it to the data. The difficulty lies in the fact that this has to be performed for every dimension independently. Suppose one aims at learning the copula transformation implicitly with a neural network, which might be a part of a larger deep net. This task calls for a dedicated network architecture, which first separates the dimensions of the data and subsequently learns the marginal transformation for every dimension. The marginal transformation learning itself can be accomplished with an one-layer sub-network where the neurons model the bins of the domain of the dimension.

We have therefore identified two ways of performing the copula transformation: explicitly by applying Eq. (5) to the data, and implicitly by embedding the architecture described above in the deep net.

For continuous data, by applying the copula transformation directly to the data rather than implicitly in the network, we reduce the complexity of the network without restricting the generality of the model. The implicit approach of learning the copula also does not solve issue 3 from Section 3.2, since one cannot control for the final marginal transformation learned by the network. On the other hand, in the case of discrete or mixed data, one is often forced to randomly break ties among the data. In such case, the dedicated architecture implicitly implementing the copula makes it possible to embed the copula transformation in the end-to-end learning process and thus to implicitly use additional information for tie breaking.

3.4 Sparsity of the Latent Space

Sparse Gaussian Information Bottleneck. Recall that the information bottleneck compresses x to a new variable t by minimising $I(x;t) - \lambda I(t;y)$. This ensures that some amount of information with respect to a second "relevance" variable y is preserved in the compression.

The assumption that x and y are jointly Gaussian-distributed leads to the *Gaussian Information Bottleneck* (Chechik et al., 2005) where the solution t can be proved to also be

Gaussian distributed, i.e. for

$$(x, y) \sim \mathcal{N}\left(0, \begin{pmatrix} \Sigma_x & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_y \end{pmatrix}\right),$$

the optimal t is a noisy projection of x of the following form:

$$t = Ax + \xi, \quad \xi \sim \mathcal{N}(0, I) \quad \Rightarrow \quad t|x \sim \mathcal{N}(Ax, I), \quad t \sim \mathcal{N}(0, A\Sigma_x A^\top + I).$$

The mutual information between x and t is then equal to: $I(x;t) = \frac{1}{2}\log|A\Sigma_x A^\top + I|$.

In the *sparse* Gaussian Information Bottleneck, we additionally assume that A is diagonal, so that the compressed t is a sparse version of x. Intuitively, sparsity follows from the observation that for a pair of random variables x, x', any full-rank projection of x' would lead to the same mutual information since $I(x, x') = I(x; Ax')$, and a reduction in mutual information can only be achieved by a rank-deficient matrix A. For diagonal projections, this immediately implies sparsity of A.

Sparse latent space of the Deep Variational Information Bottleneck. We now proceed to explain the sparsity induced in the latent space of the *copula version* of the DVIB introduced in Section 3.3. To this end, we use the Sparse Gaussian Information Bottleneck model described above. We analyse the encoder part of the DVIB, described with $I(x, t)$. Consider the general Gaussian Information Bottleneck (with x and y jointly Gaussian and a full matrix A) and assume the following:

- a deterministic pre-transformation, $z = f_\beta(x)$, is performed on x. It is parametrised by a set of parameters β,
- the projection matrix A is implicitly represented as $\mu = Az$,
- optimisation of the mutual information $I(x, t)$ as in $\min I(x;t) - \lambda I(t;y)$ is performed over μ and β.

The estimator of $I(x;t) = \frac{1}{2}\log|A\Sigma_x A^t + I|$ then becomes:

$$\hat{I}(x;t) = \frac{1}{2}\log\left|\frac{1}{n}MM^\top + I\right| \tag{6}$$

where the matrix M contains n i.i.d. samples of μ, i.e. $M = AZ$ with $Z = (z_1, \ldots, z_n)^\top$ being a matrix of dimensions $n \times p$.

If the pre-transformation f_β were such that $\frac{1}{n}MM^\top =: D$ was diagonal, then (6) would simplify to

$$\hat{I}(x;t) = \frac{1}{2}\sum_i \log(D_{ii} + 1), \tag{7}$$

which is equivalent to the Sparse Gaussian Information Bottleneck model with a diagonal covariance matrix. We can, however, approximate this case by modifying (6), such that we only consider the diagonal part of the matrix,

$$I'(x;t) = \frac{1}{2}\log\left|\mathrm{diag}(\frac{1}{n}MM^\top + I)\right|. \tag{8}$$

Note that for any positive definite matrix B, the determinant $|B|$ is always upper bounded by $\prod_i B_{ii} = |\mathrm{diag}(B)|$, which is a consequence of Hadamard's inequality. Thus, instead of minimising $\hat{I}(x;t)$, we minimise an upper bound $I'(x;t) \geq \hat{I}(x;t)$ in the Information Bottleneck cost function. Equality is obtained if the transformation f_β, which we assume to be part of an "end-to-end" optimisation procedure, indeed successfully diagonalised $D = (\frac{1}{n}MM^\top + I)$. Note that equality in the Hadamard's inequality is equivalent to $D + I$ being orthogonal, thus f_β is forced to find the "most orthogonal" representation if the inputs in the latent space. Using a highly flexible f_β (for instance, modelled by a deep neural network), we might approximate this situation reasonably well. This explains how the copula transformation translates to a low-dimensional representation of the latent space.

We demonstrate the sparse structure of the learned latent space in Section 4.

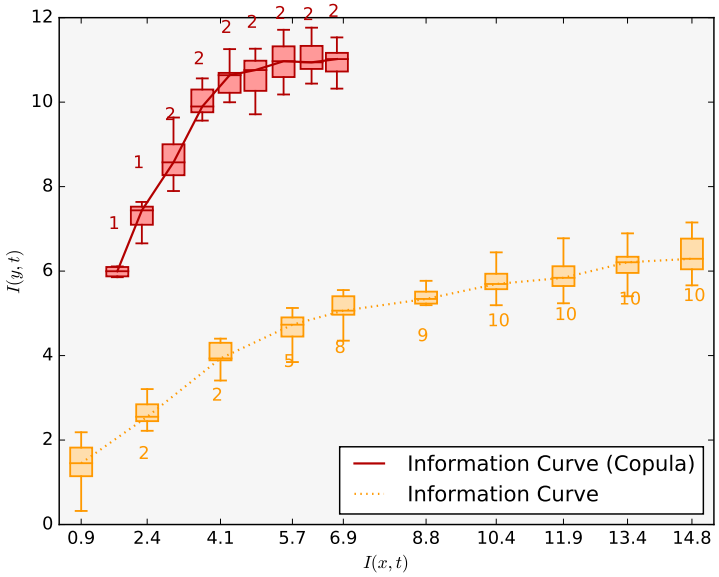


Figure 2: Information curves for the artificial experiment. The red curve describes the information curve with copula transformation whereas the orange one illustrates the plain information curve. The numbers represent the dimensions in the latent space t which are needed to reconstruct the output y.

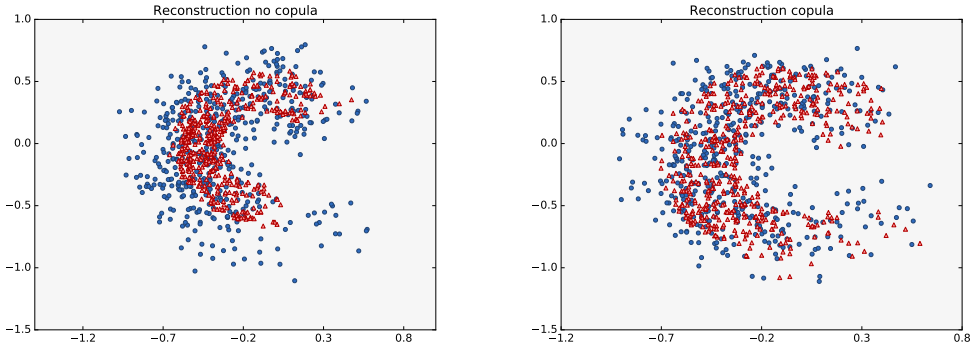
4 Experiments

4.1 Artificial Experiments

Dataset. In the artificial experiment, we compared the performance of the DVIB with its copula extension. For this reason, we generated an artificial dataset which consists of a ten-dimensional latent space t. This latent space t defined the angle and the radius of a spiral. In this experiment, both the angle and the radius were highly correlated. In order to reconstruct the details of the radial function, we had to use a higher dimensional latent space t. The random variable y was subsequently sampled from the spiral with additive noise whereas x was sampled from a uniform distribution. Both random variables x and y were further transformed to beta densities using strictly increasing transformations to show the relevance of our copula augmentation.

Test set-up. In our model, we used a latent layer with ten nodes that modelled the mean of the ten-dimensional latent space t. The variance of the latent space was set to 1 for simplicity. The stochastic encoder as well as the stochastic decoder consisted of a neural network with two fully-connected hidden layers with 50 nodes each. In addition, we used the softplus function as the activation function. Our model was trained using mini batches (size = 500) with the Adam optimiser (Kingma & Ba, 2014) for 70000 iterations using a learning rate of 0.0006. In addition, our decoder uses a Gaussian likelihood and we multiplied the λ parameter every 500 iterations by 1.06.

Results. In the first experiment, we compared the information curves between the DVIB and its copula augmentation (Figure 2). One can observe an increase in the mutual information from approximately 6 in the DVIB to approximately 11 in the copula DVIB. At the same time, only two dimensions were used in the latent space t by the copula DVIB. The version without copula did not provide competitive results despite using 10 out of 18 dimensions of the latent space t.



(a) (b)

Figure 3: Reconstruction of the output space y without the usage of the copula transformation (a) and with the copula transformation (b) and a corresponding λ value of 1480. Blue circles depict the output space Y and the red triangles — the conditionals means $\mu(y)$.

In the second experiment, we qualitatively elaborated the reconstruction capability of our model (Figure 3) with a λ value of 1480. This value was the last λ which was calculated during our training procedure. Figure 3(b) shows a detailed reconstruction of the output space y when using the copula transformation. On the other hand, the reconstruction quality of our model without the copula transformation resulted in a tight backbone which was not capable of reconstructing the whole output space y (Figure 3(a)).

4.2 Real Experiments

Dataset. As a real world experiment, we analysed the unnormalized *Communities and Crime* dataset Lyons et al. (1998) from the UCI repository[1], in order to demonstrate the significance of the proposed model. The dataset consisted of 125 predictive, 4 non-predictive and 18 target variables with 2215 samples in total. In a preprocessing step, we removed all missing values from the dataset. In the end, we used 1901 observations with 102 predictive and 18 target variables in our analysis. In particular, we looked at two cases: Disentanglement of the latent space based on the copula transformation (Experiment 1) and predictive variable compression with respect to the target variables (Experiment 2).

Test set-up. We used the same test set-up for Experiment 1 and 2. In our model, we used a latent layer with 18 nodes that modelled the mean of the 18-dimensional latent space t. For simplicity, the variance of the latent and the output space was set to 1. The stochastic encoder as well as the stochastic decoder consisted of a neural network with two fully-connected hidden layers with 100 nodes each. In addition, we used the softplus function as the activation function. The decoder used a Gaussian likelihood and λ was multiplied by 1.01 every 500 iterations. Our model was trained for 150000 iterations using mini batches with a size of 1255. As an optimiser, we used Adam (Kingma & Ba, 2014) and considered a learning rate of 0.0005.

Results. Experiment 1 (variable compression) shows compression of predictive variables with respect to the target variables. Analogously to the artificial experiment, Figure 4 compares the information curves for the model with and without copula transformation. Again, the information curve for the copula model yields larger values of mutual information for the same λ, which we attribute to the increased flexibility of the model which cannot be achieved by standard neural network architectures, as we pointed out in Section 3.3. In addition, the application of the copula transformation led to a much lower number of used dimensions in the latent space. For example, copula DVIB used only four dimensions in the

[1] http://archive.ics.uci.edu/ml/datasets/communities+and+crime+unnormalized

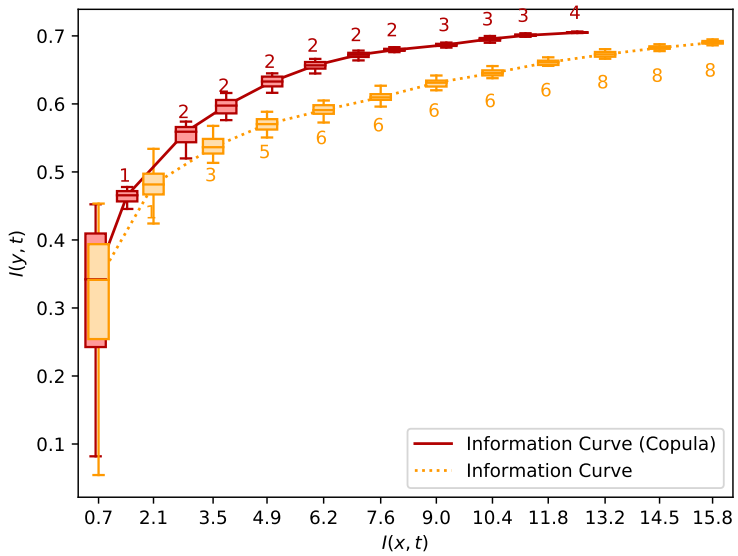
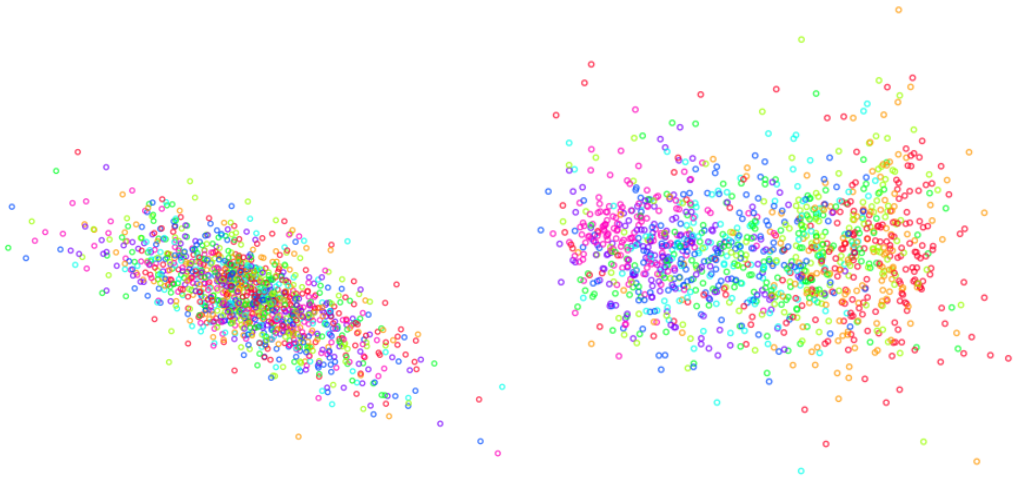


Figure 4: Information curves for the real data experiment. The red curve is the information curve with copula transformation whereas the orange one depicts the plain information curve. The numbers represent the dimensions in the latent space t which are needed to reconstruct the output y.



(a) (b)

Figure 5: Representation of the latent space t of two dimensions without (a) and with (b) the usage of the copula transformation.

latent space for the highest λ values. DVIB, however, needed eight dimensions in the latent space and nonetheless resulted in lower mutual information scores. In order to show that our information curves are significantly different, we performed a Kruskal-Wallis rank test (p-value of $1.6 * 10^{-16}$).

The second experiment (qualitative latent space analysis) illustrates the difference in the disentanglement of the latent spaces of the DVIB model with and without the copula transformation. We selected the target variable *arsons* and plotted it against the target

variable *larcenies*. In order to obtain the corresponding class labels (rainbow colours in Figure 5), we separated the values of *arsons* in eight equally-sized bins. We set λ to 22 in both models, which corresponded to the last calculated λ during our training procedure. The latent space t of DVIB, appears completely unstructured (Figure 5(a)). In contrast, we could identify a much clearer structure in the latent space with respect to our previously calculated class labels in Figure 5(b).

5 CONCLUSION

We have presented a novel approach to compact representation learning of deep latent variable models. To this end, we showed how restoring invariance properties of the Deep Variational Information Bottleneck with a copula transformation leads to disentanglement of the features in the latent space. Subsequently, we analysed how the copula transformation translates to sparsity in the latent space of the considered model. The proposed model allows for a simplified and fully non-parametric treatment of marginal distributions which has the advantage that it can be applied to distributions with arbitrary marginals. We evaluated our method on both artificial and real data.

References

A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy. Deep Variational Information Bottleneck. *ArXiv e-prints*, December 2016.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. *CoRR*, abs/1612.00410, 2016.

Jose M Alvarez and Mathieu Salzmann. Learning the number of neurons in deep networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 2270–2278. Curran Associates, Inc., 2016.

Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. *CoRR*, abs/1705.08841, 2017.

Matthew Chalk, Olivier Marre, and Gasper Tkacik. Relevant sparse codes with variational information bottleneck. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 1957–1965. Curran Associates, Inc., 2016.

Gal Chechik, Amir Globerson, Naftali Tishby, and Yair Weiss. Information bottleneck for gaussian variables. In *Journal of Machine Learning Research*, pp. 165–188, 2005.

Xi Chen, Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 2172–2180. Curran Associates, Inc., 2016.

Emily Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. *CoRR*, abs/1705.10915, 2017.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, Inc., 2014.

Shaobo Han, Xuejun Liao, David B Dunson, and Lawrence Carin. Variational gaussian copula inference. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, pp. 829–838, 2016.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arxiv*, 2016.

D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *ArXiv e-prints*, December 2014.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. *CoRR*, 2016.

M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings of the 3rd. International Conference on Face & Gesture Recognition*, FG '98, pp. 200–, Washington, DC, USA, 1998. IEEE Computer Society. ISBN 0-8186-8344-9. URL `http://dl.acm.org/citation.cfm?id=520809.796143`.

Jian Ma and Zengqi Sun. Mutual information is copula entropy. *Tsinghua Science & Technology*, 16(1):51–54, 2011.

Michael C. Mozer and Paul Smolensky. Advances in neural information processing systems 1. chapter Skeletonization: A Technique for Trimming the Fat from a Network via Relevance Assessment, pp. 107–115. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1989. ISBN 1-558-60015-9.

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. Regularizing neural networks by penalizing confident output distributions. *CoRR*, abs/1701.06548, 2017.

Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. Variational autoencoder for deep learning of images, labels and captions. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 2352–2360. Curran Associates, Inc., 2016.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pp. 1060–1069. JMLR.org, 2016. URL `http://dl.acm.org/citation.cfm?id=3045390.3045503`.

Melani Rey, Volker Roth, and Thomas Fuchs. Sparse meta-gaussian information bottleneck. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 910–918, Bejing, China, 22–24 Jun 2014. PMLR.

Mélanie Rey and Volker Roth. Meta-gaussian information bottleneck. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger (eds.), *NIPS*, pp. 1925–1933, 2012.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1278–1286, Bejing, China, 22–24 Jun 2014. PMLR.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

S. Suh and S. Choi. Gaussian Copula Variational Autoencoders for Mixed Data. *ArXiv e-prints*, April 2016.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. *CoRR*, abs/1503.02406, 2015. URL `http://arxiv.org/abs/1503.02406`.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Dustin Tran, David Blei, and Edo M Airoldi. Copula variational inference. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 28*, pp. 3564–3572. Curran Associates, Inc., 2015.

Weiran Wang, Honglak Lee, and Karen Livescu. Deep variational canonical correlation analysis. *CoRR*, abs/1610.03454, 2016. URL `http://arxiv.org/abs/1610.03454`.