Learning Sparse Latent Representations with the Deep Copula Information Bottleneck

Aleksander Wieczorek[*] **Mario Wieser**[*]**, Damian Murezzan, Volker Roth**
University of Basel, Switzerland

In this paper, we consider sparse latent space representation learning with the Deep Variational Information Bottleneck (DVIB) principle [1]. The DVIB combines the information bottleneck and the variational autoencoder methods. The *information bottleneck (IB)* [6] 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)* [3, 5] is a generative model which learns a latent representation t of x by using the variational approach. The solution t of the IB is identified with the latent space t of the VAE.

The DVIB model 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. Second, the latent space of the IB is not sparse which results in the fact that a compact feature representation is not feasible. We overcome these shortcomings by leveraging information theoretic properties of mutual information. Our contribution is therefore two-fold: we first restore the invariance properties of the information bottleneck solution in the DVIB. Subsequently, we show that the restored invariance properties allow us to exploit the sparse structure of the latent space of DVIB.

1 Model

1.1 Formulation of the DVIB

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

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

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

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

and

$$I_{\phi,\theta}(T;Y) = \mathbb{E}_{p(x,y)} \mathbb{E}_{p_\phi(t|x)} \log p_\theta(y|t) + h(Y). \tag{3}$$

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

1.2 Motivation: 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 monotone transformation

[*] equal contribution

Second workshop on Bayesian Deep Learning (NIPS 2017), Long Beach, CA, USA.

$x_j \to \tilde{x}_j$ 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.

2. On the decoder side, assuming Gaussian margins in $p_\theta(y_j|t)$ might be inappropriate for modelling 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.

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 [1] 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 $+\infty$.

1.3 Proposed Solution

The issues described in Section 1.2 can be fixed by using transformed variables ($x = (x_1, \dots, x_d)$):

$$\tilde{x}_j = \Phi^{-1}(\hat{F}(x_j)), \quad t_j = \hat{F}^{-1}(\Phi(\tilde{x}_j)), \tag{4}$$

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.

1.4 Sparsity of the Latent Space

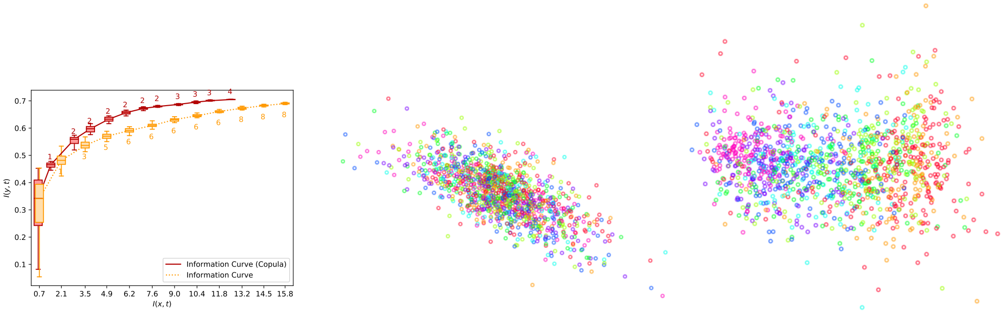
The assumption that x and y are jointly Gaussian-distributed leads to the *Gaussian Information Bottleneck* [2] where the solution t can be proved to also be Gaussian distributed, i.e. for $(x,y) \sim \mathcal{N}\left(0, \begin{pmatrix} \Sigma_x & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_y \end{pmatrix}\right)$, the optimal t is a noisy projection of x of the following form:$t = Ax + \xi$, $\xi \sim \mathcal{N}(0, I)$, $t|x \sim \mathcal{N}(Ax, I)$, $t \sim \mathcal{N}(0, A\Sigma_x A^\top + I)$. 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.

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 1.3. Consider the general Gaussian Information Bottleneck (with x and y jointly Gaussian and a full matrix A) and assume: a deterministic pre-transformation of x parametrised by β $z = f_\beta(x)$, an implicit representation of A $\mu = Az$, so that the optimisation of the mutual information $I(x,t)$ 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: $\hat{I}(x;t) = \frac{1}{2}\log\left|\frac{1}{n}MM^\top + I\right|$, where the matrix M contains n i.i.d. samples of μ, i.e. $M = AZ$ with $Z = (z_1, \dots, z_n)^\top$ being a matrix of dimensions $n \times p$. If the pre-transformation f_β were such that $D := \frac{1}{n}MM^\top$ were diagonal, then this would simplify to $\hat{I}(x;t) = \frac{1}{2}\sum_i \log(D_{ii} + 1)$, which is equivalent to the Sparse Gaussian Information Bottleneck model with a diagonal covariance matrix. We can, however, approximate this case by modifying $\hat{I}(x;t)$, such that we only consider the diagonal part of the matrix M, resulting in: $I'(x;t) = \frac{1}{2}\log\left|\text{diag}(\frac{1}{n}MM^\top + I)\right|$. Note that for any positive definite matrix B, the determinant $|B|$ is always upper bounded by $\prod_i B_{ii} = |\text{diag}(B)|$, 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 of 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.

2 Experiments

Dataset and Test set-up. We analysed the unnormalized *Communities and Crime* dataset [4] from the UCI repository[2], in order to demonstrate the significance of the proposed model. In our model, we used a latent layer with 18 nodes that modelled the mean of the 18-dimensional latent space t. The stochastic encoder as well as the stochastic decoder consisted of a neural network with two fully-connected hidden layers with 100 nodes each. 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.



(a) (b) (c)

Figure 1: (a) Information curves along with dimensionalities of the latent spaces. Representation of the latent space t of two dimensions without (b) and with (c) the copula transformation.

Results. First, we analysed the compression of predictive variables with respect to the target variables (Figure 1(a)). The information curve for the copula model yields larger values of mutual information for the same λ. In addition, the application of the copula transformation led to a much lower number of used dimensions in the latent space with higher mutual information scores. We subsequently performed a qualitative latent space analysis. Figures 1(c) and 1(b) illustrate 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*. The latent space t of DVIB, appears completely unstructured (Figure 1(b)) whereas we could identify a much clearer structure in the latent space for the copula version (Figure 1(c)).

3 Conclusion

We have presented a novel approach to compact representation learning of deep latent variable models. To this end, we showed that restoring invariance properties of mutual information in 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.

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

References

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

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

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

[4] 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, pages 200–, Washington, DC, USA, 1998. IEEE Computer Society.

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

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