Disentangling by Factorising

Hyunjik Kim[1,2], Andriy Mnih[1]
DeepMind[1], University of Oxford[2]
hyunjikk@google.com, amnih@google.com

Abstract

We introduce FactorVAE, a method that disentangles by encouraging the distribution of codes, the variational posterior averaged over the training set, to be factorial and hence independent in the dimensions. Furthermore, we propose a new measure of disentanglement that addresses some weaknesses of commonly used metrics.

1 Introduction

Learning interpretable representations of data that expose semantic meaning has important consequences for artificial intelligence. Such representations are useful not only for standard downstream tasks such as supervised learning and reinforcement learning, but also for tasks such as transfer learning and zero-shot inference where humans excel but machines struggle [13]. In particular, there have been multiple efforts in the deep learning community towards learning factors of variation in the data, commonly referred to as learning a *disentangled representation*. While there is no canonical definition for this term, we adopt the following definition: a representation where a change in one dimension corresponds to a change in one factor of variation, while being relatively invariant to change in other factors [3]. Moreover, we focus on image data in this work.

Using generative models has shown great promise in learning disentangled representations in images. Notably semi-supervised approaches that require implicit or explicit knowledge about the true underlying factors of the data have excelled at disentangling [12, 10, 22, 23]. However, ideally we would like to learn these in an unsupervised manner, due to the following reasons: 1. Humans are able to learn factors of variation unsupervised [21]. 2. Labels are costly as obtaining them requires a human in the loop. 3. Labels assigned by humans may be inconsistent and could also lead to omissions of factors that are imperceptible to the human eye.

The generative models used for unsupervised disentangling largely fall into two categories: the Variational Autoencoder (VAE) framework [11] and the Generative Adversarial Net (GAN) framework [7]. InfoGAN [5] is a notable example of the latter that learns disentangled representations by encouraging mutual information between the observations and a subset of latent variables. However it suffers from instabilities in training, and its disentangling performance is sensitive to the choice of the prior and the number of latents used [8]. The β-VAE [8] uses a VAE objective with extra penalty on the KL between the variational posterior and the prior, giving a more robust and stable method of disentangling.

One drawback of the β-VAE is that there is a strong dependency between disentanglement and reconstruction. The motivation for our work is to get a better trade-off between disentanglement and reconstruction, so as to improve the optimal disentanglement and also obtain sharper reconstructions. We achieve this goal by first analysing the source of this trade-off. We then propose FactorVAE, which modifies the objective accordingly, introducing a penalty that encourages the marginal distribution of representations to be factorial without hurting reconstruction too much. This new penalty is optimised using a discriminator network, following the divergence minimisation view of GANs [18, 20], and we show that our approach enhances disentanglement as well as reconstruction compared to the β-VAE. Moreover, to help quantify our improvements, we point out the weaknesses in existing metrics of disentanglement, and propose a new metric that addresses these shortcomings.

2 The Trade-off between Disentanglement and Reconstruction in β-VAE

We motivate our approach by analysing where the disentanglement and reconstruction trade-off arises in the β-VAE loss. First, we introduce notation and architecture of our VAE framework. We have observations $x^{(i)}, i = 1, \ldots, N$ in image space $\mathcal{X}$, and latents $z \in \mathbb{R}^D$ are real vectors interpreted as representations of the data. The generative model is defined by the standard Gaussian prior $p(z) = \mathcal{N}(0, I)$, and the decoder $p_\theta(x|z)$ parameterised by a DeconvNet with weights θ. The variational posterior is given by the encoder $q_\phi(z|x) = \prod_{j=1}^D \mathcal{N}(z_j|\mu_j(x), \sigma_j^2(x))$, parameterised by a ConvNet with weights ϕ. An important distribution for our analysis is the marginal posterior of VAEs, namely the marginal distribution of the latents/code (used interchangeably):

$$r(z) = \int p_{data}(x)q(z|x)dx = \frac{1}{N}\sum_{i=1}^{N} q(z|x^{(i)}) \tag{1}$$

We can easily sample from r by ancestral sampling: $x \sim p_{data}, z \sim q(\cdot|x)$. r is relevant for when we are looking for a disentangled representation; should the representations correspond to the independent factors of variation in the data, we would like the distribution of these representations to be factorised, i.e. independent in the dimensions: $r(z) = \prod_{j=1}^D r(z_j)$.

The β-VAE objective is as follows:

$$\sum_{i=1}^{N} \mathbb{E}_{q(z|x^{(i)})}[\log p(x^{(i)}|z)] - \beta KL(q(z|x^{(i)})||p(z)) \tag{2}$$

Note this is a variational lower bound for $\beta \geq 1$. The first term is a measure of reconstruction, and the second term is the complexity penalty that acts as a regulariser. We may break down this KL term further as follows [9, 14]:

$$KL(q(z|x^{(i)})||p(z)) = I(x;z) + KL(r(z)||p(z)) \tag{3}$$

where $I(x;z)$ is the mutual information (MI) between x and z under the joint distribution of the data and their codes $p_{data}(x)q(z|x)$. The $KL(r(z)||p(z))$ term encourages independence in the dimensions of z and hence disentanglement by pushing $r(z)$ towards $p(z)$, a factorised distribution. On the other hand, penalising the MI term $I(x;z)$ acts as an information bottleneck between x and z whose presence is necessary for generalisation, but penalising this term too heavily (high β) leads to a lack of information about x in z and hence poor reconstruction [14]. Hence initially raising β from 1, and thus further penalising both terms, leads to better disentanglement while sacrificing reconstruction. When this sacrifice is severe, there is insufficient information about the observation in the latents, hurting disentanglement as well. So there exists an optimal value of $\beta > 1$ that gives highest disentanglement, whose reconstructions are blurrier than a VAE ($\beta = 1$).

3 The Total Correlation penalty and FactorVAE

We motivate FactorVAE with the suspicion that the further penalty on the MI might be unnecessary for improved disentanglement. So instead, we keep the VAE objective and directly encourage independence in the code distribution, arriving at our new objective:

$$\sum_{i=1}^{N} \mathbb{E}_{q(z|x^{(i)})}[\log p(x^{(i)}|z)] - KL(q(z|x^{(i)})||p(z)) - \gamma KL(r(z)|| \prod_{j=1}^{D} r(z_j)) \tag{4}$$

The latter term is also known as *Total Correlation* (TC) [26], often used as a measure of independence for multiple random variables. However this term is intractable, so we need further machinery to optimise it. To do so, first note that we can easily sample from $r(z)$ using ancestral sampling described above. Moreover we can sample from $\prod_j r(z_j)$ by sampling D times from $r(z)$ then ignoring all but one dimension for each sample, or more efficiently by sampling a batch from $r(z)$ then randomly permuting across the batches for each dimension. As long as the batch is large enough, these samples will be close to sampling from $\prod_j r(z_j)$. This is a standard trick used in the independence testing literature [1, 6]. Having access to samples from both distributions allows us to minimise their KL divergence using a discriminator to approximate the density ratio that arises in the KL [19, 25]. That is to say, suppose we have a discriminator D_ψ, an MLP with weights ψ, that outputs a probability given input z. Suppose it approximates the probability that z is a sample from $r(z)$ over $\prod_j r(z_j)$. Then we have:

$$TC(z) = KL(r(z)||\prod_{j=1}^{D} r(z_j)) = \mathbb{E}_{r(z)}\left[\log \frac{r(z)}{\prod_j r(z_j)}\right] \approx \mathbb{E}_{r(z)}\left[\log \frac{D(z)}{1 - D(z)}\right] \tag{5}$$

So for FactorVAE we train the discriminator and the VAE by simultaneous gradient descent. In particular, the VAE parameters θ, ϕ are updated using the loss in Equation 4, but replacing the TC term by the right hand side of Equation 5. The discriminator is trained using samples from $r(z)$ and $\prod_j r(z_j)$ to approximate their density ratio and hence the TC.

Note that in the usual GAN literature, the divergence minimisation occurs between two distributions over the data space, which is often very high dimensional (e.g. images). So the two distributions often have disjoint support, which makes training unstable especially when the discriminator is strong. Hence it is necessary to use tricks such as using sparse discriminator updates, instance noise [24] or getting rid of the discriminator altogether as for Wasserstein-divergence [2]. For our work, we are minimising divergence between two distributions over the latent space (as in e.g. [17]), which is usually much lower dimensional and the two distributions have overlapping support. We observe that training is stable for large enough batch sizes, allowing us to use a strong discriminator with frequent updates.

4 A New metric for Disentanglement

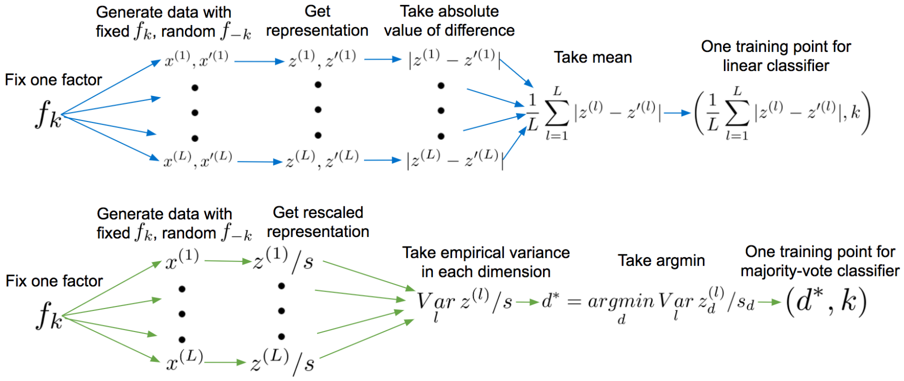
Generate data with fixed f_k, random f_{-k}
Get representation
Take absolute value of difference
Take mean
One training point for linear classifier
Fix one factor
f_k
$x^{(1)}, x'^{(1)}$
$z^{(1)}, z'^{(1)}$
$|z^{(1)} - z'^{(1)}|$
$x^{(L)}, x'^{(L)}$
$z^{(L)}, z'^{(L)}$
$|z^{(L)} - z'^{(L)}|$
$\frac{1}{L}\sum_{l=1}^{L}|z^{(l)} - z'^{(l)}|$
$\left(\frac{1}{L}\sum_{l=1}^{L}|z^{(l)} - z'^{(l)}|, k\right)$
Generate data with fixed f_k, random f_{-k}
Get rescaled representation
Take empirical variance in each dimension
Take argmin
One training point for majority-vote classifier
Fix one factor
f_k
$x^{(1)}$
$z^{(1)}/s$
$x^{(L)}$
$z^{(L)}/s$
$\underset{l}{Var}\, z^{(l)}/s$
$d^* = \underset{d}{argmin}\, \underset{l}{Var}\, z_d^{(l)}/s_d$
(d^*, k)

Figure 1: Top: Metric in [8]. Bottom: Our new metric, where $s \in \mathbb{R}^d$ is the scale (empirical standard deviation) of latent representations of the full data (or big enough random subset)

Returning to our definition of disentanglement, where a change in one dimension of the representation corresponds to a change in precisely one factor of variation, we point out that this definition is quite crude. We have implicitly ignored the possibility of: correlations among the factors, hierarchy in the factors of variation, and a many-to-one mapping between a combination of factors and a data point (over-representation). Thus our definition is limited to synthetic data with independent factors of variation. However, as we'll show in the paper, robust disentanglement is not a fully solved problem even in this setting. Part of the obstacle in achieving this first milestone lies in the absence of a sound, quantitative metric for measuring disentanglement. We point out the weaknesses in existing methods of assessing disentanglement, and introduce a new metric that addresses these problems.

A popular method of measuring disentanglement is by inspecting latent traversals: visualising the change in reconstructions as one traverses across each dimension of the latent space. The qualitative nature of this approach makes it unsuitable for comparing different algorithms. Moreover we must look at multiple latent traversals for a robust assessment of an algorithm, namely using multiple reference images, random seeds, and points during training. Having a human in the loop to assess the traversals is too time consuming and the evaluations are not reproducible.

The authors of β-VAE proposed a supervised metric that attempts to quantify disentanglement [8]. The metric is the error rate of a linear classifier that is trained as follows. Choose a factor k; generate data with this factor fixed but all other factors varying randomly; obtain their representations; take the absolute value of the pairwise differences of these representations. Then the mean of these statistics across the pairs gives one training input for the classifier, and the fixed factor k is the corresponding training output (see top of Figure 1). So if the representations were perfectly disentangled, we would see zeroes in the dimension of the training input that corresponds to the fixed factor of variation, and the linear classifier will learn to map the index of the zero value to the factor.

However this metric has several weaknesses. Firstly, the metric could be sensitive to hyperparameters of the linear classifier optimisation (optimiser, weight initialisation, number of training iterations) hence these need to be tuned. Secondly, having a linear classifier is not so intuitive - we could get representations where each factor corresponds to a linear combination of dimensions instead of a single dimension. Finally and most importantly, the metric has a failure mode where it gives 100% accuracy when it only disentangles $K-1$ factors out of K; for the remaining factor, the classifier can cheat by detecting when all dimensions for the $K-1$ factors are non-zero. An example of such a case is displayed in Figure 2.

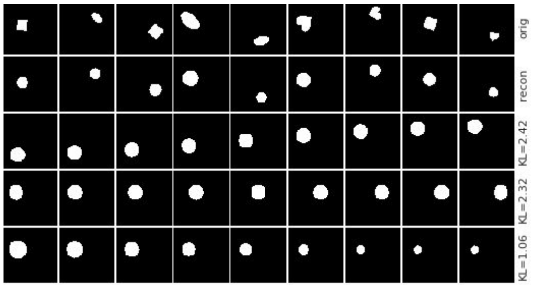

Figure 2: A model trained on the 2D Shapes data that scores 100% on metric in [8] (ignoring the shape factor). First row: originals. Second row: reconstructions. Remaining rows: reconstructions of latent traversals. The model captures x-pos,y-pos and scale but ignores orientation, yet achieves a perfect score on the metric.

So we propose an enhanced disentanglement metric as follows. Choose a factor k; generate data with this factor fixed but all other factors varying randomly; obtain their representations; rescale each dimension by its empirical standard deviation of representations over the full data (or a large enough random subset); take the empirical variance in each dimension. Then the index of the dimension with lowest variance gives one training input with training output k for a classifier. So if the representation is perfectly disentangled, the empirical variance in the dimension corresponding to the fixed factor will be 0. We rescale the representations prior to taking the argmin, so that the argmin is invariant to rescaling of the representations in each dimension. Since both inputs and outputs lie on a discrete space, the optimal classifier is the majority-vote classifier, and the metric is the error rate of the classifier. Here the classifier is a deterministic function of the training data, hence there are no optimisation hyperparameters to tune, and we claim that the metric is conceptually simpler and more intuitive than the previous metric. Most importantly it circumvents the failure mode in the latter, since the classifier needs to see the lowest variance in a latent dimension for a given factor to classify it correctly (see bottom of Figure 1).

5 Related Work

Using a discriminator to optimise a divergence encouraging independendence has been explored in a couple of recent works. The Adversarial Autoencoder (AAE) [15] removes the MI term in the VAE objective, to optimise the reconstruction error plus $KL(r(z)||p(z))$ using the density ratio trick. However they explore the AAE for semi-supervised classification or unsupervised clustering, not so much in the context of disentangling. In PixelGAN Autoencoders [14] that use the same objective, it is claimed that adding extra additive noise to the inputs of the encoder is crucial, which could be an indication that an information bottleneck is necessary and that the MI term shouldn't be dropped. Brakel et al [4] also use a discriminator to minimise the Jensen-Shannon Divergence between the distribution of codes and the product of its marginals, but use the GAN framework with deterministic encoders and decoders, and only explore their technique in the context of Independent Component Analysis source separation.

6 Preliminary Experiments

We show results of experiments on the 2D Shapes dataset [16], which are binary 64 by 64 images generated from five factors of variation: shape, x-position, y-position, scale and orientation. The encoder is a 4-layer ConvNet and the decoder is a deConvNet with the same architecture (same as in [8]). The discriminator is a 6-layer MLP with 1000 units per layer, and use five discriminator updates per VAE update (smaller MLPs and fewer discriminator updates work fine, but we noticed slight improvements up to this setting).

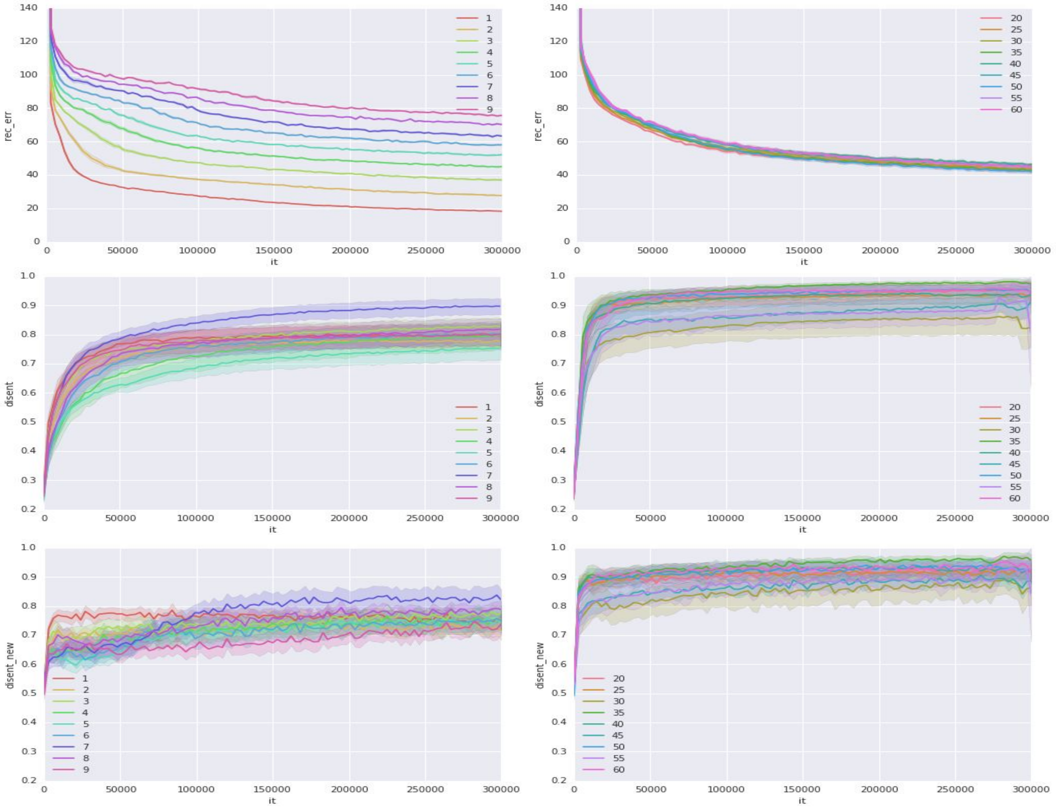

Figure 3: Reconstruction error (top), metric in [8] (middle), our metric (bottom). β-VAE (left), FactorVAE (right). The colours correspond to different values of β and γ respectively, and confidence intervals are over 10 random seeds.

From Figure 3, we see that FactorVAE gives much improved disentanglement compared to β-VAE for both metrics[1], and that we can do so without sacrificing reconstruction error too much. Note that the reconstruction error for the best disentanglement of β-VAE ($\beta = 7$) is over 60, which is significantly higher than that for FactorVAE ($\gamma = 35$), around 40. Also comparing the β-VAE scores on the two metrics, we can see that the metric in [8] is inflated compared to our new metric, due to the failure mode described in Section 4 (can be seen from visual inspection). Our method, on the contrary, robustly disentangles the four continuous factors, hence the two metrics are similar.

7 Future Work

To ensure that the discriminator is giving us the correct gradients, we wish to investigate the error in the discriminator's approximation of TC and analyse how it evolves during training, and how it is affected by the architecture of the discriminator and the frequency of discriminator updates. We will also carry out experiments for more complex data sets, and provide comparisons with InfoGAN.

[1] Both metrics ignore the shape factor for now, since neither the β-VAE nor our method can successfully model discrete factors of variation. This would require using discrete latent variables instead of Gaussians, but jointly modelling discrete and continuous factors of variation is a non-trivial problem that needs further research.

References

[1] Miguel A Arcones and Evarist Gine. On the bootstrap of u and v statistics. *The Annals of Statistics*, pages 655–674, 1992.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223, 2017.

[3] 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.

[4] Philemon Brakel and Yoshua Bengio. Learning independent features with adversarial nets for non-linear ica. *arXiv preprint arXiv:1710.05050*, 2017.

[5] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, pages 2172–2180, 2016.

[6] Gary Doran, Krikamol Muandet, Kun Zhang, and Bernhard Schölkopf. A permutation-based kernel conditional independence test. In *UAI*, pages 132–141, 2014.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[8] 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. 2016.

[9] 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.

[10] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *NIPS*, 2014.

[11] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. 2014.

[12] Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *NIPS*, 2015.

[13] 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*, pages 1–101, 2016.

[14] Alireza Makhzani and Brendan Frey. Pixelgan autoencoders. 2017.

[15] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

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

[17] L. Mescheder, S. Nowozin, and A. Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *ICML*, 2017.

[18] Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.

[19] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 2010.

[20] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *NIPS*, pages 271–279, 2016.

[21] G Perry, ET Rolls, and SM Stringer. Continuous transformation learning of translation invariant representations. *Experimental brain research*, 204(2):255–270, 2010.

[22] Scott Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee. Learning to disentangle factors of variation with manifold interaction. In *ICML*, pages 1431–1439, 2014.

[23] N Siddharth, Brooks Paige, Van de Meent, Alban Desmaison, Frank Wood, Noah D Goodman, Pushmeet Kohli, and Philip HS Torr. Learning disentangled representations with semi-supervised deep generative models. In *NIPS*, 2017.

[24] Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. Amortised map inference for image super-resolution. In *ICLR*, 2016.

[25] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64(5):1009–1044, 2012.

[26] Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of research and development*, 4(1):66–82, 1960.