Estimating the true probability distribution of categorical data with dependent variables using Variational Autoencoders

Ceren Dikmen, Jakob Heyder, Lutfi Altin, Muhammad Fasih Ullah
KTH
Stockholm, Sweden
cerend|heyder|lutfia|mufu@kth.se

***Abstract*—Recent advancements on Generative Networks has improved the accuracy of generating natural images. Focusing on VAEs, they have challenges to consider and various studies proposed different methods (Beta-VAE, InfoVAE) to mitigate the problems. These methods have been tested on image datasets. We propose to evaluate performance of recent advancements in VAE for estimating the true distribution of categorical data.**

Keywords—Probability distribution estimation, Categorical data, Variational Autoencoders

I. INTRODUCTION

Generative Networks have shown much success in recent years to estimate the true probability distribution and generate new samples of images. One most recent example is the Generative Adversarial Network (GAN) based algorithm published by NVIDIA with astonishing realistic generated samples of humans, cats of objects.[1]

Most studies focus on images, as they are very high dimensional and intuitive to validate for humans. Further categorical or discrete data can be problematic as the cost function is not simply differentiable, thus needs tricks such as the Gumbel-softmax trick[2][3] or one-hot encoding.

The most prominent approaches to Generative Networks are GANs and Variational Autoencoders(VAEs). Both have different challenges to consider, GANs are prone to mode-drop[4] and lack a standardized, objective evaluation measure[5][6] while VAEs often suffer from disentangled[7] and uninformed latent codes[8] and it is unclear how well they model the true distribution. This is due to the loss function which assumes a standard multivariate Gaussian as true prior and constrains the latent space to it by increasing the cost for information encoded that varies from a zero mean and one standard deviation per dimension in the latent space. A recent study proposed the InfoVAE family[8], which uses the maximum mean discrepancy (MMD) as probability-distribution measure instead of the KL-Divergence, which outperformed the other approaches on all their metrics. It furthermore solved the problem of disentangled and uninformative latent codes by allowing more variation.

In this study we investigate if VAEs can estimate the true distribution of a complex synthetic categorical data distribution with dependent variables and if simple Monte Carlo Simulations could be used to approximate the quality of the model. We will compare the KL-Divergence based cost functions with the MMD-VAEs with varying beta values.

We evaluate their approximation of the true distribution underlying the synthetic data and how correlated the estimate through a simple Monte Carlo simulation is with their distribution approximation.

II. METHOD

In this study we used synthetical data generated from a complex probability distribution. This has the advantage to make it easier to validate the correctness of approximating the underlying distribution. The distribution is generated by joining different marginal distributions equally weighted. Each distribution represented one feature *f* with a number of *n* possible categories. To account for dependent variables some marginal distributions were multidimensional over two and three random variables and allowed only to generate categories we labelled “even” together and categories we labelled “odd” together, but never mixed ones. This constrained was validated later on generated samples, as approximator how often this implicit rules in the data would be broken.

Furthermore we used the mean squared error (MSE) or negative log likelihood (NLL) as reconstruction loss and the KL-Divergence or MMD as divergence measure. The loss function was composed of the reconstruction loss and a beta weighted divergence measure. The performance of the model was tested comparing the generated distribution of samples to the original and applying a simple Monte Carlo simulation to compare their performance and reliability on an imaginative simulation task with the original data. The baseline to compare with is made by calculating the distance from a separate test set that is generated from the true prior distribution to the original training set. This can be thought of as the best possible way to get new samples, assuming knowing the underlying distribution.

The setup of the networks and computational flow is depicted in Figure 1.

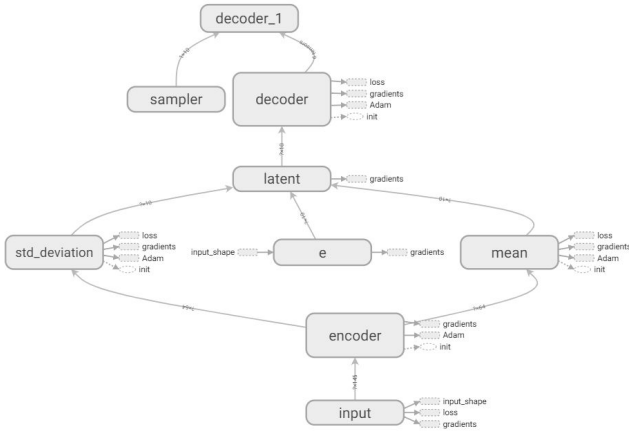

Fig. 1. Tensorboard shows the computational flow and structure of the neural networks and complete setup of the experiments.

III. RESULTS

In our experiments we used different hyperparameters such as beta values, batch size, number of iterations or divergence measure. The experiments were constrained by the hardware memory limitations given by a public access to Google Collab Jupyter Notebooks. We measured the mean of the score-distance between the Monte Carlo simulation with the original data and generated data based on the trained model, as well as the number of invalid instances, the total loss used for training and the MMD-Distance between the two distributions.

In Figure 2a and 2b we can see the mean score-distance and MMD-Distance for the original data compared to generated data with MMD as metric for divergence during training, 50.000 iterations and varying beta values. We do not see significant improvement with higher beta values, but rather a strong fluctuation.

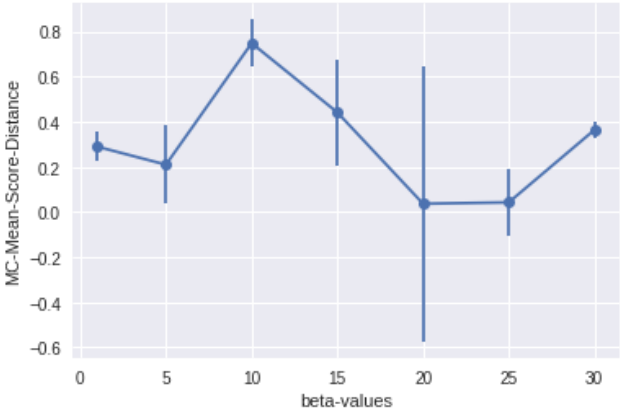

Fig. 2a. Mean of the Score-Distance between original and generated instances after 50.000 iterations with MMD-Divergence for different betas.

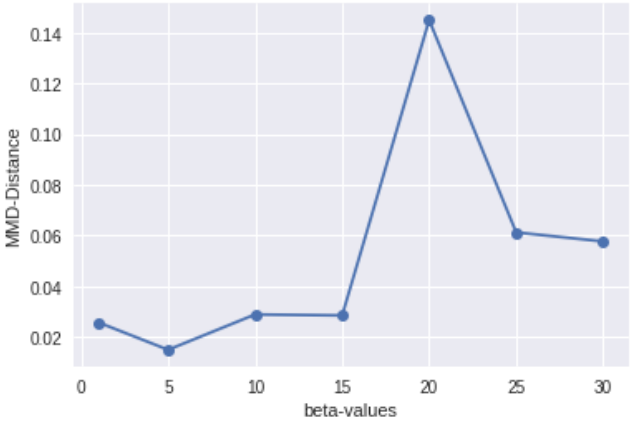

Fig. 2b. MMD-Distance between original and generated instances after 50.000 iterations with MMD-Divergence for different betas.

For both divergence metrics, MMD and ELBO we can see that increasing beta, which will weight the loss for distribution divergence higher, will increase the total loss and variation between the iterations. More iterations are needed with higher beta values. This correlation of decreasing variance and MMD-Distance with increased iterations and reduced total loss is depicted in Figures 3a-c with ELBO as metric for divergence. It takes longer for the network to reduce the divergence loss and infer the underlying distribution under the given constraint of fitting the latent space a standard multivariate Gaussian prior than learning a correct reconstruction through the reconstruction error.

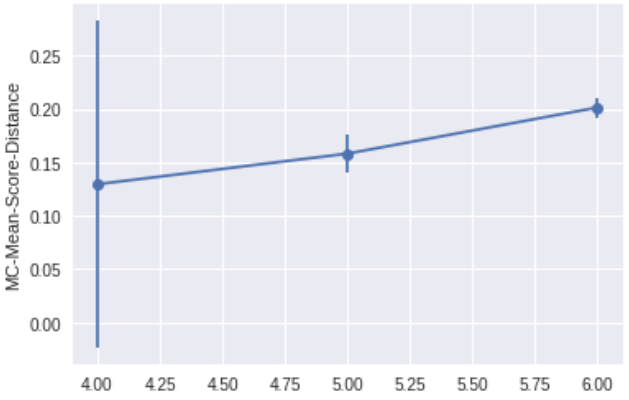

Fig. 3a. Mean score-distance between original and generated data with β=25 and different number of training steps: 10^4, 10^5, 10^6.

In Figure 3c we can see the total loss for a beta value of 25 over one million iterations. The graph shows a steady improvement by reducing the loss. At the same time the MMD-distance is continuously decreasing as more iterations the model is trained on. It shows that the model is successfully learning to fit the underlying distribution and thus decreasing the variance and MMD-distance.

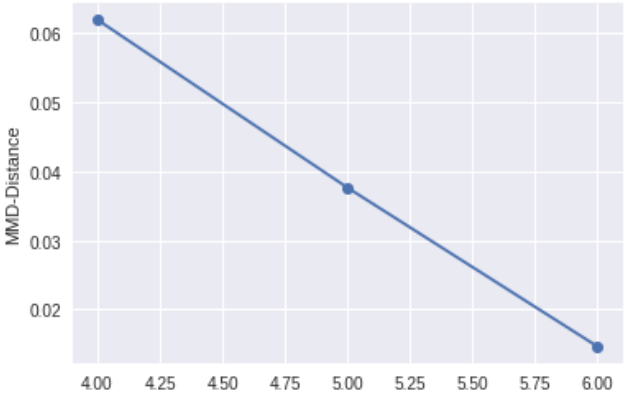

Fig. 3b. MMD distance between original and generated data with β=25 and different number of training steps: 10^4, 10^5, 10^6.

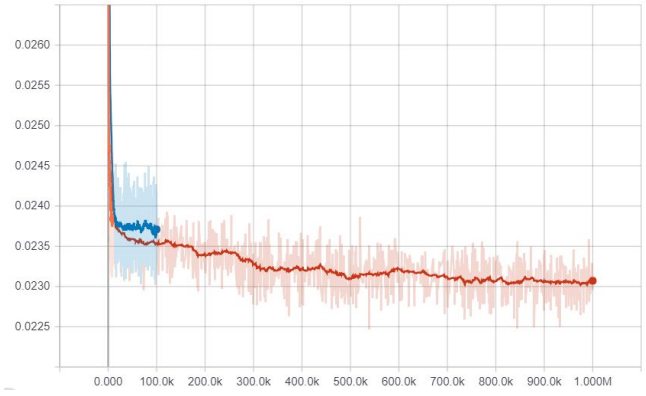

Fig. 3c. Total loss value evolution with β=25 and different number of training steps: 10^4, 10^5, 10^6.

For the comparison with the baseline, the number of iterations is kept at a reasonable level and we fixed the beta values due to restricted resources.. As depicted in Figure 4a the approach using ELBO-Divergence is archiving MC-Score distances in the area of 0.07-0.4 while the optimal baseline from a separately generated test-sets through the underlying distribution is in the range of 0.004-0.02. For some use cases this can be considered a reasonably well estimation of the data, but still needs to be further improved to reach an undistinguishable level to the original distribution baseline.

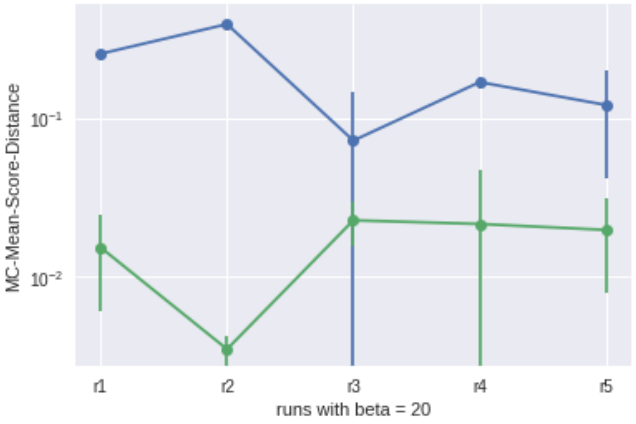

Fig. 4a. Mean score-distance between original and generated data (blue) with β=25 with 20.000 iterations and the baseline distance in green. The y-axis is in log-scale and using ELBO-Divergence.

The MMD-Distance is non-significant for the baseline test-sets and compared to the MMD-Distance of the generated data by two magnitudes improved as shown in Figure 4b on a log-scale graph. Figure 4c shows that the loss seems minimized after the number of iterations and equals out without further improvement. The networks still vary in their performance, this is also due to different measurements of model performance and reconstruction loss. Former is the MC-Score distance and latter is the MSE based on the reconstructed training sample.

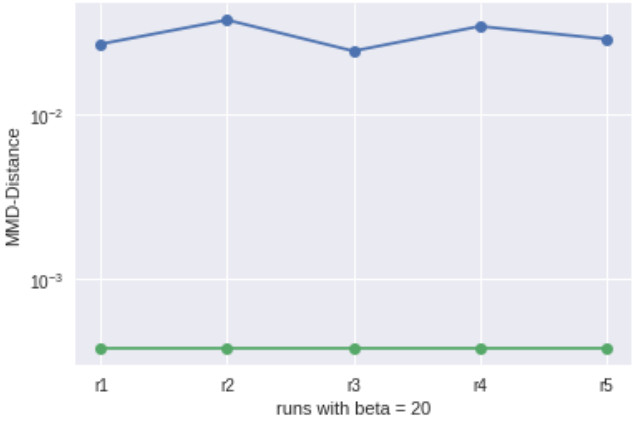

Fig. 4b. MMD distance between original and generated data with β=25 and different number after 20.000 iterations compared to the baseline (green). The y-axis is in log-scale and using ELBO-Divergence.

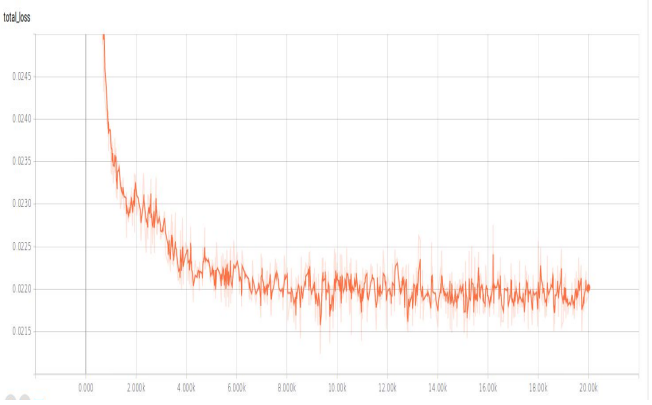

Fig. 4c Total loss with β=25 and 20.000 iterations for ELBO-Divergence.

Figures 5a-b plot MMD-distance with increased number of iterations, beta values of 25 and MMD as divergence metric. Comparing figures 3 and 5 we conclude that MMD as divergence metric performs worse compared to ELBO for categorical data used in Monte Carlo simulations.

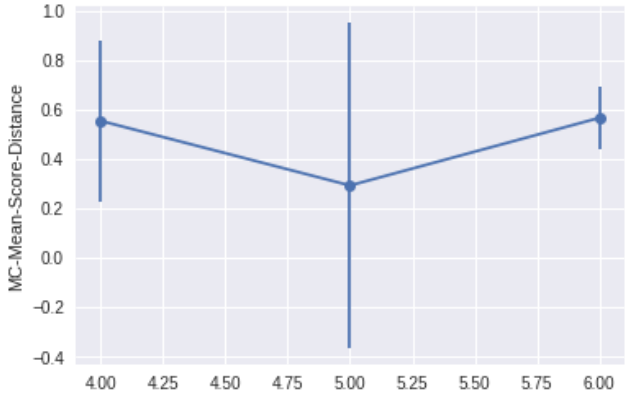

Fig. 5a. Mean score-distance between original and generated data with β=25 and different number of training steps: 10^4, 10^5, 10^6.

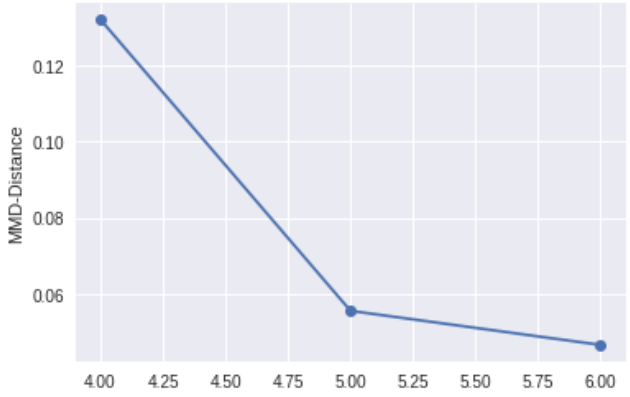

Fig. 5b. MMD distance between original and generated data with β=25 and different number of training steps: 10^4, 10^5, 10^6.

IV. DISCUSSION

In this study we investigated if VAEs can estimate a complex underlying distribution of categorical data, including dependent variables. In our results we conclude that while ELBO performs reasonably well compared to the optimal baseline, it still needs further improvement to generate indistinguishable samples to the originals. However, even with the constraint of a simple multivariate Gaussian as prior it is possible with sufficient hyperparameters to learn the underlying distribution and generate instances with less than one percent of invalid generated samples. This can not simply be generalized to even more complex distributions and will need further investigation. Other papers which investigated application for complex categorical data were mostly based on different prior distribution assumptions such as InfoCatVAE[9].

Higher beta values did not result necessarily in a loss of reconstruction and an improvement in divergence as proposed by BetaVAE[7] through disentangled latent-codes. This could mainly be due to limited computational resources, as it needs sufficient batch-sizes and iterations to learn the distribution and increased beta resulted also in much increased variance.

It remains an open question why the Maximum Mean Discrepancy (MMD) as divergence metric from the InfoVAE-family[8] does not result in an improvement compared to the original ELBO approach. In their paper it resolves the issue of entangled and uninformative latent codes by loosening the constraint on the latent space distribution. This is done by using the comparison of all Moments of two distributions as similarity measure instead of their samples likelihood as when using the KL-Divergence.

Future works could investigate this further and explore the latent code distribution, how informative it is and if the codes can be further disentangled through hyperparameter tuning.

V. REFERENCES

[1] Tero Karras and (2018). A Style-Based Generator Architecture for Generative Adversarial Networks. *CoRR, abs/1812.04948*.

[2] Jang, E., Gu, S., & Poole, B. (2015). Categorical Reparameterization with Gumbel-Softmax. Retrieved from https://arxiv.org/abs/1611.01144

[3] Maddison, C., Mnih, A., & Teh, Y. (2016). The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. Retrieved from https://arxiv.org/abs/1611.00712

[4] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., & Ozair, S. et al. (2014). Generative Adversarial Nets. Retrieved from https://papers.nips.cc/paper/5423-generative-adversarial-nets

[5] Borji, A. (2018). Pros and Cons of GAN Evaluation Measures. Retrieved from https://arxiv.org/abs/1802.03446

[6] Lucic, M., Kurach, K., Michalski, M., Gelly, S., & Bousquet, O. (2017). Are GANs Created Equal? A Large-Scale Study. Retrieved from https://arxiv.org/abs/1711.10337

[7] Burgess, C., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., & Lerchner, A. (2018). Understanding disentangling in BetaVAE. Retrieved from https://arxiv.org/abs/1804.03599

[8] Zhao, S., Song, J., & Ermon, S. (2017). InfoVAE: Information Maximizing Variational Autoencoders. Retrieved from https://arxiv.org/abs/1706.02262

[9] Pineau, E., & Lelarge, M. (2018). InfoCatVAE: Representation Learning with Categorical Variational Autoencoders. Retrieved from https://arxiv.org/abs/1806.08240