Schedae Informaticae Vol. 27 (2018): 69–79
doi: 10.4467/20838476SI.18.006.10411

Sliced Generative Models

Szymon Knop, Marcin Mazur, Jacek Tabor,
Igor Podolak, Przemysław Spurek
Faculty of Mathematics and Computer Science
Jagiellonian University, Łojasiewicza 6, 30-348 Kraków, Poland
e-mail: *szymon.knop@doctoral.uj.edu.pl*

Abstract. In this paper we discuss a class of AutoEncoder based generative models based on one dimensional sliced approach. The idea is based on the reduction of the discrimination between samples to one-dimensional case. Our experiments show that methods can be divided into two groups. First consists of methods which are a modification of standard normality tests, while the second is based on classical distances between samples. It turns out that both groups are correct generative models, but the second one gives a slightly faster decrease rate of Fréchet Inception Distance (FID).

Keywords: Generative model, AutoEncoder, Wasserstein distances

1. Introduction

In recent years a number of generative models based on AutoEncoder architecture were constructed (see, e.g., [5, 6, 10, 11]). Some of them have applied elegant geometric properties of the optimal transport (OT) problem and the Wasserstein distances. An important example is given in [6], where the authors construct Sliced-Wasserstein AutoEncoder (SWAE) – a generative model that performs well without the need for training an adversarial network but, on the other hand, with necessity of sampling from the prior distribution $P_{\mathcal{Z}}$ on the latent $\mathcal{Z}$. Specifically, the method applied there uses the sliced Wasserstein distance between the distribution of encoded training samples (z_i) and $P_{\mathcal{Z}}$ [6]. SWAE has an efficient numerical solution that provides similar capabilities to Wasserstein AutoEncoders (WAE-MMD) [11] and Variational AutoEncoders [5]. A typical choice for $P_{\mathcal{Z}}$ is the Gaussian distribution $N(0, 1)$ even though SWAE is valid for any prior distribution. In this case, there is no need to sample from $P_{\mathcal{Z}}$, as long as we can analytically calculate a closed formula for the distance between a given sample (x_i) and $N(0,1)$.

In our paper we follow the idea of [6] and make a comparison of few AutoEncoder based generative models, for which the loss functions are given by appropriately chosen sliced distanced between (z_i) and $N(0,1)$ that can be expressed in a closed form. Specifically, we use respective one-dimensional "measures of normality", including the 2nd Wasserstein [6] or the Cramer-Wold [10] distances, as well those derived from some classical one dimensional goodness of fit tests for normality, i.e the Cramér-von Mises and the Kolmogorov-Smirnov. Let us also note that our approach is, up to some extent, related to that of [8], where the authors propose a method for training generative AutoEncoders by explicitly testing $P_{\mathcal{Z}}$ via the Shapiro-Wilk test for (one-dimensional) normality, applied to a "vectorized" (multidimensional) sample (z_i).

Consequently, we use the following models:

- Sliced Wasserstein AutoEncoder (SWAE) [6],
- Sliced Closed Form Wasserstein AutoEncoder (SCFWAE) – an upgrade of SWAE,
- Sliced Cramer-Wold AutoEncoder (SCWAE), based on one dimensional Cramer-Wold distance [10],
- Sliced Cramér-von Mises AutoEncoder (SCvMAE)using Cramér-von Mises normality test,
- Sliced Kolmogorov-Smirnov AutoEncoder (SKSAE), based on Kolmogorov-Smirnov normality test.

There is also an important novelty which we have adopted from [10], namely we use the logarithm-like modification of the cost function. The main idea is that instead of considering the cost function of the form

$$\mathrm{RecError} + \lambda \cdot \mathrm{NormalityIndex},$$

which needs a grid search over λ for the proper weighting of reconstruction error RecError and divergence from normality we can, typically with similar or better results, use

$$\mathrm{RecError} + \log(\mathrm{NormalityIndex}).$$

Thanks to this formulation, the cost function, from the optimization point of view, does not change with rescaling of the normality index by a constant λ (in this case cost functions differ only by a constant $\log \lambda$, which results in the same gradient).

Our experiments show that applied methods can be divided into two groups given their generalization properties. The first consists of those which are a modification of standard normality tests: SCvMAE, SKSAE, see Fig. 1, while the second is based on classical distances between samples: SWAE, SCFWAE, SCWAE, see Fig. 2. Methods from both groups are correct generative models, but those from the second one give a slightly faster decrease rate of Fréchet Inception Distance FID [4].

2. Related works

The field of representation learning was initially driven by supervised approaches, with impressive results using large labelled datasets. Unsupervised generative modeling, in contrast, used to be a domain governed by probabilistic approaches focusing on low-dimensional data. The situation was changed with introduction of Variational AutoEncoders (VAE) [5], which were the first AutoEncoder based generative models. As a deep learning techniques for learning latent representations, VAE are used to draw images, achieve state-of-the-art results in semi-supervised learning, as well as interpolate between sentences.

One of the the most important aspect in generative models is computational complexity and effectiveness of a distance between the true and the model distribution. Originally in VAE this computation was carried out using variational methods. An important improvement was brought by using the Wasserstein metric to measure the mentioned distance, which relaxed the need for variational methods and led to the construction Wasserstein AutoEncoder (WAE) [11].

The next contribution into this research trend was made in [6], where the authors used a sliced version of the Wasserstein distance, instead of the JS-divergency as in WAE-GAN or the maximum mean discrepancy as in WAE-MMD, to penalize dissimilarity between the distribution of encoded training samples and the prior on the latent space. The obtained generative model was called the Sliced-Wasserstein AutoEncoder (SWAE).

The other related concept can be found in [10], where the authors constructed the Cramer-Wold AutoEncoder (CWAE), by replacing the sliced Wasserstein distance in SWAE by the newly introduced CW-distance between distributions, which based on the Cramer-Wold Theorem [1]. It should be noticed here that, despite the fact that CWAE can be also considered as a version of WAE-MMD method (with a choice of a specific kernel), it involved a closed formula of the CW-distance that came from the application of a sliced approach. Thus, CWAE can be seen as a borderline model between SWAE and WAE-MMD.

With reference to the above mentioned models, in the next section we derive the detailed concept of this paper.

3. Model

For convenience of the reader and to establish notation let us start from a classical AutoEncoder (AE) architecture. Let $X = (x_i)_{i=1..n} \subset \mathbb{R}^N$ be a given data set, which can be considered as sample from (true but unknown) data distribution P_X. The basic aim of AE is to transport the data to a (typically, but not necessarily) less dimensional latent space $\mathcal{Z} = \mathbb{R}^D$ with reconstruction error as small as possible. Thus, we search for an encoder $\mathcal{E}\colon \mathbb{R}^N \to \mathcal{Z}$ and decoder $\mathcal{D}\colon \mathcal{Z} \to \mathbb{R}^N$ functions, which minimize the

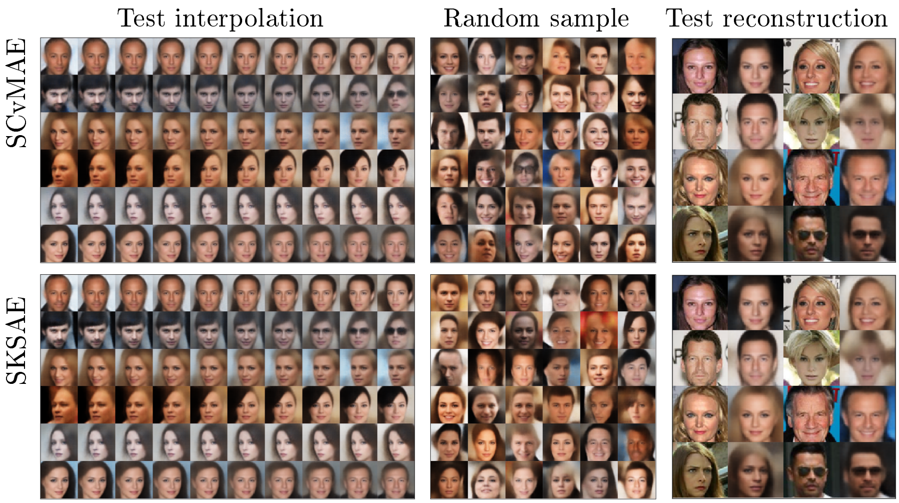

Figure 1. Results of SCvMAE and SKSAE models trained on CELEB A dataset. In "test reconstructions" odd rows correspond to the real test points.

reconstruction error on the data set X:

$$MSE(X;\mathcal{E},\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} \|x_i - \mathcal{D}(\mathcal{E}x_i)\|^2.$$

In turn, AutoEncoder based generative model is a modification of AE model by introducing a cost function that forces the model to be generative, i.e., ensures that the data transported to the latent space $\mathcal{Z}$ come from the (typically Gaussian) prior distribution $P_{\mathcal{Z}}$. A usual way to obtain this is through adding to $MSE(X;\mathcal{E},\mathcal{D})$ a regularized (using appropriately chosen hyper-parameter $\lambda > 0$) term that penalizes dissimilarity between the distribution of the encoded data $P_{\mathcal{E}(X)}$ and $P_{\mathcal{Z}}$:

$$COST(X;\mathcal{E},\mathcal{D}) = MSE(X;\mathcal{E},\mathcal{D}) + \lambda \cdot d(P_{\mathcal{E}(X)}, P_{\mathcal{Z}}). \tag{1}$$

The main idea of WAE was based on the use of the Jensen-Shannon divergence (in WAE-GAN) or the maximum mean discrepancy (in WAE-MMD) as $d(P_{\mathcal{E}(X)}, P_{\mathcal{Z}})$, which required sampling from $P_{\mathcal{Z}}$. Note that the Wasserstein metric was applied there to measure only the distance between P_X and the model distribution $P_{\mathcal{D}(\mathcal{E}(X))}$ (this approach is, in fact, a generalization of the reconstruction error $MSE(X;\mathcal{E},\mathcal{D})$ and coincide with it in the case of 2nd Wasserstein metric).

As mentioned in the introduction in this paper, we apply a modification of the cost function, which uses logarithm of the dissimilarity measure instead of (potential grid search over) hyperparameter λ:

$$COST(X;\mathcal{E},\mathcal{D}) = MSE(X;\mathcal{E},\mathcal{D}) + \log(d(P_{\mathcal{E}(X)}, P_{\mathcal{Z}})). \tag{2}$$

The modification introduced in SWAE relied on the use of the sliced Wasserstein distance to express $d(P_{\mathcal{E}(X)}, P_{\mathcal{Z}})$. The main idea was to take the mean of the

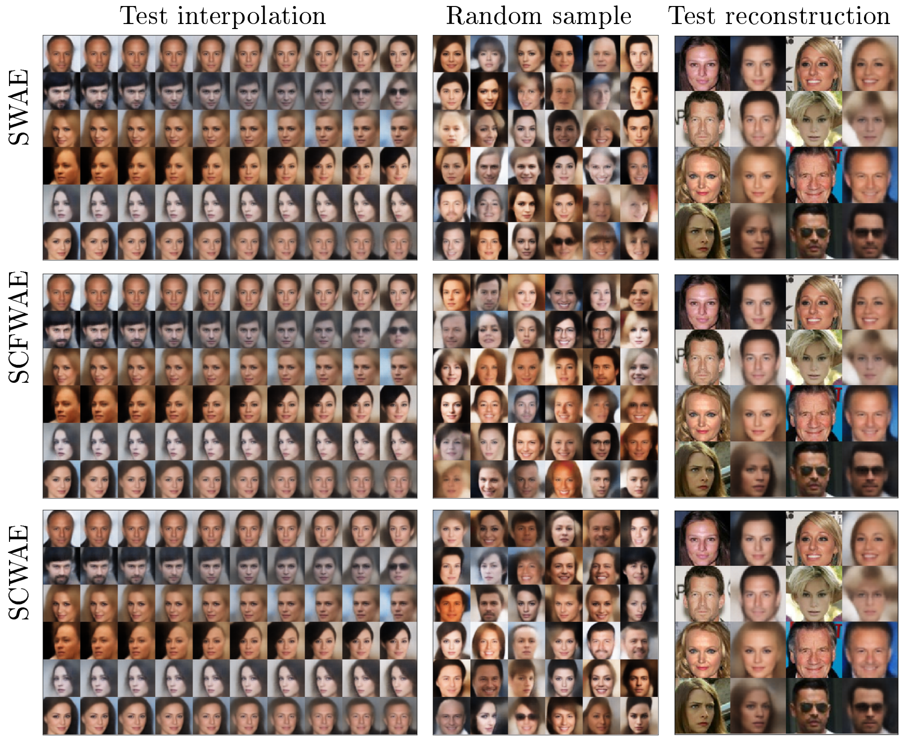

Figure 2. Results of SWAE, SCFWAE and SCWAE models trained on CELEB A dataset. In "test reconstructions" odd rows correspond to the real test points.

Wasserstein distances between one-dimensional projections of $P_{\mathcal{E}(X)}$ and $P_{\mathcal{Z}}$ on a sampled collection of one-dimensional directions. Note that SWAE, similarly to WAE, also needed sampling from $P_{\mathcal{Z}}$. Consequently in SWAE two types of sampling were applied: sampling over one-dimensional projections and sampling from the prior distribution $P_{\mathcal{Z}}$. The method is effective, but as we show in SCWAE model, it is possible to improve on it by reducing one of the above samplings by using distance between sample and the Gaussian distribution.

To the best of our knowledge, CWAE was the first WAE-like concept that required no sampling. Assuming the Gaussian prior $P_{\mathcal{Z}}$, it used (newly defined) the Cramer-Wold metric to represent $d(P_{\mathcal{E}(X)}, P_{\mathcal{Z}})$, which was expressed in an elegant closed form as the distance of a sample from standard multivariate normal distribution $N(0, I)$.

As it was mentioned before, in this paper we examine few variants of sliced distances, which possess computable closed form when considered as a measure of non-normality of a given sample, applied as a penalization term $d(P_{\mathcal{E}(X)}, P_{\mathcal{Z}})$, where $P_{\mathcal{Z}} = N(0, I)$. Specifically, assuming that we have k one-dimensional projections on the spaces spanned by the unit vectors $v_i \in \mathbb{R}^D$ for $i = 1, \ldots, k$, we define:

$$d(P_{\mathcal{E}(X)}, P_{\mathcal{Z}}) = \frac{1}{k} \sum_{i=1}^{k} d_S(v_i^T X, N(0,1)), \tag{3}$$

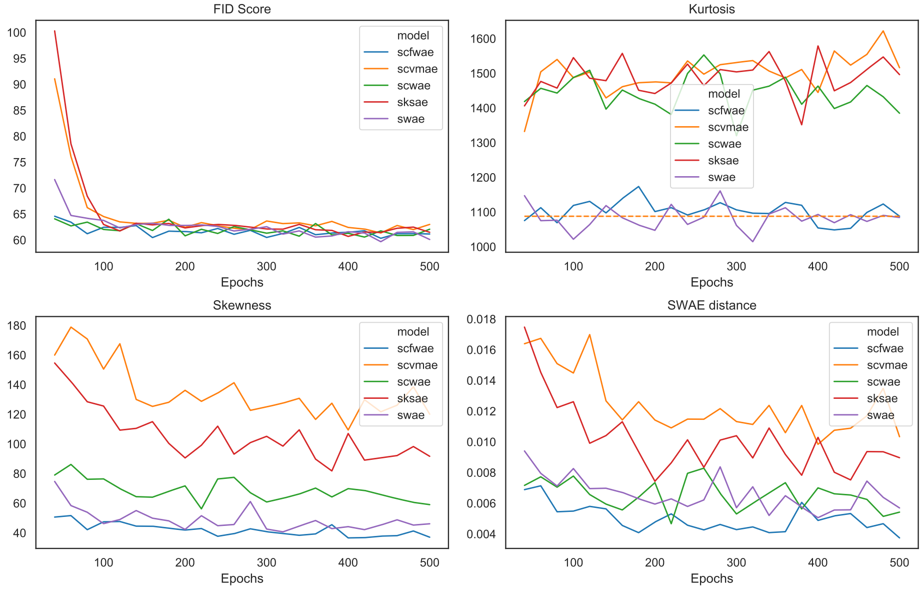

Figure 3. Metrics assessing normality of the model output distributions, during training: FID score, Mardia's skewness, kurtosis and classical SWAE distance of models SCFWAE, SCWAE, SCvMAE, SKSAE and SWAE, on the CELEB A test set. Optimal value of kurtosis (i.e. for normal distribution) is given by a dash line.

where d_S denotes a specified one-dimensional distance function (note that if a random variable $Z \in \mathbb{R}^D$ has the $N(0, I)$ distribution, then $v_i^T Z$ has the $N(0, 1)$ distribution).

4. Dissimilarity measures

In this section we make few choices of d_S's, which were used (via (2) and (3)) to construct generative AutoEncoders that are discussed in this paper.

Sliced Wasserstein AutoEncoder (SWAE). In the original SWAE paper [6], to express d_S the authors use the square of the 2-nd Wasserstein distance between the (empirical) distributions generated by the respective samples.

This leads to the following formula:

$$d_S(Y, Z) = \int_0^1 (P_Y^{-1}(t) - P_Z^{-1}(t))^2\, dt = \int_0^1 \left(\sum_{i=1}^n (y_{(i)} - z_{(i)}) \mathbf{1}_{\frac{i-1}{n} < t \leq \frac{i}{n}} \right)^2 dt$$

$$= \frac{1}{n} \sum_{i=1}^n (y_{(i)} - z_{(i)})^2,$$

where $P_*^{-1}(t) = \inf\{x \in \mathbb{R} : P_*(x) \geq t\}$ for $t \in (0,1)$, whereas $(y_{(1)},\dots,y_{(n)})$ is an ordered sample $Y = (y_1,\dots,y_n)$ and $(z_{(1)},\dots,z_{(n)})$ represents an ordered sample $Z = (z_1,\dots,z_n)$ derived from $N(0,1)$.

Sliced Closed Form Wasserstein AutoEncoder (SCFWAE). In the original SWAE paper authors have used Wasserstein distance between samples [6]. We show in SCFWAE a model that we can simplify it by using distance between sample and Gaussian density distribution (consequently, no sampling from the normal distribution is necessary). We define d_S as the square of the 2nd Wasserstein distance:

$$\begin{aligned}
d_S(Y, N(0,1)) &= \int_0^1 (P_Y^{-1}(t) - P_0^{-1}(t))^2\,dt = \int_0^1 \left(\sum_{i=1}^n y_{(i)}\mathbf{1}_{\frac{i-1}{n}<t\leq\frac{i}{n}} - P_0^{-1}(t)\right)^2 dt \\
&= \frac{1}{n}\sum_{i=1}^n y_{(i)}^2 - 2\sum_{i=1}^n y_{(i)} \int_{\frac{i-1}{n}}^{\frac{i}{n}} P_0^{-1}(t)\,dt + \int_{-\infty}^{\infty} y^2 \cdot p_0(y)\,dy \\
&= \frac{1}{n}\sum_{i=1}^n y_{(i)}^2 - 2\sum_{i=1}^n y_{(i)} \int_{Q_{\frac{i-1}{n}}}^{Q_{\frac{i}{n}}} y \cdot p_0(y)\,dy + 1 \\
&= 1 + \frac{1}{n}\sum_{i=1}^n y_{(i)}^2 - \sqrt{\frac{2}{\pi}}\sum_{i=1}^n y_{(i)} \int_{Q_{\frac{i-1}{n}}}^{Q_{\frac{i}{n}}} y \cdot \exp(-\frac{y^2}{2})\,dy \\
&= 1 + \frac{1}{n}\sum_{i=1}^n y_{(i)}^2 + \sqrt{\frac{2}{\pi}}\sum_{i=1}^n y_{(i)} (\exp(-\frac{1}{2}Q_{\frac{i}{n}}^2) - \exp(-\frac{1}{2}Q_{\frac{i-1}{n}}^2)),
\end{aligned}$$

where P_0, p_0, and Q_r denote the distribution function, the density function and the r-th quantile of $N(0,1)$.

Sliced Cramer-Wold AutoEncoder (SCWAE). Following [10], as d_S we choose the square of the one dimensional Cramer-Wold distance, which is defined as an ℓ_2 distance between a sample $Y = (y_1,\dots,y_n) \subset \mathbb{R}$ and $N(0,1)$, both smoothen using a Gaussian kernel $N(0,\gamma)$, where $\gamma = (\frac{4}{3n})^{2/5}$ is a bandwidth constant given by the Silverman's rule of thumb (see [9]). This leads to the following formula:

$$\begin{aligned}
d_S(Y, N(0,1)) &= \|\frac{1}{n}\sum_{i=1}^n p_{y_i,\gamma} - p_{0,1+\gamma}\|_2^2 = \frac{1}{n^2}\langle \sum_{i=1}^n p_{y_i,\gamma}, \sum_{i=1}^n p_{y_i,\gamma}\rangle_2 \\
&+ \langle p_{0,1+\gamma}, p_{0,1+\gamma}\rangle_2 - \frac{2}{n}\langle \sum_{i=1}^n p_{y_i,\gamma}, p_{0,1+\gamma}\rangle_2 \\
&= \frac{1}{n^2}\sum_{i,j=1}^n p_{y_i-y_j,2\gamma}(0) + p_{0,2+2\gamma}(0) - \frac{2}{n}\sum_{i=1}^n p_{y_i,1+2\gamma}(0),
\end{aligned}$$

where by $p_{m,\sigma}$ we denote the density function of $N(m,\sigma)$.

In addition to the classic distances used in generative models, we can use various dissimilarity measures related to classical statistical tests. In the literature there are many tests for normality, which work well in the case of one dimensional datasets. In the paper we verify a possibility of application of that classical statistical models in deep generative architectures.

Sliced Cramér-von Mises AutoEncoder (SCvMAE). The first statistical model we apply is the Cramér-von Mises test for normality. It can be easily derived from an application of the Wasserstein distance. Indeed, basing on the known fact that if Y is a random variable then the variable $P_Y(Y)$ has the continuous uniform distribution $U(0,1)$, as d_S we use the square of the 2nd Wasserstein distance between the distribution of $P_Y(Y)$ and $U(0,1)$, i.e.:

$$\begin{aligned} d_S(Y,N(0,1)) &= \int_0^1 (P_Z^{-1}(t) - P_1^{-1}(t))^2\,dt \\ &= \int_0^1 \left(\sum_{i=1}^n z_{(i)}\mathbf{1}_{\frac{i-1}{n}<t\leq\frac{i}{n}} - P_1^{-1}(t)\right)^2 dt \\ &= \frac{1}{n}\sum_{i=1}^n z_{(i)}^2 - 2\sum_{i=1}^n z_{(i)}\int_{\frac{i-1}{n}}^{\frac{i}{n}} t\,dt + \frac{1}{4} + \frac{1}{12} \qquad (4) \\ &= \frac{1}{n}\sum_{i=1}^n z_{(i)}^2 - \frac{1}{n^2}\sum_{i=1}^n z_{(i)}\cdot(i^2-(i-1)^2) + \frac{1}{3} \\ &= \frac{1}{n}\sum_{i=1}^n z_{(i)}^2 + \frac{1}{n^2}\sum_{i=1}^n z_{(i)}(2i-1) + \frac{1}{3}, \end{aligned}$$

where P_1 is the distribution function of $U(0,1)$ and $(z_{(1)},\ldots,z_{(n)})$ is an ordered sample $Z = (P_1(y_1),\ldots,P_1(y_n))$. Then it is easy to verify (see, e.g., [7]) that (4) coincides with the Cramér-von Mises distance between P_Y and P_0, which is used in the Cramér-von Mises goodness of fit test for normality.

Sliced Kolmogorov-Smirnov AutoEncoder (SKSAE). Our last choice of d_S is a clasical Kolmogorov-Smirnov distance, which is used as a statistics in the Kolmogorov-Smirnov goodness of fit test for normality. It is expressed (see, e.g., [2]) by the following formula:

$$d_S(Y,N(0,1)) = \sup_y \left|P_Y(y) - P_0(y)\right| = \max_i \left|\tfrac{i}{n} - P_0(y_{(i)})\right|.$$

5. Experiments

In this section we shall empirically validate proposed models on standard benchmarks for generative models CELEB A, CIFAR-10 and MNIST. We will compare different approaches to sliced generative models SCFWAE, SCWAE, SCvMAE, SKSAE, SWAE with classical SWAE [6]. As we shall see, all the above methods can be divided in to two groups. The first contains methods which are all modifications of classical normality tests: SCvMAE, SKSAE, while the second one those based on classical distances between multidimensional samples: SCFWAE, SCWAE and classical SWAE. It can be noticed that the second class of methods gives a slightly better results.

In the experiments we use two basic architecture types. Experiments on MNIST use a feed-forward network for both encoder and decoder, and a 20 neuron latent

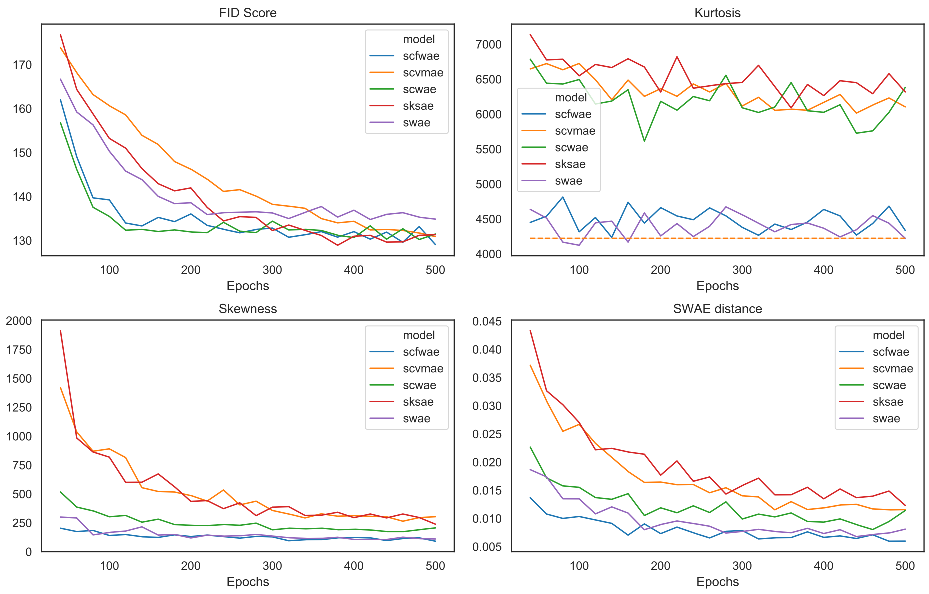

Figure 4. Metrics assessing normality of the model output distributions, during training: FID score, Mardia's skewness, kurtosis and classical SWAE distance of models SCFWAE, SCWAE, SCvMAE, SKSAE and SWAE, on the Cifar 10 test set. Optimal value of kurtosis, (i.e. for normal distribution) is given by a dash line.

layer, all using ReLU activations. For CIFAR-10 and CELEB A data sets we use convolution-deconvolution architectures.

The quality of a generative model is typically evaluated by examining generated samples or by interpolating between samples in the latent space. We present such a comparison between all approaches in Fig. 1 and Fig. 2. The experiment shows that there are no perceptual differences between considered models. In order to quantitatively compare all above slicing methods we use three measures. First of all, we use the Fréchet Inception Distance (FID) [4], which is the most popular measure of generalization in deep generative models.

Next, following experiments from [10], we verified standard normal distribution in the latent by using statistical normality tests, i.e. Mardia tests [3]. More precisely we use skewness $b_{1,D}(\cdot)$ and kurtosis $b_{2,D}(\cdot)$ of a sample $X = (x_i)_{i=1..n} \subset \mathbb{R}^D$:

$$b_{1,D}(X) = \frac{1}{n^2}\sum_{j,k}(x_j^T x_k)^3 \text{ and } b_{2,D}(X) = \frac{1}{n}\sum_{j}\|x_j\|^4$$

are close to that of standard normal density. The expected Mardia's skewness and kurtosis for standard multivariate normal distribution is 0 and $D(D+2)$, respectively.

Results are presented in Figure 3, Figure 4 and Table 1. In Figure 3 we report for CELEB A data set the value of FID score, Mardia's skewness and kurtosis during learning process of SCFWAE, SCWAE, SCvMAE, SKSAE, SWAE (measured on the validation data set). Methods based on modification of classical normality tests: SCvMAE, SKSAE obtain a sightly worse skewness and kurtosis in the case of both

Table 1. Comparison between different models output distributions and the normal distribution, together with reconstruction error. All model outputs except AE are similarly close to the normal distribution. Normality is assessed by comparing Mardia's skewness, kurtosis (normalized), and the reconstruction error. For reference FID scores are provided as well (except for MNIST, where it is not defined).

Data set	Method	SWAE	SKSAE	SCWAE	SCvMAE	SCFWAE
MNIST	Skewness	35.86	57.34	34.19	59.22	37.41
	Kurtosis (normalized)	-57.46	35.33	-10.29	23.82	-31.93
	Reconstruction error	5.37	5.01	5.35	5.04	5.42
CIFAR10	Skewness	110.49	238.52	206.50	303.45	91.42
	Kurtosis (normalized)	-0.96	2093.68	2159.31	1879.98	111.21
	Reconstruction error	27.02	24.93	27.29	25.60	26.35
	FID score error	134.87	131.32	131.48	130.89	129.07
CelebA	Skewness	46.14	91.68	59.07	120.09	37.07
	Kurtosis (normalized)	-3.60	408.24	296.99	428.18	0.17
	Reconstruction error	115.68	115.57	115.30	115.62	115.26
	FID score error	60.10	61.49	62.09	63.01	61.16

data-sets. On the other hand all methods gives similar level of FID score but it can be seen that SCFWAE, SCWAE and classical SWAE faster convergence.

6. Conclusions

In this paper, we have compared a few different approaches to construct sliced AutoEncoder based generative models. In particular, we used classical one-dimensional distances between samples and arbitrary fixed density distribution, some of them derived from classical (one-dimensional) goodness of fit tests for normality. Moreover, we have constructed SCFWAE – a simplified version of SWAE, where there is no necessity to sample from the normal prior. Our experiments show that all considered method are correct generative models, but the methods based on the Wasserstein and the Cramer-Wold distances have slightly faster decrease rate of the FID score.

Acknowledgements

The work of P. Spurek was supported by the National Centre of Science (Poland) Grant No. UMO-2015/19/D/ST6/01472. The work of J. Tabor was supported by the National Centre of Science (Poland) Grant No. UMO-2017/25/B/ST6/01271.

7. References

[1] H. Cramér and H. Wold. Some theorems on distribution functions. *London Math. Soc.*, 11:290–294, 1936.

[2] M. Hazewinkel, ed. *Kolmogorov–Smirnov test.* Encyclopedia of Mathematics. Springer Science+Business Media B.V. / Kluwer Academic Publishers, 2001.

[3] N. Henze. Invariant tests for multivariate normality: a critical review. *Statist. Papers*, 43(4):467–506, 2002.

[4] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. *arXiv:1706.08500*, 2017.

[5] D.P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv:1312.6114*, 2014.

[6] S. Kolouri, P.E. Pope, C.E. Martin, and G.K. Rohde. Sliced wasserstein autoencoders. 2018.

[7] M. Mazur and P. Kościelniak. On some goodness of fit tests for normality based on the optimal transport distance. *submitted.*

[8] A. Palmer, D. Dey, and J. Bi. Reforming generative autoencoders via goodnessoffit hypothesis testing. *UAI*, 2018.

[9] B.W. Silverman. *Density estimation for statistics and data analysis.* Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986.

[10] Jacek Tabor, Szymon Knop, Przemysław Spurek, Igor Podolak, Marcin Mazur, and Stanisław Jastrzębski. Cramer-wold autoencoder. *arXiv preprint arXiv:1805.09235*, 2018.

[11] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf. Wasserstein autoencoders. *arXiv preprint arXiv:1711.01558*, 2017.