# Generatywne AutoEncodery

April 24, 2018

## 1 Introduction

Deep generative models have achieved impressive success in recent years. In particular, Generative Adversarial Networks (GANs) [5] and Variational Autoencoders (VAEs) [7], as powerful frameworks for deep generative model learning, have largely been considered as two distinct paradigms.

Our aim is to present an AutoEncoder based alternative to VAE. More precisely we present an easy to implement generative AutoEncoder. The construction comes from our study of the following important scientific question:

Does the generative AE based model has to be so complicated as VAE?

The problem is that the basic aim of VAE is to just enforce that the autoencoder is normal in the latent space. Why to do this one has to use variational approach and nontrivial optimization? Clearly, to deal with such task, a first basic natural approach is to take an AE, and add to cost a part measuring the distance from normality (in the case of VAE this is in a sense done by variational approach and KL-divergence).

However, in practice the above idea cannot be applied as there is no measure of normality which is well-suited with dealing with highly dimensional data with dimension $D \geq 20$. Considered the best, the most popular measures of normality known in the statistical society are BHEP (and its special case Bowman-Foster) and Mardia skewness and curtosis test [6]. However, both this tests are not suitable for use as a measure of normality in high dimensions, see Remark 1.1.

Thus to realize our scientific aim we decided to construct an index of normality well-adapted to highly dimensional data. Our idea comes from the well-known Cramer-Wold Theorem [2], which states that a Borel probability

measure on Euclidean space is determined by the values it assigns to all half-spaces (equivalently, by its projections to lines through the origin).

This theorem is a basic step in many methods which rely on the reduction of the high-dimensional problem to many separate one-dimensional, in particular it serves as a crucial step in the proof of universal approximation theorem [3]. Since for one-dimensional data we have reliable kernel density estimation, we can easily use any measure of the distance from the standard univariate normal density. By integrating over all one dimensional projections to lines we obtain our final measure of non-normality $\mathrm{NI}_{\mathrm{CW}}(\cdot)$. Thus we arrive at the normality index which has the following properties:

- $\mathrm{NI}_{\mathrm{CW}}(X) > 0$, if $\mathrm{NI}_{\mathrm{CW}}(X)$ is close to zero then the data $X$ is close to the normal distribution,
- for large dimensions $D \geq 20$ it has simple asymptotic form.

Thus our final generative model is the standard AE with the addition of $\mathrm{NI}_{\mathrm{CW}}(\cdot)$.

In our experiments we apply only the limiting formula (except for the one illustrative experiment with 2-dimensional case). It occurs that due to simplicity GAE, compared to VAE, learns better and obtains normality at the smaller cost of reconstruction error.

## 1.1 Normality tests

For the survey on the normality tests we refer the reader to [1, 6]. We discuss here the most important tests.

The BHEP and Bowman-Foster test

We fix $\beta$ (a smoothing parameter, with default setting $\beta = 0.5$), and put

$$T_{n,\beta} = \frac{1}{n^2}\sum_{j,k=1}^{n} \exp(-\frac{\beta^2}{2}\|x_j - x_k\|^2) - \frac{2}{n}(1+\beta^2)^{-d/2}\sum_{j=1}^{n}\exp(-\frac{\beta^2\|x_j\|^2}{2(1+\beta^2)}) + (1+2\beta^2)^{-d/2}.$$

Now the closer $T_{n,\beta}$ is to zero, the more given sample is considered to be close to normal distribution.

The BHEP test can be seen as the application of the standard kernel density estimation, when one measures the $L^2$ distance between the regularized sample and regularized normal density [6]:

$$T_{n,\beta} = \frac{(2\pi)^{D/2}}{\beta^D}\|N(0, \tfrac{2\beta^2+1}{2\beta^2}I) - \frac{1}{n}\sum_{i=1}^{n} N(x_i, \tfrac{1}{2\beta^2}I)\|_2. \qquad (1)$$

Thus $\beta$ has the role of smoothing parameter.

Bowman-Foster test is special case of BHEP, where the parameter $\beta$ is chosen so that the associated kernel density estimation in (1) is optimal (in the sense of the Silverman's rule of thumb[1] [8, 4]):

$$\beta_{\text{B.-F.}} = \frac{1}{\sqrt{2}h_{opt}} \text{ where } h_{opt} = (\tfrac{4}{n(D+2)})^{1/(D+4)}.$$

Observe, that for large dimension $D$, even for large sample sizes, $h_{opt}$ becomes close to 1.

Mardia tests

Mardia's skewness $b_{1,D}(\cdot)$ and curtosis $b_{2,D}(\cdot)$ of a sample $X = (x_i)_{i=1..n} \subset \mathbb{R}^D$ are given by

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

The expected Mardia's skewness is 0 for a multivariate normal distribution and higher values indicate a more severe departure from normality, while the expected Mardia's kurtosis is $D(D+2)$ for a multivarite normal distribution. Thus the high values of either $b_{1,D}$ or $|b_{2,D} - D(D+2)|$ implicate that the sample is not normal.

*Remark* 1.1. At the end of this section let us explain why both this tests cannot be successfully applied as a valid normality index in high dimensions.

- *BHEP test* can be seen a the generalized application of kernel density estimation, is consistent, and consequently it can distinguish normal density from any other (for sufficiently large sample). It works well in small dimensions $D < 10$, but as show our preliminary studies, dramatically fails for large $D$ and standard sample size. The reason is that the number of datapoints needed to estimate reliably the density by kenrel methods increases faster then exponentially with the dimension [8].

- *Mardia tests* measure whether the data skewness and curtosis coincide with that of normal density. In our preliminary research they performed better as compared to BHEP, but they, contrary to BHEP, have the failure that they cannot be used as a distance from normality, as they are not consistent, i.e. do not distinguish normal density from some other spherical

[1] In the first edition of Silverman's book is the wrong formula which is ?poprawiona in the following editions.

densities [6]. The lack of consistency is not a tragic problem in the standard tests of normality, where we want to study some predefined datasets. The real problem arises in the optimization since then the neural net can easily learn to pass the test while not being normal.

# 2 Cramer-Wold Normality Index

## 2.1 Basic idea

If $X$ has density $f$, then we use the notation $f_v$ to denote the density of $v^T X$.

**Theorem 2.1** (Cramer-Wold Theorem)**.** *Let $f, g$ be densities (or in general integrable functions). If*

$$f_v = g_v \text{ for every } v \in S_{D-1}(0,1),$$

*then*

$$f = g.$$

As a direct corollary of Cramer-Wold Theorem we obtain that the density $g$ on $\mathbb{R}^D$ is equal to the standard normal density $N(0, I)$ iff its projection on the line though zero is the one dimensional density $N(0, 1)$. A natural informal consequence is that a set $X \subset \mathbb{R}^D$ was generated from the standard multivariate normal density $N(0, I)$ iff $v^T X$ comes from the normal density $N(0, 1)$ for every $v : \|v\| = 1$. So we reduce the problem to the one-dimensional case, where as we know the classical kernel density estimation works very well.

To formalize the above we thus need to be able to calculate the distance from $N(0, 1)$ for scalar data $S \subset \mathbb{R}$. We use the simplest idea and apply the $L^2$ norm between the kernel density estimation on $S$ and $N(0, 1)$ (with the kernal width assumed for the standard deviation 1).

## 2.2 Calculations

Dla dwóch gęstości $f, g$ (zakładamy, że jeden ma kow. bliska $I$) kładziemy

$$d^2(f, g) = \int_{v \in S(0,1)} \|f_v - g_v\|^2 dv$$

Naszym celem jest policzenie powyższego współczynnika w szczególnej sytuacji gdy mamy próbkę $X = (x_i)_{i=1..N} \subset \mathbb{R}^D$ i rozkład normalny $N(0, I)$. Stosujemy kernelowa estymacje w sytuacji jednowymiarowej kładąc $h$ ze wzoru Silvermana

$$h = h_{opt} = (\frac{4}{3N})^{1/5}.$$

dostajemy wzór na indeks normalności dla zbioru $X$ (im bliższy zera, tym lepiej):

$$NI(X) = \int_{v:\|v\|=1} \|f_v - g_v\|^2 dv = \int_{v:\|v\|=1} \|\frac{1}{N}\sum_{i=1}^{N} N(v^T x_i, h) - N(0,1)\|^2 dv.$$

Teraz mamy

$$\|\frac{1}{N}\sum_{i=1}^{N} N(x_i, h^2) - N(0,1)\|^2$$

$$= \frac{1}{N^2}\sum_{i,j=1}^{N} N(x_i - x_j, 2h^2)(0) + N(0,2)(0) - \frac{2}{N}\sum_{i=1}^{N} N(x_i, 1+h^2)(0).$$

Jeżeli zdefiniujemy funkcję pomocniczą

$$\Phi(z, H) = \int_{v \in S} N(v^T z, H)(0) dv,$$

to nasz końcowy wzór na indeks normalności redukuje się do

$$NI(X) = \frac{1}{N^2}\sum_{i,j=1}^{N} \Phi(y_i - y_j, 2h^2) + \Phi(0,2) - \frac{2}{N}\sum_{i=1}^{N} \Phi(y_i, 1+h^2)(0).$$

## 2.3 Wzór asymptotyczny na $\Phi$

W konsekwencji wystarczy nam policzyc

$$\Phi(z, H) = \int_{v \in S} N(v^T z, H)(0) dv.$$

Po nietrywialnych przeliczeniach można pokazać, że

$$\Phi(z, H) = \frac{2\pi^{\frac{D-1}{2}}}{\sqrt{2H}} \frac{1}{\Gamma(\frac{D}{2})} {}_1F_1\left(\frac{1}{2}; \frac{D}{2}; -\|z\|^2/(2H)\right),$$

gdzie ${}_1F_1$ to odpowiednia funkcja specjalna (patrz `https://en.wikipedia.org/wiki/Confluent_hypergeometric_function`, tam oznaczone przez $M$)

Wzór ten powyżej wartości teoretycznej, jest dla nas praktycznie mało wartościowy, gdyż funkcji ${}_1F_1$ nie ma w TensorFlow. Oznacza to, że przydatne będzie przybliżenie. Skorzystam z dziwnoście sytuacji wysokowymiarowej, gdzie rozkład normalny $N(0,I)$ koncentruje się w okolicy sfery o promieniu $\sqrt{D}$, co oznacza, że $N(0,\frac{1}{D}I)$ koncentruje się wokół $S(0,1)=\{v : \|v\|=1\}$, a jak zauważyliśmy wcześniej to nas właśnie interesuje.

**Twierdzenie 2.1.** *Mamy w dużych wymiarach*

$$\Phi(z,H) \approx \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{H+\|z\|^2/D}}.$$

*Proof.* Kładziemy $r=\|z\|$, i liczymy:

$$\Phi(z,H) \approx \int_{\mathbb{R}^D} N_1(\langle v,z\rangle,H)(0)\cdot N_D(0,\tfrac{1}{D}I)(v)dv$$

$$= \int_{-\infty}^{\infty} N_1(0,H)(rs)\frac{N_D(0,\frac{1}{D}I)(s\frac{z}{\|z\|})}{N_{D-1}(0,\frac{1}{D}I)(0)}\int_{\mathbb{R}^{D-1}} N_{D-1}(0,\tfrac{1}{D}I)(w)dw\,ds$$

$$= \int_{-\infty}^{\infty} N_1(0,H)(rs)\frac{N_D(0,\frac{1}{D}I)(s\frac{z}{\|z\|})}{N_{D-1}(0,\frac{1}{D}I)}ds$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi H}}\exp(-\frac{1}{2h}(rs)^2)\cdot\frac{\sqrt{D}}{\sqrt{2\pi}}\exp(-\frac{1}{2}Ds^2)ds$$

$$= \frac{\sqrt{D}}{2\pi\sqrt{H}}\int_{-\infty}^{\infty}\exp(-\frac{1}{2}(r^2/H+D)s^2)ds$$

$$= \frac{\sqrt{D}}{2\pi\sqrt{H}}\frac{1}{N(0,\frac{H}{r^2+HD})(0)}\int_{-\infty}^{\infty} N(0,\tfrac{H}{r^2+HD})(s)ds$$

$$= \frac{\sqrt{D}}{2\pi\sqrt{H}}\sqrt{2\pi}\sqrt{\frac{H}{r^2+HD}} = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{H+r^2/D}}.$$

□

W końcowej definicji indeksu opuszczamy $1/\sqrt{2\pi}$ bo stała, i na koniec dostajemy nasz indeks normalności zbioru $X = (x_i)_{i=1..N} \subset \mathbb{R}^D$ już w formie gotowej do wrzucenia do AE:

$$\mathrm{NI}_{\mathrm{CW}}(X) = \frac{1}{N^2} \sum_{i,j=1}^{N} \frac{1}{\sqrt{2h^2 + \|x_i - x_j\|^2/D}} + \frac{1}{\sqrt{2}} - \frac{2}{N} \sum_{i=1}^{N} \frac{1}{\sqrt{1 + h^2 + \|x_i\|^2/D}},$$

gdzie $h = (\frac{4}{3N})^{1/5}$. Jeżeli $\mathrm{NI}_{\mathrm{CW}}(X)$ jest bliskie zera, to zbiór $X$ jest bliski bycia rozkładem normalnym.

# References


[1] AW Bowman and PJ Foster. Adaptive smoothing and density-based tests of multivariate normality. *Journal of the American Statistical Association*, 88(422):529–537, 1993.

[2] Harald Cramér and Herman Wold. Some theorems on distribution functions. *Journal of the London Mathematical Society*, 1(4):290–294, 1936.

[3] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[4] Paul Deheuvels. Estimation non paramétrique de la densité par histogrammes généralisés. *Rev. Statist. Appl*, 25(3):5–42, 1977.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

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

[7] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 2014.

[8] Bernard W Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.