2021 XVII International Symposium Problems of Redundancy in Information and Control Systems (REDUNDANCY) | 978-1-6654-3308-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/REDUNDANCY52534.2021.9606448

Variational Autoencoders with Euclidean and Hyperbolic Latent Spaces for Population Genetics

1st Igor Bogdanov
International laboratory of statistical and computational genomics
HSE University
Moscow, Russia
iebogdanov@hse.ru

2nd Vladimir Shchur
International laboratory of statistical and computational genomics
HSE University
Moscow, Russia
vshchur@hse.ru

***Abstract*—Population structure inference is one of the main problems of population genetics. Genetic variation might give a clue on relations between populations as well as to identify population components in a single individual. Currently, principle component analysis (PCA) is one of standard tools for genetic data structure visualisation. In this work we present the application of variational autoencoders (VAE) with Euclidean and hyperbolic latent spaces and compare these approaches with PCA. In contrast to the PCA, VAE allows to find nonlinear dependencies in the data, and hyperbolic geometry is better suited for data with hierarchical structure. We show that VAEs have more power to separate population components in some complicated population scenarios.**

I. Introduction

Exploratory data analysis is one of the crucial steps in any data-oriented field. With the unprecedent availability of genetic data, populaion genetics is one of the most dynamic research fields nowadays. Visualusation of population structure based on the genetic variation plays an important role in most of the studies. There are two most used approaches to visualize the population genetics data: Structure/Admixture [1] [2] and principle component analysis [3]. Structure and similar methods use Bayesian algorithm to determine population components for each genome. PCA is a geometric method for qualitative analysis and visualization of genetic data. In machine learning the t-SNE algorithm [4] is also used to visualize two-dimensional representations. But t-SNE doesn't preserve the global structure, distances between clusters are not always meaningful [5].

Geometric methods for data analysis became of great interest in recent years, and showed their power in many applications [6] [3] [7]. In particular, combination of geometric methods and deep learning might give important insights in the data analysis. In this work we adopted variational autoencoders for genetic data analysis. We developped this method independently and in parallel with [8], and hyperbolic VAE for genetic variation is the novel approach in the field.

Genetic data are generated by natural genealogical processes, which lead to hierarchical relationships between individuals of the same species. Therefore, hyperbolic geometry and variational autoencoder may be an appropriateapproach for genetic data analysis. Variational autoencoder (VAE) [9] is a deep learning approach that can identify nonlinear relations within data. It consists of two fully connected neural networks: encoder and decoder. The encoder takes an object x as input and maps it to a low-dimensional distribution space on a latent space, which is usually bounded by a normal distribution. The decoder takes the latent z representation for each object and maps it back to the original space. At the end of the training we will be interested in the representation of objects in the latent space.

In the simplest case (e.g., viruses), individuals in the population are connected by a family tree. The geometry of trees is similar to the geometry of a hyperbolic plane, since trees are a 0-hyperbolic spaces [10]. The use of hyperbolic geometry in machine learning was first proposed in article [11] due to the fact that data often have a complex hierarchical structure, methods that use Euclidean spaces do not take these properties into account. The surface area of hyperbolic space grows exponentially with increasing radius, this is equivalent to the exponential growth of the number of leaves of a tree in relation to the depth of the tree. This means that in the case of Euclidean spaces, problems may arise because of this, machine learning algorithms may consider the data similar, although they are not, overfitting may occur [11]. Hyperbolic geometry is able to eliminate these disadvantages, so it is recommended to use it on data in which there is a hierarchical structure, family trees are one of such examples.

Variational autoencoders with hyperbolic latent space were proposed in [12]. It also considers two generalizations of the normal distribution on hyperbolic space, namely the Poincare ball, which are used in latent space and the construction of VAE.

II. Methods

A. VAE

Let $X = \{x_1, x_2, ..., x_N\}$ be a set of objects (individuals) consisting of genotypes. For each object x_i, we introduce an additional latent variable z_i and construct a joint probabilistic distribution $p(X, Z|\theta)$, Z is the space of latent variables, θ

V. Shchur is supported by the Russian Science Foundation under grant 20-71-00143.

are parameters for the distribution of objects X. Generative model:

$$p(x, z|\theta) = p(x|z, \theta)p(z|\theta),$$

$p(z|\theta)$ is a prior distribution on latent variables,it has a standard normal distribution. First multiplier is called a decoder, it maps the latent vectors z_i to the data space.

Also introduce the encoder model. In order to find the latent representation of z_i by the object x_i, we need to solve the maximization problem $\int p(x, z|\theta)dz$, but the integral cannot be calculated analytically. The solution to this is the approximation of $p(z|x, \theta)$ using a neural network $q(z|x, \varphi)$ an encoder. The parameters φ are the weights and biases of the neural network. The encoder accepts objects x as input, and outputs normal distribution parameters for $q(z|x, \varphi)$. Training takes place using the evidence lower bound (ELBO) optimization:

$$\arg\max_{\theta,\varphi} E_{q(z|x,\varphi)}[\log p(x|z, \theta)] - KL(q(z|x, \varphi)||p(z)),$$

KL is Kullback-Leibler divergence for objects of latent distributions relative to the standard normal distribution $p(z)$. And the first term is usually binary cross-entropy between true and generated objects.

B. Hyperbolic VAE

$\mathbb{H}^d$ is a d-dimensional hyperbolic space, complete, simply connected d-dimensional Riemannian manifold with constant negative curvature c. $\mathbb{H}^d$ can be constructed using various equivalent models of hyperbolic geometry. We will consider the implementation of a hyperbolic space, which is called a Poincare ball, because the Poincare ball is better suited for gradient optimization, and therefore for training neural networks. Poincare ball is the Riemannian manifold $\mathbb{B}_c^d = (B_c^d, g_p^c)$, B_c^d - open ball of radius $\frac{1}{\sqrt{c}}$ and g_p^c is an metric tensor, $g_p^c = (\lambda_z^c)^2 g_e(z)$, g_e - Euclidean metric tensor with usual dot product. Distance on Poincare ball:

$$d_p^c(z, y) = \frac{1}{\sqrt{c}} \cosh^{-1}(1 + 2c\frac{||z-y||^2}{(1-c||z||^2)(1-c||y||^2)})$$

Define Mobius addition of $z, y \in \mathbb{B}_c^d$:

$$z \oplus_c y = \frac{(1+2c\langle z,y\rangle+c||y||^2)z+(1-c||z||^2)y}{1+2c\langle z,y\rangle+c^2||z||^2\cdot||y||^2}$$

Using the introduced Mobius addition, we will set the exponential and logarithmic maps on the Poincare ball [13]:

$$exp_z^c(x) = z \oplus_c (\tanh(\sqrt{c}\frac{\lambda_z^c \cdot ||x||}{2})\frac{x}{\sqrt{c}||x|}$$

$$\log_z^c(x) = \frac{2}{\sqrt{c}\lambda_z^c} \tanh^{-1}(\sqrt{c} \cdot || - z \oplus_c x||) \cdot \frac{-z\oplus_c x}{||-z\oplus_c x||}$$

We will be interested in a variational autoencoder with a hyperbolic hidden space, namely a Poincare ball [12]. We consider the problem of mapping objects into a low-dimensional Poincare ball $\mathbb{B}_c^d$, as well as studying the mapping from this latent hidden space $Z = \mathbb{B}_c^d$ back to the observation space X.The hyperbolic variational autoencoder differs from the standard VAE by the choice of the prior and posterior distributions defined on $\mathbb{B}_c^d$, as well as the setting of the g_φ encoder and f_θ decoder, which take into account the geometry of the hidden space. To define distributions on a Poincare ball, we consider a generalization of normal distributions on this space, which is called a Wrapped normal distribution. It is obtained by considering the image obtained by taking the exponential map of the usual normal distribution on the tangent space, where the center is the mean value. The elements on $z \in \mathbb{B}_c^d$ are obtained as $z = \exp_\mu^c(\frac{x}{\lambda_\mu^c})$ where $x \sim N(\cdot|0, \Sigma)$. The prior distribution on Z is given as Wrapped normal distribution with mean zero $p(z) = N_{\mathbb{B}_c^d}(\cdot|0, \sigma_0^2)$, the variance is selected from $\{N_{\mathbb{B}_c^d}(\cdot|\mu, \sigma^2) \mid \mu \in \mathbb{B}_c^d, \sigma \in \mathbb{R}_+\}$.

The encoder predicts the parameters of the a posterior normal distribution on the Poincare ball $\mu \in \mathbb{B}_c^d$ and $\sigma \in \mathbb{R}_+$. The mean μ is obtained as an image of the exp_0^c and σ via the softplus function.For the decoder, it is proposed to use the operator $f_{a,p}^c : \mathbb{B}_c^d \to \mathbb{R}^p$ on a Poincare ball, it is an analog of an affine transformation in Euclidean space:

$$f_{a,p}^c(z) = sign(\langle a,\ \log_p^c(z)\rangle_p)\ ||a||_p\ d_p^c(z, H_{a,p}^c)$$
$$H_{a,p}^c = \{z \in B_c^d | \langle a, \log_p^c(z)\rangle = 0\} = exp_p^c(\{a\}^\perp)$$

III. Results

We applied our method to a data set of 1000 Genomes Project (55 AISNP panel) [14] as well as of simulated data generated with msprime [15]. The dataset of 1000 genomes consists of a table in which the rows are people (2504 individuals), and the columns are genetic variants, namely SNP. The study of population structure was carried out using the PCA, VAE and hyperbolic VAE methods.

(a) PCA (b) VAE (c) HVAE

Fig. 1: Performance of PCA, VAE, HVAE on the real genomes representing five populations from 1000 Genomes Project.

The PCA was able to separate Africa, East Asia and Europe well, but it had problems with America and South Asia.

This is the result of the work of the standard VAE - it is the representation of data in two-dimensional latent space after training the neural network. VAE did better than PCA, America and South Asia can now be divided.

The hyperbolic VAE separated Africa, South Asia, and most of the European population. It becomes difficult to interpret the two remaining populations, but since when approaching the border of the Poincare ball, the distances between objects increase sharply, it can be assumed that most of these populations can be divided.

Experiments were also carried out on data from the simulator. The data was generated using the msprime library. Initially, there were two ancestral populations, and 500 generations ago, some individuals from each of these populations separated and

formed a new third population. The result of work for a dataset with 300 individuals and 10,000 SNPs:

(a) PCA (b) VAE (c) HVAE

Fig. 2: Performance of PCA, VAE and HVAE on simulated data with three populations. There are two founding population and an admixed population.

We will also give an example when the PCA can no longer cope with the division of populations. Consider the following population tree in which we have four present-day populations A, B, C and D (100 individuals in each). And some individuals from populations A, B and C combine with individuals from population D, forming new populations of AD, BD and CD.

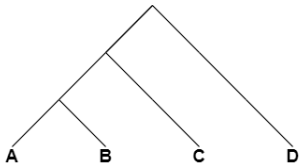
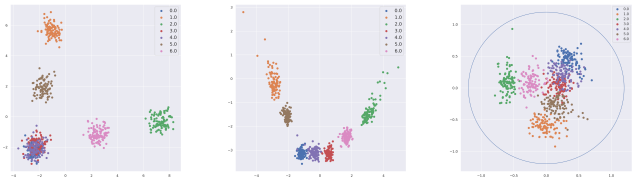


Fig. 3: Founding populations used for the second simulation.



(a) PCA (b) VAE (c) HVAE

Fig. 4: Performance of PCA, VAE and HVAE on simulated data with seven populations. There are four founding population and three admixed population

PCA was unable to separate populations A, D and AD, but VAE did.

Hyperbolic VAE also divided all the populations, but the clusters have a large spread on the Poincare ball. Source code is available https://github.com/igor-bogdanov/ProjectVAE_Bogdanov .

IV. Conclusion

This work has shown that the use of variational autocoders with a hyperbolic hidden space is an appropriate approach for studying the population structure from genetic data.

It was demonstrated that the variational autoencoder is able to find new dependencies in genetic data compared to the principal component analysis. This approach may affect the interpretation of old results that were previously considered only using PCA.

Variational autoencoders are also generative models, so in the future they can be used to simulate data, which will increase the size of the training sample.

A serious disadvantage of autoencoders is their training time, but information about new, nonlinear dependencies is more valuable, so the running time of algorithms is not a key factor in choosing a model.

Also, a further research issue is the interpretation and improvement of the results of hyperbolic VAE. At the moment, it may be difficult to visualize individuals on real data due to the peculiarities of hyperbolic geometry, so it is necessary to conduct more experiments with different neural network architecture and different settings of hyperparameters, for example, the curvature of the latent space.

References

[1] J. K. Pritchard, M. Stephens, and P. Donnelly, "Inference of population structure using multilocus genotype data," *Genetics*, vol. 155, no. 2, pp. 945–959, 2000.

[2] D. H. Alexander, J. Novembre, and K. Lange, "Fast model-based estimation of ancestry in unrelated individuals," *Genome research*, vol. 19, no. 9, pp. 1655–1664, 2009.

[3] N. Patterson, A. L. Price, and D. Reich, "Population structure and eigenanalysis," *PLoS genetics*, vol. 2, no. 12, p. e190, 2006.

[4] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[5] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell, "Dimensionality reduction for visualizing single-cell data using umap," *Nature biotechnology*, vol. 37, no. 1, pp. 38–44, 2019.

[6] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature genetics*, vol. 38, no. 8, pp. 904–909, 2006.

[7] O. François and F. Jay, "Factor analysis of ancient population genomic samples," *Nature communications*, vol. 11, no. 1, pp. 1–11, 2020.

[8] C. Battey, G. C. Coffing, and A. D. Kern, "Visualizing population structure with variational autoencoders," *G3*, vol. 11, no. 1, pp. 1–11, 2021.

[9] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[10] R. Sonthalia and A. C. Gilbert, "Tree! i am no tree! i am a low dimensional hyperbolic embedding," *arXiv preprint arXiv:2005.03847*, 2020.

[11] M. Nickel and D. Kiela, "Poincaré embeddings for learning hierarchical representations," *Advances in neural information processing systems*, vol. 30, pp. 6338–6347, 2017.

[12] E. Mathieu, C. L. Lan, C. J. Maddison, R. Tomioka, and Y. W. Teh, "Continuous hierarchical representations with poincar\'e variational auto-encoders," *arXiv preprint arXiv:1901.06033*, 2019.

[13] O.-E. Ganea, G. Bécigneul, and T. Hofmann, "Hyperbolic neural networks," *arXiv preprint arXiv:1805.09112*, 2018.

[14] A. J. Pakstis, L. Kang, L. Liu, Z. Zhang, T. Jin, E. L. Grigorenko, F. R. Wendt, B. Budowle, S. Hadi, M. S. Al Qahtani *et al.*, "Increasing the reference populations for the 55 aisnp panel: the need and benefits," *International journal of legal medicine*, vol. 131, no. 4, pp. 913–917, 2017.

[15] J. Kelleher, A. M. Etheridge, and G. McVean, "Efficient coalescent simulation and genealogical analysis for large sample sizes," *PLoS computational biology*, vol. 12, no. 5, p. e1004842, 2016.

V. Reviews

Thanks for the reviews. This article presents the results about the current state of work that we want to present at the conference.

Review 1

Significance	Technical Level	Novelty	Presentation	Recommendation
Very Important (4)	Technically sound (3)	Some Novelty (2)	Average (3)	Borderline (3)

Strengths (What are the key strengths of this paper?)

The authors apply modern methods--EVAE and HVAE--on an interesting dataset--1000 Genomes Project. The paper is ok written.

Review 1, Weakness: 1. The state of the art is very weak: PCA is from 1902 and is a linear method. Why do you compare with PCA? What about tSNE (also good with heavy tails)? What about other methods with hyperbolic distance? For example, nonmetric MDS (https://www.sciencedirect.com/science/article/pii/S2589004221001930)?

Thanks for the remark, in population genetics, PCA is the main geometric method for analyzing and visualizing genetic data, it is a simple and fast algorithm, so our approaches are compared with this method. The t-SNE algorithm shows good results in visualizing representations, but distances between clusters are not always meangful. We added this part about t-SNE to the article.

Review 1, Weakness: 2. The numerical evaluation is too weak. The authors provide only pictures and their views on them. What about numerical quantities (distance preservation, classification quality on the top of low dimensional representations)?

Other

3. The source code is not provided

Thanks, we will take this remarks about details of the experiments and quantitative analysis into account in our future work. Source code is available at github (https://github.com/igor-bogdanov/ProjectVAE_Bogdanov).

Review 1, Weakness: 4. The Russian-style editing does not always work in English. The dash is ”–” in Latex and does not need spaces around. But try to use ”is a/are” instead. For example, ”g_e is a Euclidean metric tensor” instead of ”g_e – ...”. See http://www.ega-math.narod.ru/Quant/ABS.htm for instructions.

5. Split affiliations into two lines, please: $>>$ International laboratory of statistical and computational genomics

$- >$

$>>$ International laboratory of statistical and computational genomics

Thanks, fixed.

Review 1, Comments and Recommendation: Please, extend competitors and provide quantitative analysis in the experiments section to improve the quality of the paper.

Review 2, Weakness: 1. The idea is simple and looks like ”tried it and it worked”

Review 2

Significance	Technical Level	Novelty	Presentation	Recommendation
Average Importance (3)	Not very convincing (2)	Little Novelty (1)	Good (4)	Weak Reject (2)

Strengths (What are the key strengths of this paper?)

1. Good mathematical description of VAE and HVAE
2. The article is written in a simple and understandable language

Thanks, we think that this idea is an extension of the PCA algorithm: VAE is a non-linear analogue of PCA and hyperbolic geometry is better suited for genetic data.

Review 2, Weakness: 2. Small amount of experiments and comparisons

3. Any metrics missing

4. Any experiment details missing

Thanks, we will take this remark into account in our future work.

Review 2, Weakness:1. Small amount of experiments and comparisons. It is a good chance, that simple autoencoder with custom activation function will be better, than PCA and compatible with VAE on these type of the data

3. There are only visual results and conclusions. Some clustering metrics may be used to prove that HVAE outperforms PCA and VAE. Also, 3D plot (maybe) will be better for HVAE visualization, because it is not clear that HVAE is better than VAE

4. Small amount of experiment details. What is the VAE hyperparameters? How much data was used for training and testing? Were the figures drawn using train data or test data? What the metrics of the trained VAE? Did the VAE converges or not?

Thanks, 3D plots can be used, but they are poorly represented on paper. For VAE hyperparameters are number and size of hidden layers, for HVAE also the curvature of hyperbolic space. The 1000 Genomes Project contains 2504 individuals, the simulated datasets contain 300 and 700 individuals, respectively. Metric learning is ELBO this is the sum of of binary cross-entropy and KL divergence. We added all this information into the manuscript.