Dimensionality Reduction using Contrastive Learning

UPPSALA UNIVERSITET

Filip Thor filip.thor@it.uu.se Carl Nettelblad carl.nettelblad.it.uu.se

Division of Scientific Computing, Department of Information Technology Science for Life Laboratory (SciLifeLab)

Introduction

Interpreting the genetic variation in a dataset through visualization techniques can be a powerful first step when studying population structure. This analysis requires a method of dimensionality reduction, where principal component analysis (PCA) has seen the most frequent use. We develop SNP-based dimensionality reduction models using deep learning techniques, which can produce more information-rich embeddings.

Contemporary Methods

PCA is a linear method that computes a lower dimensional representation with the aim to preserve as much variance as possible. It has been shown to capture global patterns.

UMAP and t-SNE are non-linear methods that promote clustering through an attraction-repulsion algorithm. They perform well locally, but can fail to capture global relationships. Projecting new samples to an existing embedding is non-trivial with these methods.

Why Use Deep Learning?

With neural networks we can use more information in the data, such as the ordering and position of the SNPs.

We find an explicit mapping $f : \mathbb{R}^D \to \mathbb{R}^d$, thus projecting new samples onto a learned embedding is easy. The framework is versatile, and a good higher-dimensional feature extraction should be a good foundation for future work on phenotype prediction.

Two Deep Learning Methods for Dimensionality Reduction

The **Autoencoder** (Figure 1) is a classic model for deep learning-based dimensionality reduction. The overall aim is to train the model to reconstruct the input as accurately as possible, after having been transformed to a lower-dimensional representation.

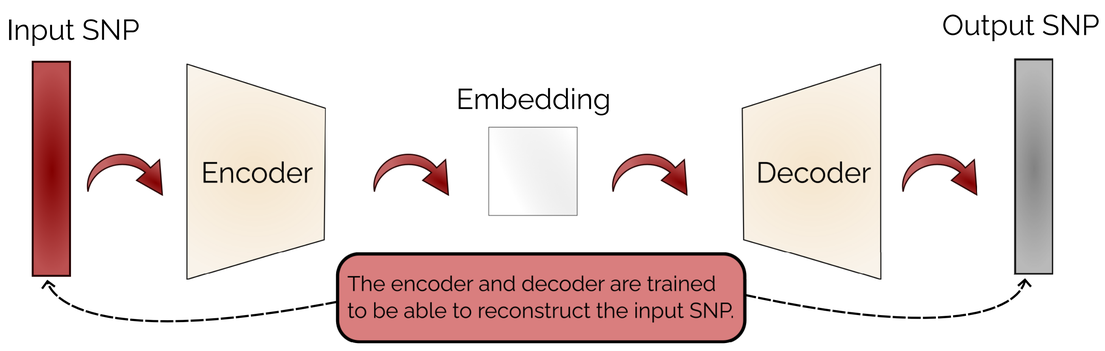

Figure 1. Autoencoder concept. **Motivation:** *"A good reconstruction requires a good embedding"*

An alternative, **Contrastive Learning** (Figure 2), minimizes distances in the embedding space between similar samples and maximizes distances to dissimilar ones.

This is done by assigning each sample positive and negative samples. The network is trained to embed the positive samples closer than the negatives. The positives are defined as some data augmentation of the original, and the negatives are taken as any other sample.

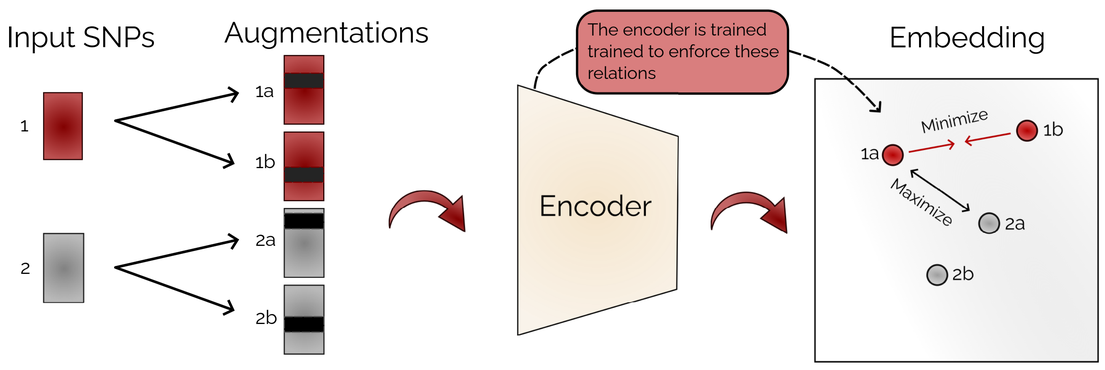

Figure 2. Contrastive learning concept. **Motivation:** *"Similar samples should be mapped similarly"*

Note: Neither of the methods requires the sample labels or relations, only the SNP data.

Example: PCA-like 2D Embeddings

Setting the embedding dimension to 2, we can produce PCA-like visualizations. Figure 3 shows an example for a dog dataset (1355 samples, 161 breeds, 23 clades, 150k SNPs) [4] where similar colors indicate samples belonging to the same clade, and same color and marker indicate the same breed.

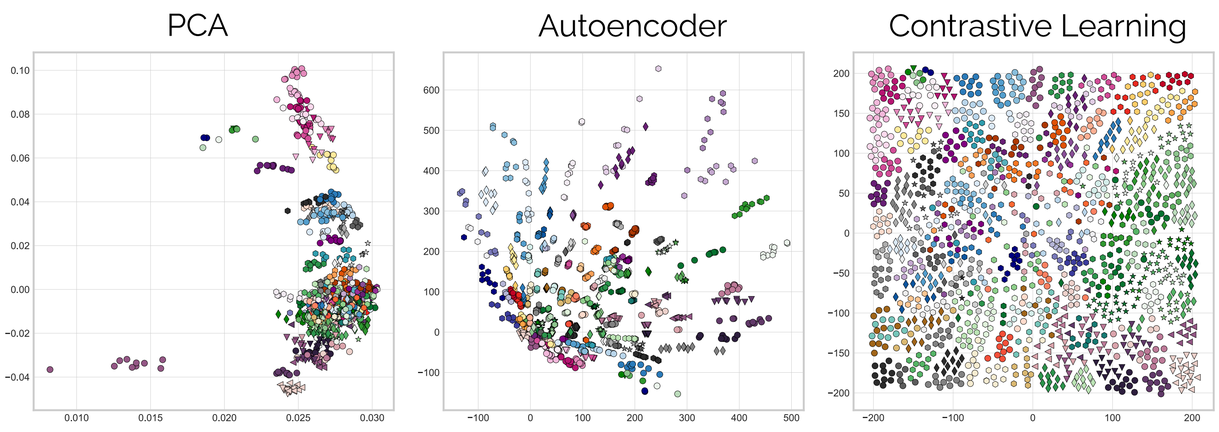

Figure 3. Dimensionality reduction using PCA and our deep learning methods on a dog dataset.

Our methods are able to untangle and visualize the data more clearly. We are also able to customize and set conditions on the embeddings. In the contrastive learning case, we have a weak restriction on the samples to get mapped to the [-200,200] square.

Example on Human Data, and Extension to Genetic Clustering

Figure 4 shows a similar example using the Human Origins dataset (160k SNPs, 2067 samples), a diverse dataset with 8 superpopulations [3]. Here, we compare t-SNE and both of our methods. We see that our methods results in a more pronounced clustering of the data, while also retaining global structure, such as continental origin.

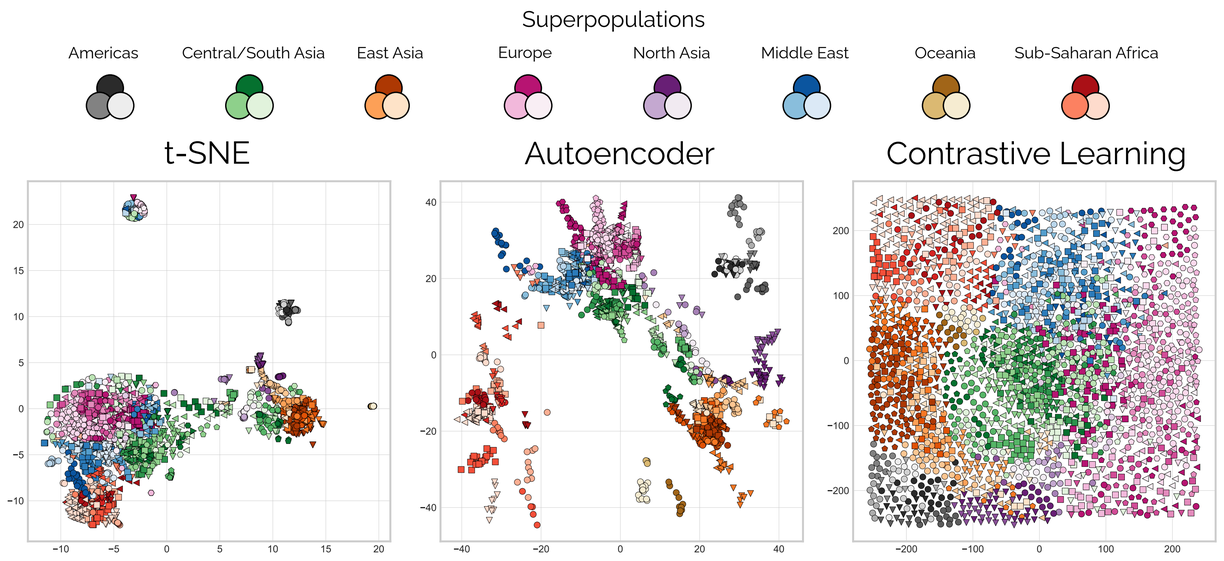

Figure 4. Dimensionality reduction using t-SNE and our deep learning methods on a human dataset.

The contrastive learning implementation is built on top of the deep learning framework GCAE [2]. That work showed that if the embedding dimension for the Autoencoder model is increased, we can instead produce ADMIXTURE-like clustering. Figure 5 is an example for $k = 5$ clusters on the Human Origins dataset, where the Autoencoder outputs a result similar to the commonly used ADMIXTURE [1] software.

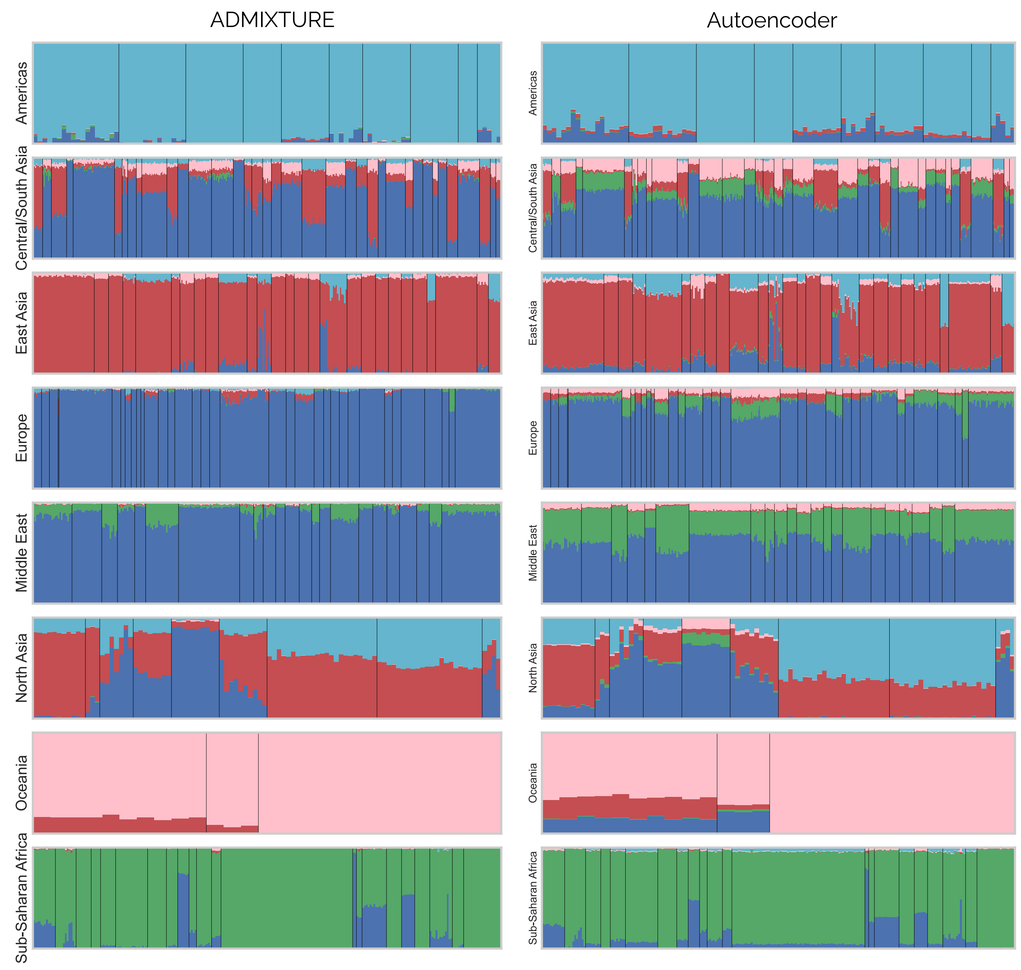

Figure 5. Comparison of ADMIXTURE and Autoencoder clustering of the Human Origins data.

Future Work on Wheat Data

We want to work with the NIAB Diverse MAGIC wheat dataset [5]. It consists of 16 founders, and 504 recombinant inbred lines (RILs). The aim would be to embed the RILs into a 16 dimensional space, where ideally the values in axis k would signify the genetic contribution from founder k. RILS within the same family should be able to cluster well since they have had the same breeding pattern. Due to the numerous matings in a MAGIC pedigree, it would make more sense to create embeddings for short genome windows, rather than the full genome. Work is ongoing to produce these in a reliable manner.

References

[1] D. H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655–1664, 2009. doi: 10.1101/gr.094052.109.

[2] K. Ausmees and C. Nettelblad. A deep learning framework for characterization of genotype data. *G3 Genes|Genomes|Genetics*, 12(3), 01 2022. ISSN 2160-1836. doi: 10.1093/g3journal/jkac020. jkac020.

[3] I. Lazaridis, D. Nadel, G. Rollefson, et al. Genomic insights into the origin of farming in the ancient near east. *Nature*, 536 (7617):419–424, Aug. 2016. ISSN 0028-0836. doi: 10.1038/nature19310.

[4] H. G. Parker, D. L. Dreger, M. Rimbault, et al. Genomic analyses reveal the influence of geographic origin, migration, and hybridization on modern dog breed development. *Cell Reports*, 19(4):697–708, 2017. ISSN 2211-1247. doi: https://doi.org/10.1016/j.celrep.2017.03.079.

[5] M. F. Scott, N. Fradgley, A. R. Bentley, et al. Limited haplotype diversity underlies polygenic trait architecture across 70 years of wheat breeding. *bioRxiv*, 2020. doi: 10.1101/2020.09.15.296533.

Acknowledgements

Compute resources provided by SNIC through National Supercomputer Centre NSC) at Linköping University under Project Berzelius 2022-17 and 2022-165. Project funded by Formas, The Swedish government research council for sustainable development, Grant no 2020-00712, Deep Learning for Analyzing Population Structure and Genotype-Phenotype Mapping.