Contrastive Learning for Dimensionality Reduction of SNP Genotype Data

Filip Thor filip.thor@it.uu.se Max Kovalenko max.kovalenko@it.uu.se Carl Nettelblad carl.nettelblad@it.uu.se

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

UPPSALA UNIVERSITET

Code and Additional Figures:

Introduction

Dimensionality reduction is usually a first step when studying the population structure of a genetic dataset. We develop SNP-based dimensionality reduction models using deep learning. In this work, we show how using contrastive learning and data augmentation strategies can train dimensionality reduction models with good generalization properties.

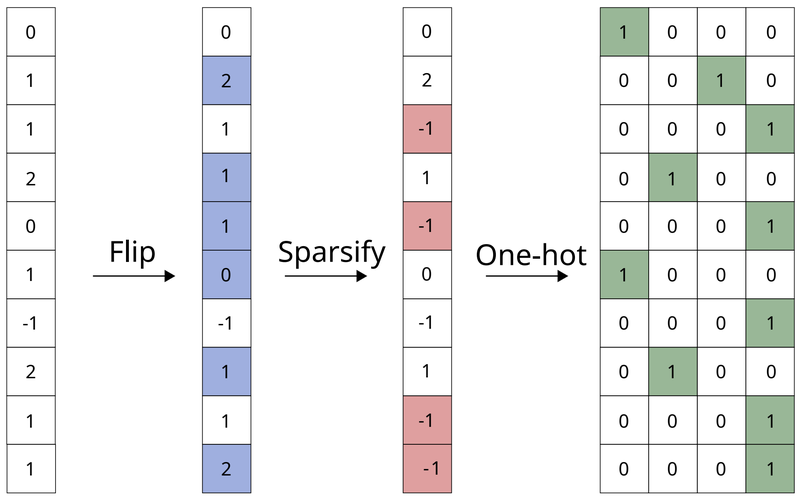


Figure 1. Our augmentation strategy for SNP vectors consists of flipping some variants and setting some as missing - both at random. The SNP data is then one-hot encoded. The sample is slightly altered, but still approximately represents the same individual.

Contrastive Learning for Dimensionality Reduction

We utilize **Contrastive Learning** (Figure 2), which minimizes distances in the embedding space between similar samples and maximizes distances to dissimilar ones, while using no information on population labels.

This is done by assigning each sample positive and negative samples. The network is trained by penalizing large embedding distances to the postive and small distances to the negatives. The positives are defined as augmented versions of the original (Figure 1), and the negatives are taken as some random sample in the batch. Since genetically similar samples will have similar augmented versions, the encoder will after training produce embeddings that reflect this similarity, mapping samples with similar genetic makeup close. The implementation is built on top of the deep learning framework GCAE [1].

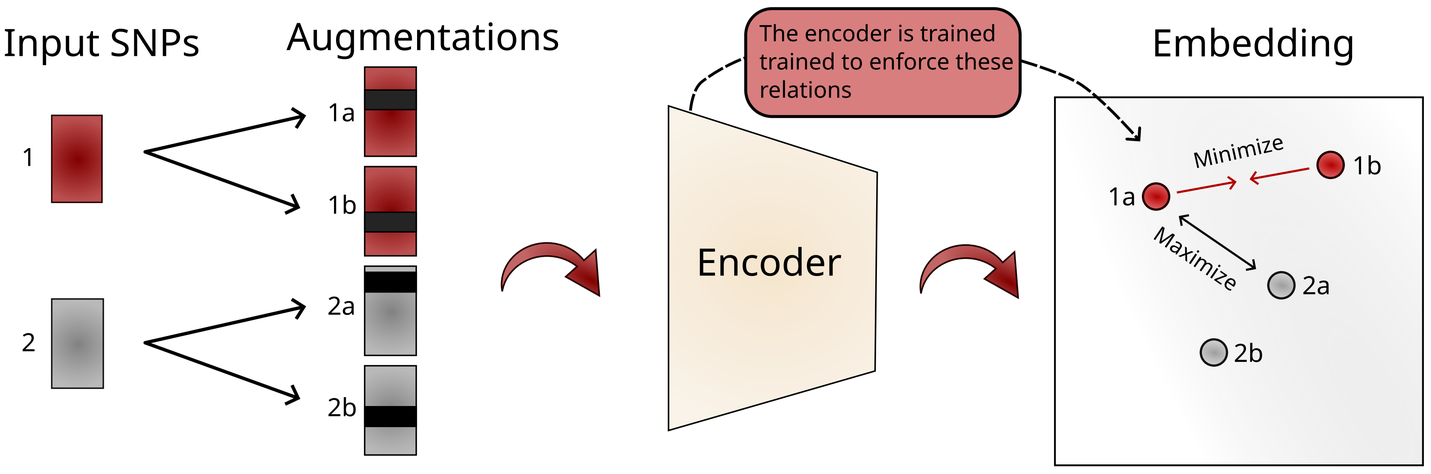
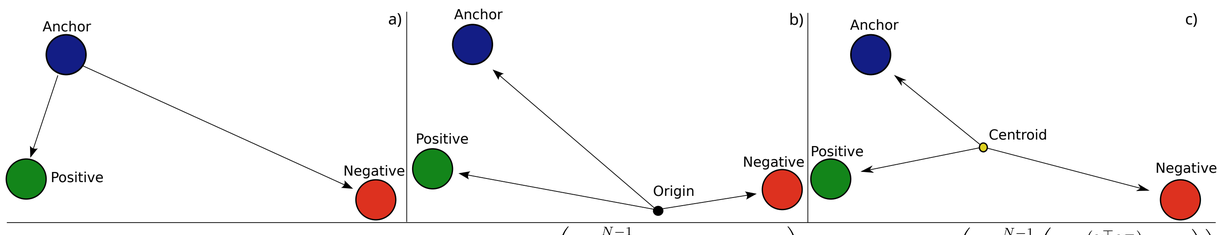


Figure 2. Contrastive learning: an encoder neural network model is trained to embed two slightly altered versions of the same sample closer to each other than other samples. This trains the model to identify characteristic traits that make samples similar, as defined by our augmentation scheme.

Embeddings Without a Genter of Gravity

Many methods produce embeddings with a clear bias towards the origin of the coordinate system (0,0). Some populations may have a group of founders from which others have diverged, however, we do not want a method that a-priori assigns special properties to the origin or any other coordinate in an embedding. We have developed two main model choices to address this.

The model is trained using a centroid-based loss function. We propose to measure distances between samples w.r.t a weighted average of their coordinates, instead of the origin. Figure 3 shows an illustration of how distances are computed in common contrastive losses, and our proposed centroid loss in c).



$$\mathcal{L}_{triplet}(z, z^+, z^-) = \max\left(|z - z^+| - |z - z^-| + \alpha, 0\right) \quad \mathcal{L}_{N\text{-}pair}(Z) = \log\left(1 + \sum_{i=1}^{N-1} \exp(z^\top z_i^- - z^\top z^+)\right) \quad \mathcal{L}_{centroid}(Z) = \log\left(1 + \sum_{i=1}^{N-1}\left(\frac{\exp\left(\bar{z}_i^\top \bar{z}_i^-\right)}{\exp\left(\bar{z}_i^\top \bar{z}_i^+\right)} - e^{-2}\right)\right)$$

Figure 3. Illustration of how samples are compared in the different contrastive loss functions. a): Triplet loss b): N-pair loss c): our proposed centroid-based N-pair loss. A new centroid will be computed and used for each triplet within each batch of samples.

We embed samples on the 3D-sphere, enabling samples to have neighbors in all directions, and use the Equal Earth map projection to visualize the embedding in 2D. Since the sphere can be rotated arbitrarily before projecting, the origin of the projected coordinate system does not carry any meaning. This makes it harder to misinterpret some populations as genetic outliers, or ”different”.

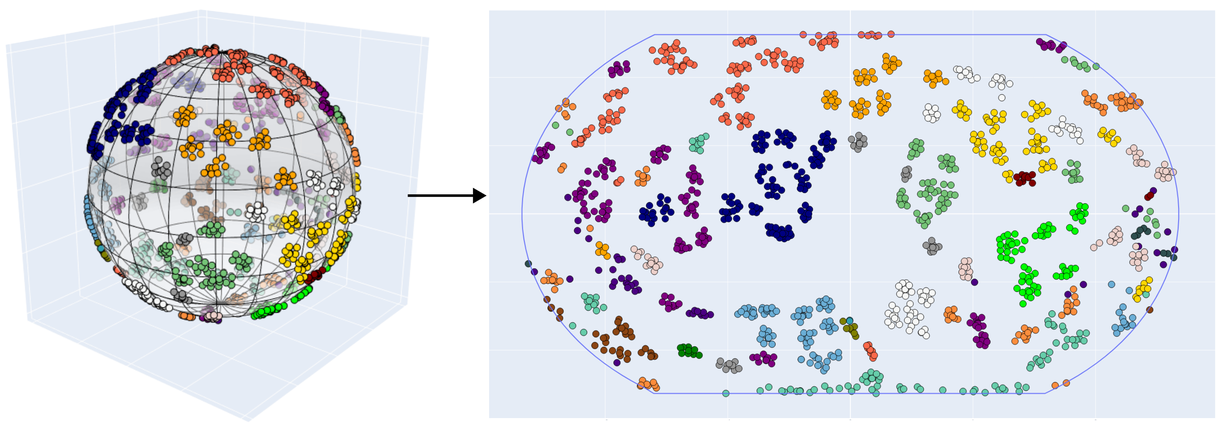


Figure 4. Embedding of the dataset from [3] with 1355 dogs, 150k SNPS on the 3D sphere created using our centroid loss, and the Equal Earth map projection.

Robustness to Missing and Unseen Data

To evaluate the performance of the embeddings, we train a KNN classifier, and report the classification accuracy in Table 1. We compare PCA, t-SNE, triplet, and centroid, where all but t-SNE are trained using 80% of the data and the remaining 20% is kept for validation. This is done both for a dog dataset [3] (1355 samples, 150k SNPs), and the Human Origins dataset [2] (2067 samples, 160k SNPS).

For ”Human masked”, the validation data has been divided into 4 groups, where within each group, the same 20% of markers are set as missing. They are then naively imputed to the most common genotype. This simulates a case with a dense reference set, and some samples genotyped with different SNP chips.

Table 1. Classification performance of the different methods using a KNN classifier with $k = 3$ on the embedding coordinates. All methods use the same 80/20 train/test split for both projection and the classification model, except for t-SNE which needs to use all samples in the projection.

Method	Dataset	Subpop acc.	validation	Superpop acc.	validation
PCA	Dog	0.3572	0.3358	0.5535	0.5351
t-SNE	Dog	**0.9506**	**0.9410**	**0.9852**	**0.9815**
Triplet	Dog	0.7284	0.7897	0.8871	0.9188
Centroid	dog	0.9269	0.9114	0.9661	0.9557
PCA	Human	0.4359	0.4203	0.8771	0.8913
t-SNE	Human	**0.7842**	**0.7609**	**0.9574**	**0.9589**
Centroid	Human	0.5694	0.4758	0.9289	0.9203
PCA	Hu. masked	0.3657	0.0845	0.8157	0.5821
t-SNE	Hu. masked	0.3851	0.2657	0.8171	0.6932
Centroid	Hu. masked	**0.5660**	**0.4565**	**0.9231**	**0.9010**

Robustness to Missing and Unseen Data Cont.

Figure 5 shows embeddings of the Human Origins dataset using contrastive learning with our centroid loss and t-SNE, in the case where some variants of the validation samples are masked. The top plots are labeled by population, and the bottom by the missingness group. t-SNE groups the samples by population, but also by the chip used to genotype the samples, which is unwanted behavior not exhibited by our method.

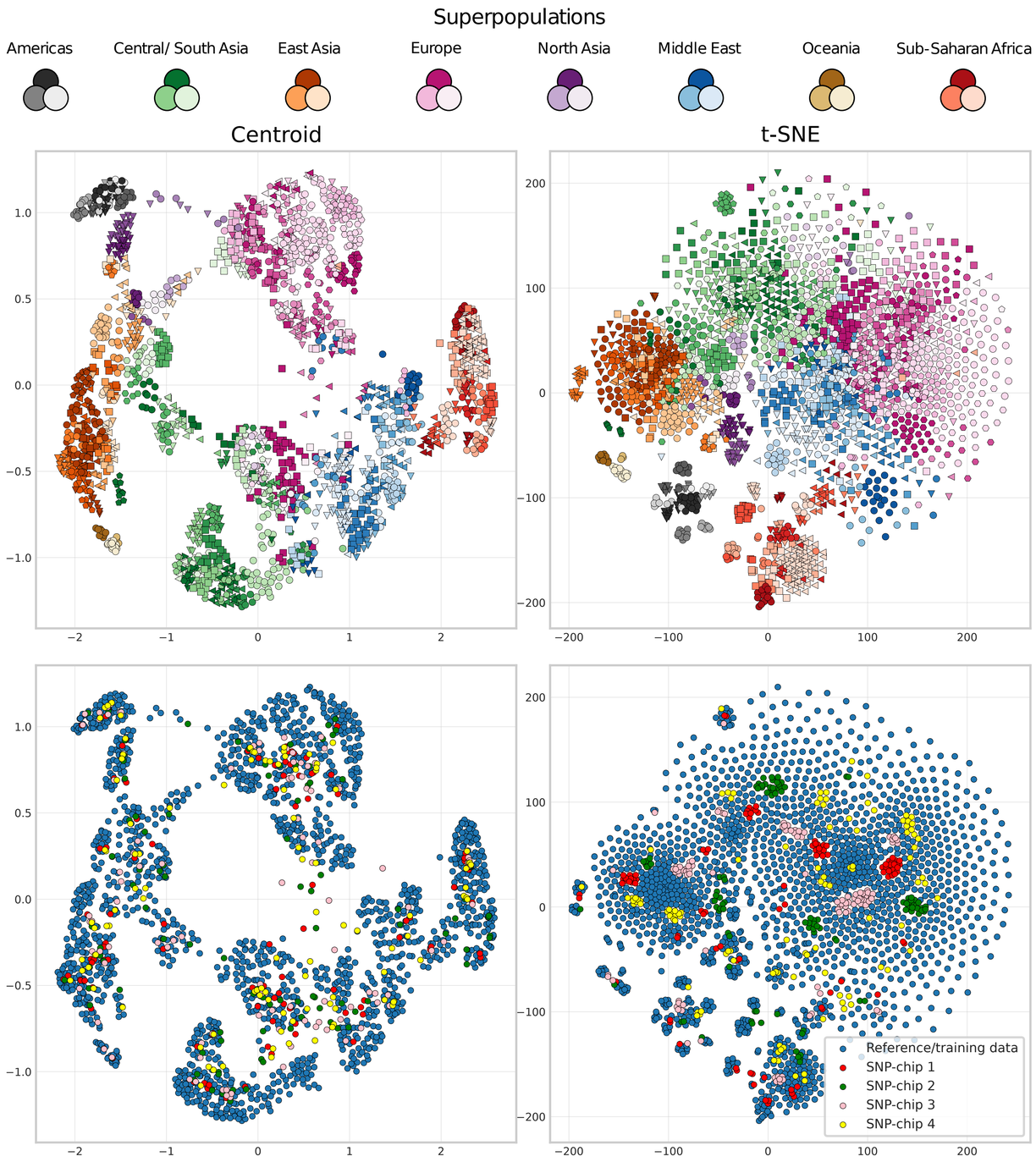


Figure 5. Dimensionality reduction results comparing our method, and t-SNE for the synthetically masked case for the Human Origins dataset. The top figures are labeled by population label, and the bottom by the masking group.

Conclusions and Future Work

The contrastive learning implementation performs slightly worse than t-SNE in peak population classification accuracy. However, we observe good generalization properties, both in terms of samples not seen during training quantified by good validation accuracy, but also on being robust to data with some spurious pattern that we do not want the embedding method to focus on.

We are currently working on training models on larger datasets, e.g. the UK BioBank data with hundreds of thousands of samples, and using our models for genomic prediction and genotype-phenotype mapping.

Acknowledgements and References

Compute resources provided by SNIC through the National Supercomputer Centre (NSC) at Linköping University under Project Berzelius 2022-17, 2022-165, 2023-258, and 2024-121. 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.

[1] 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.

[2] 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.

[3] 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.