Dimension reduction adapted to paleogenomics

Séverine Liégeois, Olivier François, Flora Jay

Dimension reduction adapted to paleogenomics

Séverine Liegeois[1], Olivier François[2], Flora Jay[1]
[1] Laboratoire de recherche en informatique, U. Paris Sud, CNRS
[2] Laboratoire TIMC-IMAG, Grenoble, CNRS

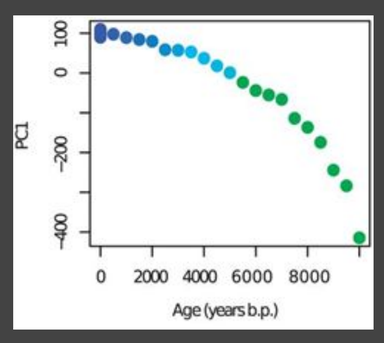
ABSTRACT

Paleogenomics studies genomes of individuals who lived hundreds or thousands years ago, and allows the reconstruction of complex demographic events, such as past migrations or cultural diffusions. However, most statistical tools used in the field do not model the **temporal heterogeneity** of samples.
We present here several methods based on latent variable models to analyze and visualize high dimensional modern and paleogenomic data while **accounting explicitly for time**.

PROBLEMATIC

Genotypic data are generally visualized in a low dimensional representation thanks to **principal component analysis (PCA)** that searches for axes (PCs) along which projected observations show the highest variance.

In the absence of any structure, if samples are collected at different timepoints in a constant size population, the first PC represents the samples on a **gradient** according to their age, and the next PCs exhibit wave patterns [1]. The same phenomenon appears when PCA is applied on spatial data (i.e. samples collected in different geographic locations), and this can **bias interpretations and blur the structure signal**.



While spatial drift is well studied and methods have been proposed to correct for it, temporal drift is still understudied. Thus, we adapted three statistical methods to **model and correct for time effects**.

I. MODELS

tFA (based on [2]):
$G = UV^T + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \Sigma)$

tLFMM (based on [3]) **,**
tRCA (based on [4]):
$G = UV^T + ZW^T + E$

time

columns = corrected axes

The models separate the **time effects** and the hidden **population structure**, which is represented by latent variables in UV^T.
tLFMM is solved by ridge regularization penalizing W, and **tRCA** is solved by a generalized eigenvalue problem.

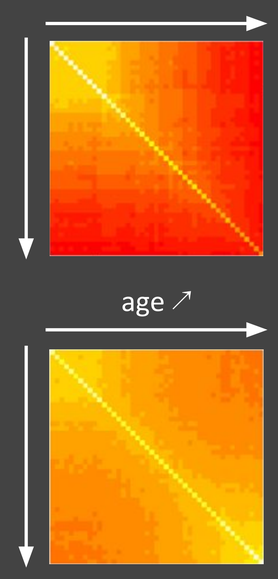
Genotypic data matrix G:

$$G = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 2 \\ 1 & 0 & 0 & 2 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 2 & 0 & 1 & \cdots & 1 \end{bmatrix}$$

L genetic loci (columns) — N samples (rows)

II. COVARIANCE

Time effects are modeled through a custom **temporal covariance matrix**.



Brownian:
$\text{cov}(\text{sample}_i, \text{sample}_j) = \min(\text{age}_i, \text{age}_j)$

Corresponds to the data covariance computed on a raw dataset.

Exponential:
$\text{cov}(\text{sample}_i, \text{sample}_j) = \exp[\ -d(\text{age}_i, \text{age}_j) / \boldsymbol{\theta}\ ]$

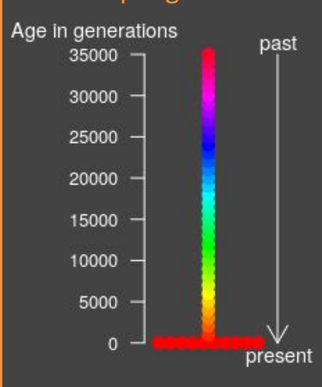
Corresponds to the data covariance matrix computed on a mean-centered dataset.

In the models, **Σ** is the custom temporal covariance matrix and **Z** contains the first *q* eigenvectors of the custom temporal covariance matrix.

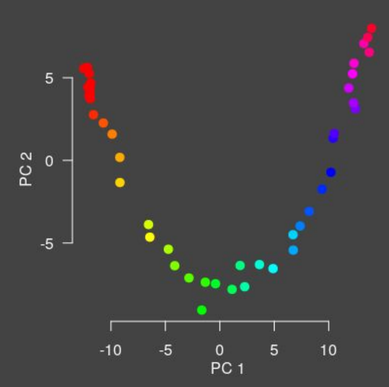
III. RESULTS

The models were tested on different **simulated scenarios**. The best results were obtained with **tLFMM** using the **Brownian** covariance matrix:

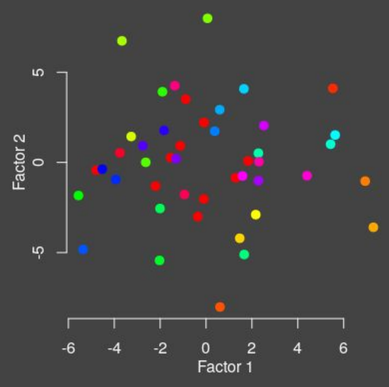
Sampling scheme:



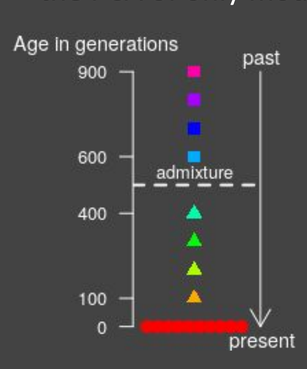
Standard PCA:

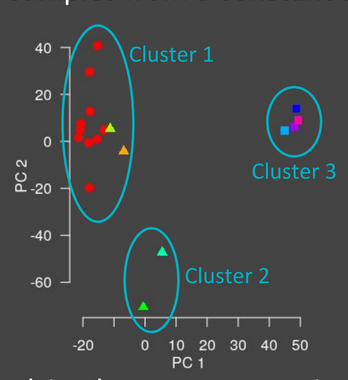


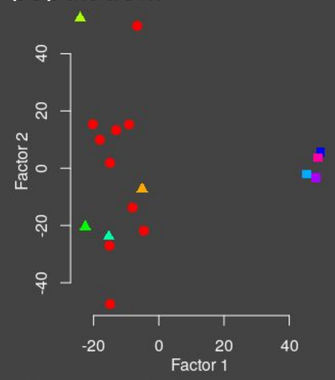
Axes corrected for time:



In the absence of any structure, as expected, **tLFMM Brownian** corrects the pattern due to time in the standard PCA biplot: the result is a point cloud like the PCA of only modern samples from a constant size population.







When admixture occured in the past, correction methods gather admixed individuals and separate them from samples taken before admixture, instead of forming three clusters like the standard PCA.

All 3 methods rely on hyper-parameters that must be chosen beforehand.

IV. MISSING DATA

Because of postmortem DNA degradation, ancient samples usually harbor lots of **missing data** compared to contemporary samples. The methods were tested for different percentages of missing values in the simulated ancient samples.

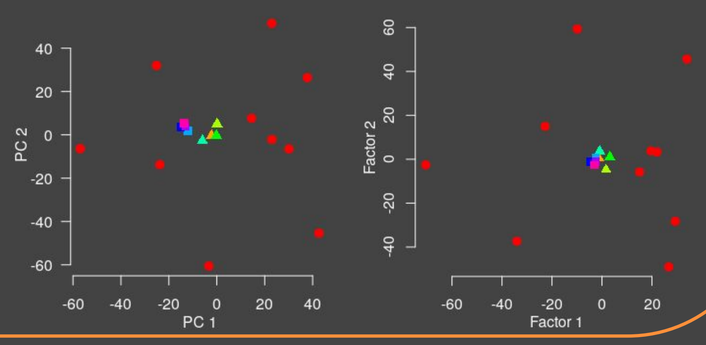
modern samples / ancient samples

$$G = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 2 \\ 1 & 0 & 0 & 2 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 2 & \times & 1 & \cdots & 1 \\ 0 & 1 & 0 & 1 & \cdots & \times \\ 1 & \times & 0 & 1 & \cdots & 0 \end{bmatrix}$$

→ imputation → time correction

As the percentage increases, imputation does not retain the population structure. Thus, temporal drift and structure cannot be separated:

Admixture scenario with 20% of missing values in the ancient samples, after imputation.
Left: standard PCA.
Right: tLFMM Brownian.



CONCLUSION / FUTURE WORKS

- Temporal drift impacts data visualization of time-separated samples. This can be corrected with our three methods in simulated datasets.
- The models are very sensitive to hyper-parameters. They can be estimated by calculating the data reconstruction error by cross-validation, but other metrics need to be found to test the methods and tune the hyper-parameters.
- The proportion of missing data impacts data structure and visualization. Other imputation algorithms need to be tested.
- Models must include both temporal and spatial autocorrelation.

References:
[1] Skoglund, P. *et al*. Investigating population history using temporal genetic differentiation. *Mol Biol Evol* **31**, 2516-2527 (2014).
[2] Frichot, E. *et al.* Correcting principal component maps for effects of spatial autocorrelation in population genetic data. *Front Genet* **3**, (2012).
[3] Caye, K. & Francois, O. LFMM 2.0: Latent factor models for confounder adjustment in genome and epigenome-wide association studies. (2018). doi:10.1101/25589
[4] Kalaitzis, A. A. & Lawrence, N. D. Residual component analysis: Generalising PCA for more flexible inference in linear-Gaussian models. *Proceedings of the 29th International Conference on Machine Learning*, 209-216 (2012).

Acknowledgements:
LRI for funding
Cyril Furtlehner (LRI) for discussions.

Contact:
sliegeois@yahoo.fr (Séverine Liegeois)
flora.jay@lri.fr (Flora Jay)