Evaluation of methods handling missing data in PCA on genotype data: applications for ancient DNA

Kristiina Ausmees
Department of Information Technology
Uppsala University, 752 37, Uppsala, Sweden
Email: kristiina.ausmees@it.uu.se

Abstract

Principal Component Analysis (PCA) is a method of projecting data onto a basis that maximizes its variance, possibly revealing previously unseen patterns or features. PCA can be used to reduce the dimensionality of multivariate data, and is widely applied in visualization of genetic information. In the field of ancient DNA, it is common to use PCA to show genetic affinities of ancient samples in the context of modern variation. Due to the low quality and sequence coverage often exhibited by ancient samples, such analysis is not straightforward, particularly when performing joint visualization of multiple individuals with non-overlapping sequence data. The PCA transform is based on variances of allele frequencies among pairs of individuals, and discrepancies in overlap may therefore have large effects on scores. As the relative distances between scores are used to infer genetic similarity, it is important to distinguish between the effects of the particular set of markers used and actual genetic affinities. This work addresses the problem of using an existing PCA model to estimate scores of new observations with missing data. We address the particular application of visualizing genotype data, and evaluate approaches commonly used in population genetic analyses as well as other methods from the literature. The methods considered are that of trimmed scores, projection to the model plane, performing PCA individually on samples and subsequently merging them using Procrustes transformation, as well as the two least-squares based methods trimmed score regression and known data regression. Using empirical ancient data, we demonstrate the use of the different methods, and show that discrepancies in the set of loci considered for different samples can have pronounced effects on estimated scores. We also present an evaluation of the methods based on modern data with varying levels of simulated sparsity, concluding that their relative performance is highly data-dependent.

I. Introduction

When using Principal Component Analysis (PCA) to illustrate ancient samples in the context of modern variation, genotypes of a set of modern reference individuals together with those of the ancient target sample are transformed to lower dimensions and visualized together, with relative distances interpreted as reflecting genetic similarity. In the case of multiple target individuals, joint analysis is often not straightforward due to low sequence coverage resulting in limited overlap between samples. Since the PCA transform is based on variances of allele frequencies among pairs of individuals, discrepancies in overlap may have large effects on transformed scores. When using PCA to assess quality of imputed data, such effects may have particularly large impact on interpretation. In this case there are multiple data points for each individual and it is important to distinguish effects of projection methodology and actual genetic affinities.

In this report we discuss different methods of handling missing values in PCA, focusing on the particular application of dimensionality reduction to genotype data. We consider methods that have been applied in analyses of genome data previously, as well as additional ones from the PCA literature. The methods of trimmed scores (TRI) [1] and projection to the model plane (PMP) [2] correspond to two options available for dealing with missing genotypes offered by the software `SMARTPCA` from `EIGENSOFT 7.2.1` [3, 4], which is widely applied in the field of genomics. A method that has seen use in the field of aDNA is based on performing separate PCAs for each target sample based on the markers overlapping the modern reference, and subsequently merging the individual projections using Procrustes analysis (PRCR) [5, 6]. In addition to these, we consider the methods of trimmed score regression (TSR) and known data regression (KDR) introduced in [1]. The use of the methods is demonstrated on empirical ancient data from [7], in which PCA was used to evaluate imputation performance. We also compare their estimation error on four panels of modern panels of genotype data.

II. Data

We consider four panels of human genotype data, summarized in Table I, for illustration and evaluation of methods in this report. For all data sets, we used called diploid genotypes, filtered to only include biallelic single-nucleotide polymorphisms (SNPs). Rare variants were removed based on minor allele frequency (MAF), and highly correlated variants were further removed based on allelic r^2, the squared correlation coefficient measure of linkage disequilibrium (LD) between two loci [8]. For pairs of SNPs within a certain window size, with an r^2 value greater than a threshold, one member of the pair was removed. The pre-processing is intended to reflect that which is common to perform in practical applications, as the presence of rare variants and LD can have an impact on the interpretation of PCA results, discussed in e.g. [9, 10].

Panel	Samples	Markers
1	429	115659
2	2067	160858
3	2548	165434
4	2548	15757

TABLE I

NUMBER OF INDIVIDUALS AND MARKERS OF THE FOUR REFERENCE PANELS OF MODERN DATA. PANEL 1 WAS USED TO DEFINE A PCA MODEL IN THE DEMONSTRATION OF THE METHODS ON EMPIRICAL ANCIENT DATA IN SECTION III. VALIDATION SETS WERE GENERATED FOR EACH PANEL AND USED FOR EVALUATION OF ESTIMATION ERROR OF THE DIFFERENT METHODS IN SECTION IV.

		Markers overlapping reference		
Legend	Individual	High quality	Imputed	Low coverage
	ans17	114727	97719	37363
	sf12	115596	95863	38013
	LBK	73812	95409	35500
	Lochbour	77256	95360	38385
	ne1	91506	94748	37567

TABLE II

THE ANCIENT INDIVIDUALS USED FOR DEMONSTRATION OF THE METHODS. FOR EACH INDIVIDUAL THERE WERE THREE SAMPLES CONSISTING OF HIGH QUALITY, IMPUTED AND LOW COVERAGE GENOTYPES. FOR EACH SAMPLE, THE NUMBER OF MARKERS OVERLAPPING THE MODERN REFERENCE USED IN PCA (PANEL 1) IS GIVEN. THE INDICATED COLORS SERVE AS A LEGEND FOR ALL PLOTS IN SECTION III.

Panel 1 consisted of 429 samples from four European populations from the Affymetrix Human Origins data set [11]. A MAF threshold of 1% was enforced, and a window of genetic distance of 1 centimorgan was used with an r^2 threshold of 0.2 for LD pruning, resulting in a total of 115659 markers. The second panel was based on the fully public Affymetrix Human Origins modern samples of [12], with the same filtering procedure as described above, containing a total of 160858 markers. Comprising 2067 individuals from 166 populations around the world, the samples in this set exhibit a larger degree of genetic diversity than Panel 1. The final two panels were based on the 1000 Genomes [13] release 2019031 set of called genotypes from chromosome 20. With 2548 samples from 26 populations, it also represents a set of world-wide genetic variation. Unlike panels 1 and 2, which derive from genotyping arrays, the 1000 Genomes data is a result of whole-genome sequencing technologies. Genotypes were filtered for a minimum MAF of 5%, resulting in 165434 markers for Panel 3. To generate Panel 4, LD pruning was also performed with an r^2 threshold of 0.2 in a window of 100 kilobases, resulting in 15757 markers.

A. Demonstration data

For illustrating the use of the methods explained in Section III, we consider the same scenario and data as in [7], in which ancient samples were projected onto a PCA model defined by a modern reference panel. Panel 1 described above was used as the reference, and the samples to be projected were the five ancient individuals ans17, sf12, LBK, Loschbour and ne1. For each ancient individual, there were three samples: high-quality (HQ), low-coverage and imputed. The HQ genotypes were called from high-coverage data, and subsequently filtered to keep only those with a high probability of being correct. The low-coverage genotypes were called from read data that had been downsampled to 1x coverage. This data was also used as input to imputation, which yielded the third sample for each individual. Table II shows the number of markers which overlapped those of Panel 1 for each of the three samples per individual. We refer to [7] for further details about the ancient individuals and how the sample data was generated.

B. Evaluation data

The data used for evaluation of methods in Section IV was generated from the four panels of modern individuals in Table I. For each panel, a validation set was defined by randomly selecting a number of individuals to remove, stratified by population. Sparse genotypes of the validation samples were subsequently simulated by randomly and incrementally removing
10, 20, ..., 90 percent of the data. This procedure was repeated 10 times, resulting in 10 validation sets per panel. For Panel 1, the size of each validation set was 10 samples, and for panels 2,3 and 4, 20% of the total number of samples were selected.

III. METHODS

We use the notation from [1] and [2] in which a PCA model is described as:

$$X = TP^T$$

where X is the $n \times k$ data matrix of n observations (samples) and k variables (markers), P is the $k \times k$ loadings matrix and T is the $n \times k$ score matrix. The columns of the loadings matrix are the eigenvectors of the covariance matrix of X, ordered by decreasing magnitude of eigenvalue. The scores matrix contains the projected coordinates of the samples of X. Matrices are denoted by upper-case letters and column vectors by lower-case letters, and the notation $P_{1:A}$ is used to describe the matrix consisting of the A first columns of P.

Standard PCA finds a solution to

$$\min_{P_{1:A}} ||X - T_{1:A}P_{1:A}^T||^2$$

where $||\cdot||$ denotes the Frobenius norm, $T_{1:A} = XP_{1:A}$ and $P_{1:A}^T P_{1:A} = I$. This formulation expresses the minimization of the reconstruction error when using A dimensions to represent the data and is equivalent to finding a projection of the data X to A dimensions that maximizes the variance of the projected data. In the case of two-dimensional visualization, $A = 2$.

For handling missing data in PCA, we consider the problem of estimating scores τ for a new observation z, from an existing PCA model based on X. In our application, X is the panel of modern samples and z is an ancient sample. The observation z may have missing values, with z^* and $z^\#$ respectively denoting the vector taken at only the indices where the value is observed and missing. Accordingly, P^* then denotes the matrix P takes only at the rows corresponding to indices of z^*. With τ denoting the scores based on the complete data of an observation, and $\tilde{\tau}$ an estimate based on the incomplete data, the estimation error is measured by the difference $\tilde{\tau} - \tau$.

In the remainder this section, the methods for handling missing data in PCA considered in this report are defined. We demonstrate the different methods on the data described in Section II-A. Unless otherwise stated, we assume that the data matrix X has been standardized by mean-centering and scaling to unit variance, and that the sample data z has been standardized accordingly. All methods were implemented in Python, available in the package `mpca` [1].

A. Trimmed scores (TRI) and projection to the model plane (PMP)

A simple method of handling unobserved data is to fill in missing variables with their unconditional mean value, i.e. zero, and perform the usual PCA projection step:

$$\tau_{1:A} = {P_{1:A}}^T z$$

which results in the estimator

$$\hat{\tau}_{1:A} = {P^*_{1:A}}^T z^*$$

of the TRI method. This method entails only taking into account loadings that relate to variables that are observed, and corresponds to the usual least squares estimator, assuming $z^\# = 0$. The estimated scores are referred to as the trimmed scores.

Another option is to consider the PCA model expressed only in terms of observed variables:

$$z^* = P^*_{1:A}\tau_{1:A}$$

The least-squares solution to the equation above corresponds to the PMP estimator

$$\hat{\tau}_{1:A} = ({P^*_{1:A}}^T P^*_{1:A})^{-1} {P^*_{1:A}}^T z^*$$

The estimation errors for the TRI and PMP methods are derived in [1, 2]. While errors due to loss of orthogonality of the loading vectors occur in both models, the TRI method can result in large estimation errors when data corresponding to large loadings is missing. As it may be the case that there is a relatively small subset of SNPs that are very influential in the PCA, this can potentially result in large discrepancies of estimated scores based on non-overlapping data. Figure 1 shows the results of using the TRI and PMP methods on the demonstration data. The TRI method yielded significantly larger discrepancies in scores between the high quality, imputed and low-coverage data for each sample than the PMP method, suggesting that it indeed seems to be more sensitive to missing data.

[1] https://github.com/kausmees/mpca

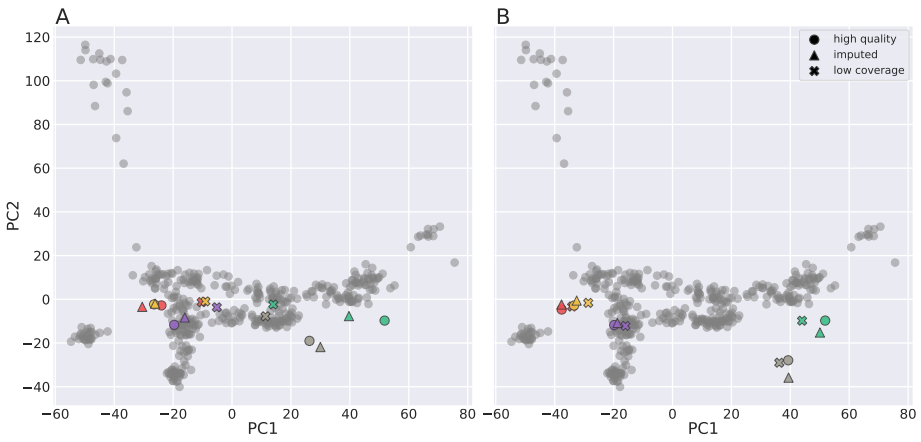

Fig. 1. Comparison of the TRI and PMP methods for projecting ancient samples with missing data to a PCA model defined by Panel 1. Plot A shows the TRI method, which imputes missing genotypes with the unconditional mean value 0. Plot B shows results using the PMP method, which estimates the scores that minimize the sum of the squared residuals of the PCA reconstruction, considering only observed genotypes.

B. Trimmed score regression (TSR) and known data regression (KDR) methods

The TSR method is similar to TRI in that it is based on the trimmed scores obtained from setting unobserved values $z^{\#}$ to the unconditional mean value 0, but it involves an additional step in which the trimmed scores of X are used to further improve the score estimator for z by least squares regression. With X^* representing the data matrix X taken at the indices of observed data z^*, the trimmed scores of X are $T^*_{1:A} = X^* P^*_{1:A}$. A regression model is defined to find a transformation B that maps the trimmed scores to the reference scores by solving the least-squares problem:

$$\min_B ||T^*_{1:A} B - T_{1:A}||^2$$

and applying the same transformation to the trimmed scores $\tau^*_{1:A} = {P^*_{1:A}}^T z^*$ results in the TSR estimator:

$$\hat{\tau}_{1:A} = \hat{B}^T \tau^*_{1:A}$$

In the KDR method, a similarity transform C is also found by solving a regression problem, but instead of using the trimmed scores of the reference data X, it is based on the overlapping data X^* itself:

$$\min_C ||X^* C - T_{1:A}||^2$$

Thus, while the TSR method finds a transformation that maps scores resulting from partial data to the reference scores $T_{1:A}$, the KDR method can be interpreted as estimating the transformation matrix that yields the reference scores based on the rows of X that correspond to observed genotypes in z. Application of the transformation to the observed data z^* yields the KDR estimator:

$$\hat{\tau}_{1:A} = \hat{C}^T z^*$$

C. Individual PCA and Procrustes transform

A method that is commonly used in the field of ancient DNA is based on performing separate PCAs for each target sample, based on the SNPs overlapping the modern reference, and subsequently merging the individual projections using Procrustes analysis. For an ancient sample z with observed data z^*, the individual PCA is performed based on X^*: the modern panel taken at indices corresponding to observed genotypes of z. This PCA model can be expressed as

$$X^* = \tilde{T}\tilde{P}^T$$

and the scores of X^* and z in this model are thus given by $\tilde{T}_{1:2} = X^* \tilde{P}_{1:2}$ and $\tilde{\tau}_{1:2} = \tilde{P}^T_{1:2} z^*$. Multiple individual projections can subsequently be merged by matching the scores of the individual PCA to the scores $T_{1:2}$ of a reference PCA performed on X. This is done by finding a transform consisting of rotation, reflection, translation and uniform scaling that reduces the squared distance between the two sets of points, with the row corresponding to the ancient sample in the reference given the scores $[0, 0]^T$. As we assume that both sets of points are centered, the translation can be skipped and the problem expressed as:

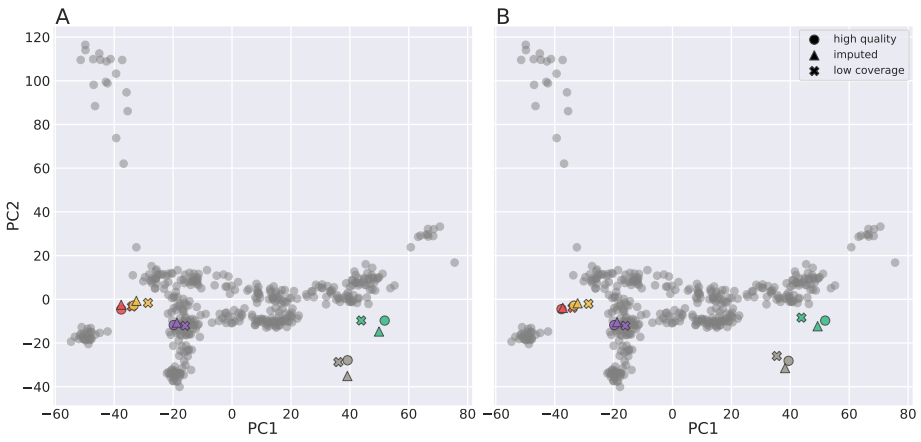

Fig. 2. A: TSR and B: KDR.

$$\min_{Q,\alpha} \left\| \alpha \begin{bmatrix} \tilde{T}_{1:2} \\ \tilde{\tau}_{1:2} \end{bmatrix} Q^T - \begin{bmatrix} T_{1:2} \\ 0\ 0 \end{bmatrix} \right\|^2$$

where α is an isotropic scaling factor and Q is an orthogonal matrix representing rotation and reflection ($Q^TQ = I$). The problem of finding optimal values $\hat{\alpha}, \hat{Q}$ has a closed form solution, e.g. [14], and the resulting transformed scores $\hat{\tau}_{1:2} = \hat{\alpha}\hat{Q}\tilde{\tau}_{1:2}$ are used for each ancient sample. For the modern reference samples, it is common to display the average of transformed scores $\hat{T}_{1:2} = \hat{\alpha}\tilde{T}_{1:2}\hat{Q}^T$ over the individual transformations. This is also the method used in this report. Figure 3 shows the results of using the method of individual PCA and Procrustes transformation

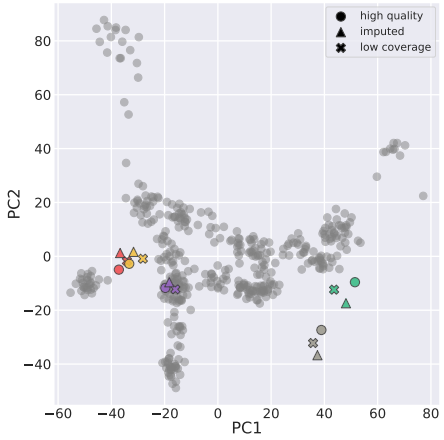

Fig. 3. Projection of the demonstration samples using the method of individual PCA and Procrustes transformation.

In order to illustrate the effects of the set of markers that overlaps the reference, we consider an alternative scenario in which the same set of markers is used for the different samples for each individual. We exclude the low coverage data for this example due to its sparsity, and intersect the imputed and high quality markers for each individual prior to performing PCA and Procrustes transformation. Figure 4 shows the results for the strategies of using all available markers (A) and intersected markers (B). It is visible that the intersection has had a large effect, with much larger differences in scores in the case where distinct sets of markers were used for the two samples for one individual. Thus, some of the discrepancies between scores of HQ and imputed data can be explained by the effects of markers used in the PCA. We exemplified this on the method of individual PCA and Procrustes transform since it is commonly used for aDNA, but note that the effects are similar for all methods considered.

IV. Evaluation

In order to evaluate the performance of the methods, data with simulated missing genotypes was used. As described in Section II-B, validation sets were generated for each of the four panels of modern individuals by selecting a group of samples

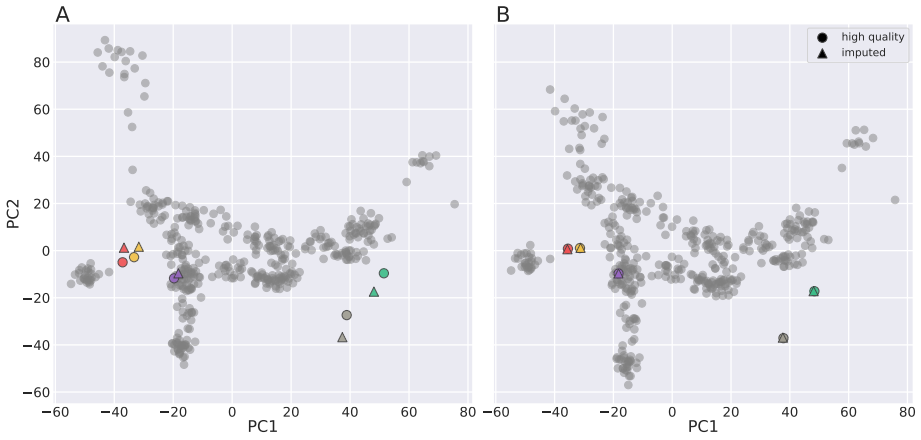

Fig. 4. Comparison of two methods of merging data when illustrating ancient samples in the context of modern variation using the method of individual PCA and Procrustes transformation. Plot A shows the results of individually intersecting every sample with the modern panel, and plot B when the two data sets for an individual are first intersected, resulting in identical sets of SNPs for the imputed and HQ data for an individual.

for which SNPs were randomly and incrementally set to missing. The remaining individuals were used to define a PCA model, and scores were subsequently estimated using the different methods. Error of estimated scores for a sample was calculated w.r.t. the PCA scores obtained when using the complete data for that sample. For each validation set and missingness level, the mean squared error (MSE) over the validation samples was calculated, and we report the average MSE over the ten validation sets. As the magnitudes of the PCA scores vary between the different panels, we do not address the absolute sizes of the errors in estimated scores, and only focus on comparing the relative performance of the methods on a given panel.

Figures 5-8 show the average and standard deviation of the MSE, in log scale, for Panels 1-4, respectively. In each case, plot A shows error for all methods on the entire range of missing data, and Plot B a zoomed-in version with the level of removed markers in the range 10% to 50%. The average MSE is also shown in Tables III - VI. For all four panels, the most simple strategy of TRI gave significantly higher errors than the other methods. Likewise, the methods of PMP and TSR had very similar performance for all experiments. For the remaining methods, differences in behaviour were visible for the different panels. For Panel 1, all methods excluding TRI exhibited similar performance, although it is visible in Plot B that the PRCR method had higher errors, and that the KDR estimator slightly outperformed the others for lower levels of missing data. For Panel 2, the PRCR method led to the lowest error, and KDR resulted in slightly worse performance than the other regression-based methods, particularly in the lower ranges of missing data. The results for Panel 3 had a similar pattern to those of Panel 1, although the advantage of the KDR method was particularly pronounced, and consistently so throughout the range of missing markers. For Panel 4, much of the differences between methods was removed. The PMP and TSR methods outperformed the others, and the KDR method seemed to be particularly sensitive to increasing levels of removed markers.

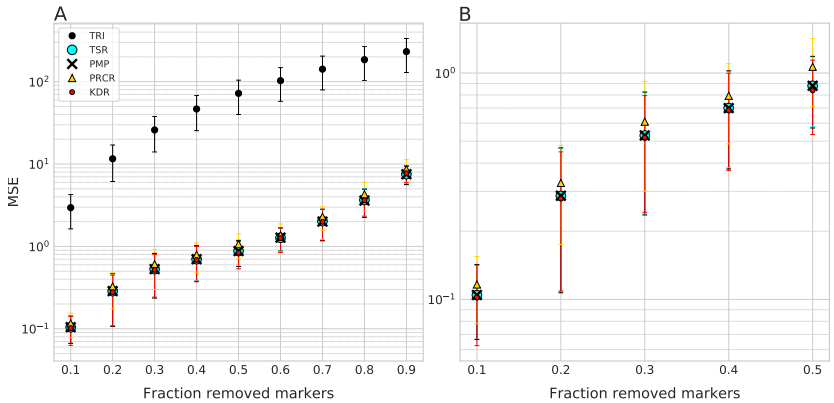

Fig. 5. Mean-squared error in log scale of estimated scores for the different methods applied to validation data derived from Panel 1. Mean and standard deviation of MSE over 10 validation sets is shown for different levels of removed markers. Plot B is a zoomed-in version of Plot A, focusing on the lower range of removed markers.

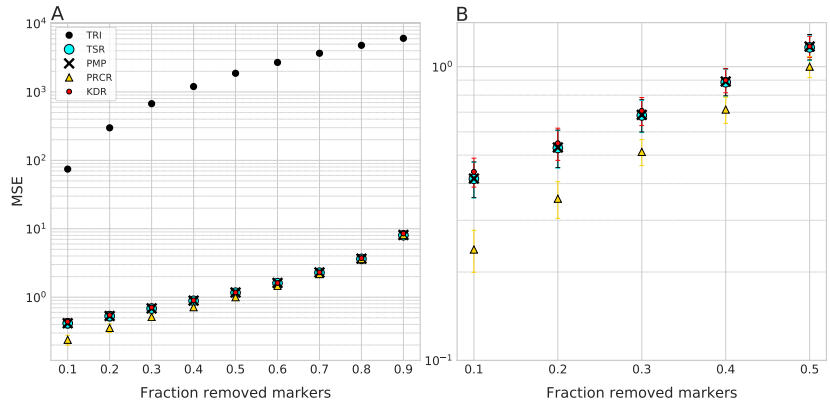

Fig. 6. Mean-squared error in log scale of estimated scores for the different methods applied to validation data derived from Panel 2. Mean and standard deviation of MSE over 10 validation sets is shown for different levels of removed markers. Plot B is a zoomed-in version of Plot A, focusing on the lower range of removed markers.

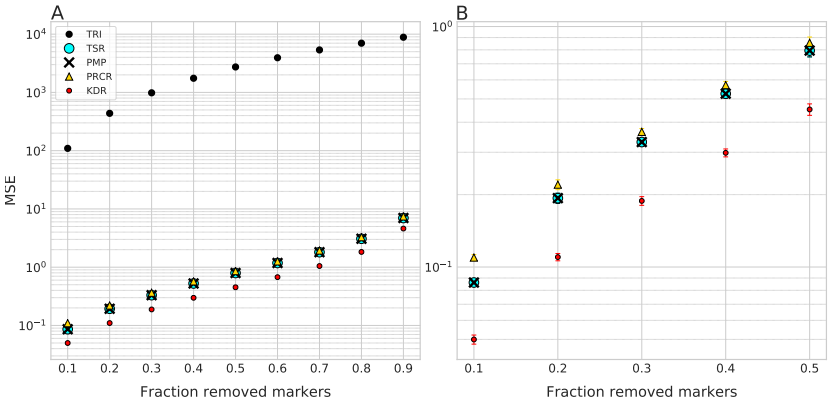

Fig. 7. Mean-squared error in log scale of estimated scores for the different methods applied to validation data derived from Panel 3. Mean and standard deviation of MSE over 10 validation sets is shown for different levels of removed markers. Plot B is a zoomed-in version of Plot A, focusing on the lower range of removed markers.

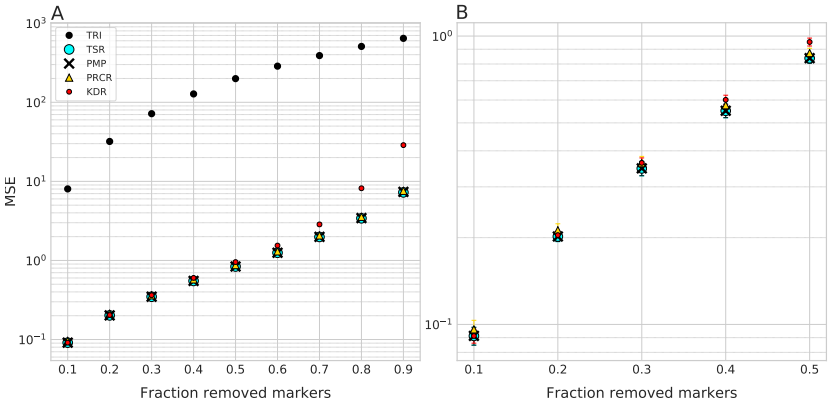

Fig. 8. Mean-squared error in log scale of estimated scores for the different methods applied to validation data derived from Panel 4. Mean and standard deviation of MSE over 10 validation sets is shown for different levels of removed markers. Plot B is a zoomed-in version of Plot A, focusing on the lower range of removed markers.

	Fraction missing markers								
method	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
TRI	2.9556	11.5965	25.9498	46.6305	72.1931	102.9136	141.9701	185.0642	231.8197
TSR	0.1045	0.2866	0.5299	0.7002	0.8784	1.2733	2.0288	3.6615	7.5438
PMP	0.1046	0.2870	0.5301	0.7000	0.8786	1.2787	2.0021	3.6059	7.5155
PRCR	0.1164	0.3268	0.6098	0.7933	1.0658	1.3663	2.2960	4.3023	8.6122
KDR	0.1020	0.2787	0.5186	0.6849	0.8380	1.2603	1.9942	3.6276	7.7415

TABLE III
AVERAGE MEAN-SQUARED ERROR OF ESTIMATED SCORES FOR DIFFERENT LEVELS OF MISSING MARKERS ON PANEL 1.

	Fraction missing markers								
method	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
TRI	74.3457	298.5430	673.1185	1197.2966	1872.0528	2696.0904	3670.5408	4794.9632	6068.3306
TSR	0.4218	0.5368	0.6877	0.8930	1.1720	1.6159	2.2818	3.6502	8.0971
PMP	0.4227	0.5388	0.6910	0.8982	1.1797	1.6272	2.2986	3.6782	8.1762
PRCR	0.2427	0.3609	0.5181	0.7244	1.0105	1.4711	2.1821	3.6242	8.1259
KDR	0.4430	0.5547	0.7097	0.9038	1.1771	1.6169	2.3006	3.7302	8.5851

TABLE IV
AVERAGE MEAN-SQUARED ERROR OF ESTIMATED SCORES FOR DIFFERENT LEVELS OF MISSING MARKERS ON PANEL 2.

	Fraction missing markers								
method	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
TRI	109.5756	437.4337	984.3316	1749.8383	2733.5829	3935.5569	5356.4348	6996.3098	8853.9575
TSR	0.0862	0.1933	0.3308	0.5256	0.7944	1.1803	1.8093	3.0756	6.9787
PMP	0.0862	0.1934	0.3309	0.5257	0.7945	1.1808	1.8103	3.0782	6.9932
PRCR	0.1095	0.2201	0.3648	0.5709	0.8560	1.2584	1.9333	3.2702	7.4281
KDR	0.0500	0.1100	0.1883	0.2984	0.4521	0.6747	1.0475	1.8278	4.6060

TABLE V
AVERAGE MEAN-SQUARED ERROR OF ESTIMATED SCORES FOR DIFFERENT LEVELS OF MISSING MARKERS ON PANEL 3.

	Fraction missing markers								
method	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
TRI	8.0286	31.9769	71.7534	127.5397	199.1566	286.4095	389.6636	508.8017	643.7422
TSR	0.0912	0.2018	0.3475	0.5505	0.8377	1.2497	1.9825	3.4123	7.2545
PMP	0.0913	0.2018	0.3478	0.5508	0.8382	1.2514	1.9881	3.4350	7.3679
PRCR	0.0961	0.2122	0.3623	0.5755	0.8731	1.3026	2.0702	3.5610	7.5925
KDR	0.0912	0.2041	0.3633	0.6009	0.9537	1.5423	2.8557	8.1802	28.7300

TABLE VI
AVERAGE MEAN-SQUARED ERROR OF ESTIMATED SCORES FOR DIFFERENT LEVELS OF MISSING MARKERS ON PANEL 4.

V. Discussion

Five methods are considered in this report. TRI and KDR can be described as performing imputation of unobserved SNPs. In the former, missing variables are replaced with the unconditional mean of the data used to define the PCA model. The KDR method is shown in [1] to be equivalent to the conditional mean replacement method in [2], in which missing variables $z^{\#}$ are imputed with the conditional mean, assuming z follows a multivariate normal distribution with mean vector $E(z) = 0$ and covariance matrix $S = covar(X)$. In the PMP method, no value is given to missing data, and the PCA model is re-expressed in terms of the observed variables and their corresponding loadings only. The TSR method is based on using the effects of the missing variables of z to the samples used to define the PCA, using the trimmed scores of X to inform the estimation of scores for z. The PRCR method is conceptually similar to the TSR method. The main differences lie in the fact that in the former, a new PCA model is defined based on observed data, rather than using the loadings corresponding to them from the original PCA model. Also, the Procrustes transformation is a constrained least squares, only allowing rotation and isotropic scaling, whereas in the TSR method, a general affine transformation is performed.

The results of method evaluation were to a large extent consistent with those found in other experiments, with TRI resulting in significantly larger score errors and the other methods showing more similar behaviour. In [1], evaluation on three industrial data sets showed that the KDR method was statistically superior to the others, although they noted that the TSR method was practically equivalent. They also found that the PMP method gave similar performance in most cases, but maintained that it did have worse performance than the other regression-based methods overall. In [2], they found that all methods considered behaved similarly for lower levels of missing data, but found that conditional mean replacement, which is equivalent to KDR, resulted in lowest error when critical combinations of variables were missing.

The experiments in this report, however, did not show that the KDR method was superior throughout, as it was outperformed by other methods for panels 2 and 4. One possible explanation for this may be numerical errors due to the properties of the matrices used in the calculations. The KDR method requires the inversion of a $(k - m) \times (k - m)$ matrix, where k is the number of variables, and m is the number of unobserved ones, whereas the PMP and TSR methods only require the inversion of a $A \times A$ matrix. As noted in both [1] and [2], the regression-based methods are sensitive to ill-conditioning due to missing data, and may require regularized algorithms. In our experiments, the KDR method was the only one for which ridge regression gave improved performance, suggesting it might indeed have suffered from ill-conditioning. All results shown for KDR are using the regularized method.

Some of the discrepancies in results compared to those of other studies could also be explained by the unique properties of the data considered here. The covariance structure of X affects how missing variables, particularly combinations of them, influence the PCA. This is illustrated by the fact that the LD pruning resulted in large differences in performance for panels 3 and 4. The data in this report also differs from the industrial data sets considered in [1] and [2] because it has much more variables than samples. This could be a possible explanation to the fact that we saw highly similar performance for the PMP and TSR methods throughout, although the TSR method did have slightly lower error in most cases. The PRCR method was generally outperformed by the regression-based methods, with the exception of the case of Panel 2. For levels of missing markers over 20% for panel 4, the KDR method also gave higher errors than PRCR.

VI. Conclusion

This report addresses the problem of estimating scores for samples with partially unobserved data, given an existing PCA model. We focus on the particular application of performing PCA on genotype data, and use empirical ancient data as a demonstrative example. In this scenario, a panel of modern samples is used to define a PCA model, and ancient samples that have partially unobserved data are projected onto it. The empirical examples illustrate that the effects of inconsistent marker sets can be significant. Intersection of genotypes prior to analysis may reduce bias due to differences in overlap with the modern panel, but this approach suffers from the obvious downside of disregarding information. In the context of aDNA, where sparsity of data is a common issue, the use of efficient methods for handling missing data in PCA is thus of particular importance.

Overall, our results show that the relative performance of the methods is highly data dependent. Drawing general conclusions is not straightforward based on the limited experiments in this report, and more work would be required to identify the effects of different properties of the data on the efficacy of the methods. The PMP method corresponds to the `lsqproject` option of the widely used software SMARTPCA, and our results indicate that this is preferable over the default option, which implements the TRI method. The use of the KDR method when using PCA to evaluate imputation performance in [7] was motivated by the results on Panel 1, which was used as the modern reference in that application. If such panel-specific experiments cannot be performed prior to selecting a method, our results suggest that the PMP and TSR methods are two options with consistent relative performance for different data sets, with the TSR method giving slightly lower errors in almost all experiments performed.

References

[1] Francisco Arteaga and Alberto Ferrer. "Dealing with missing data in MSPC: several methods, different interpretations, some examples". In: *Journal of Chemometrics* 16.8-10 (2002), pp. 408–418.

[2] Philip R.C. Nelson, Paul A. Taylor, and John F. MacGregor. "Missing data methods in PCA and PLS: Score calculations with incomplete observations". In: *Chemometrics and Intelligent Laboratory Systems* 35.1 (1996), pp. 45–65.

[3] Nick Patterson, Alkes L. Price, and David Reich. "Population structure and eigenanalysis". In: *PLoS Genet* 2.12 (Dec. 2006), pp. 1–20.

[4] Alkes L. Price et al. "Principal components analysis corrects for stratification in genome-wide association studies". In: *Nature Genetics* 38 (July 2006), pp. 904–909.

[5] Chaolong Wang et al. "Comparing spatial maps of human population-genetic variation using Procrustes analysis". In: *Stat Appl Genet Mol Biol* 9.1 (Jan. 2010), pp. 13–13.

[6] Pontus Skoglund et al. "Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe". In: *Science* 336.6080 (2012), pp. 466–469.

[7] Kristiina Ausmees et al. *An Empirical Evaluation of Genotype Imputation of Ancient DNA*. Tech. rep. 2019-008. Department of Information Technology, Uppsala University. Oct. 2019.

[8] W. G. Hill and Alan Robertson. "Linkage disequilibrium in finite populations". In: *Theoretical and Applied Genetics* 38.6 (June 1968), pp. 226–231.

[9] Iain Mathieson and Gil McVean. "Differential confounding of rare and common variants in spatially structured populations". In: *Nature genetics* 44.3 (Feb. 2012), pp. 243–246.

[10] Fei Zou et al. "Quantification of population structure using correlated SNPs by shrinkage principal components". In: *Human heredity* 70.1 (2010), pp. 9–22.

[11] Nick Patterson et al. "Ancient admixture in human history". In: *Genetics* 192.3 (Nov. 2012), pp. 1065–1093.

[12] Iosif Lazaridis et al. "Genomic insights into the origin of farming in the ancient Near East". In: *Nature* 536 (July 2016), pp. 419–424.

[13] The 1000 Genomes Project Consortium. "A global reference for human genetic variation". In: *Nature* 526 (Sept. 2015), pp. 68–74.

[14] J.C. Gower and G.B. Dijksterhuis. *Procrustes Problems*. Oxford statistical science series. Oxford University Press, 2004.