Statistical Methodology: Principal Component Analysis for Genomic Data Analysis

Mohammad Shahnewaz Morshed

March 23, 2024

1 Methodology

1.1 Problem Formulation

Genomic datasets obtained from whole-genome sequencing (WGS), reduced-representation sequencing (RAD-seq), or transcriptomic assays (RNA-seq) are intrinsically high-dimensional, often comprising tens of thousands to millions of features such as single nucleotide polymorphisms (SNPs), short k-mers, or gene expression profiles. These data exhibit strong correlation structures, including linkage disequilibrium among SNPs (where nearby genetic variants tend to be inherited together; (10)) and coordinated expression patterns among genes (where groups of genes are expressed together due to shared regulation; (3)). Such dependencies create redundancy in the predictor space, meaning that many genomic features carry overlapping information. In this setting, correlated variables inflate certain directions of variation and make it difficult to separate the unique contribution of each feature. As a result, the effective dimensionality of the dataset is reduced, but conventional models often assume predictors are independent. This mismatch leads to instability in parameter estimates and reduces the reliability of inference. In regression, redundancy appears as multicollinearity, where coefficients become unstable and highly sensitive to small changes in the data. In clustering or distance-based methods, redundancy distorts similarity measures by overweighting correlated variables, which can obscure the true biological structure.

When analyzed with conventional regression or clustering methods, these properties introduce several statistical problems. Multicollinearity makes the covariance matrix of predictors nearly singular, which leads to unstable coefficient estimates and inflated variances ((2)). Hidden correlations among variables can also distort standard errors, increasing Type I error rates and creating spurious associations. As the number of predictors increases relative to the number of samples, distances between observations become less meaningful, clusters lose separability, and noise begins to dominate the signal. High-dimensional asymptotic results, such as the Marchenko–Pastur law (which gives the expected distribution of eigenvalues in large random covariance matrices; (6)), demonstrate that artificial components may arise when the ratio of predictors to samples (p/n) is very high. Moreover, differences in measurement scale can skew covariance-based analyses, as variables with larger variances tend to dominate the decomposition and mask the contribution of smaller-scale features.

Principal Component Analysis (PCA) provides a well-established statistical solution by projecting high-dimensional genomic data into an orthogonal, low-dimensional space that captures the greatest possible variance ((5)). This approach reduces collinearity, produces uncorrelated summary variables, and generates interpretable axes of variation. In population genomics, PCA has been extensively applied to uncover population structure, detect admixture, and identify batch effects ((8); (7)). PCA can also be used to test whether patterns of genetic variation correspond to ecological groups, for example by comparing populations sampled from different environments. This makes PCA a useful tool for drawing biologically meaningful conclusions.

1.2 Statistical Characterization of Genomic Predictors

Let $X \in \mathbb{R}^{n \times p}$ denote the genomic feature matrix, where n is the number of individuals and p is the number of features. For SNP genotypes, the entries $X_{ij} \in \{0, 1, 2\}$ represent the number of copies of the minor allele carried by individual i at locus j. For RNA-seq or k-mer features, raw counts are first normalized using variance-stabilizing transformations or log-counts per million (log-CPM) to account for sequencing depth and variance heterogeneity (1).

All features are mean-centered and typically standardized to have unit variance before PCA. For genotype data, Patterson's scaling is applied:

$$\tilde{X}_{ij} = \frac{X_{ij} - 2\hat{p}_j}{\sqrt{2\hat{p}_j(1-\hat{p}_j)}},$$

where $\hat{p}_j$ is the estimated allele frequency at locus j. This adjustment ensures that the variance of each SNP reflects expectations under neutral genetic drift (8).

PCA is then performed using singular value decomposition (SVD):

$$X = USV^{\top},$$

where the columns of V are the eigenvectors (also called loadings), US contains the sample scores (the coordinates of individuals in the principal component space), and the eigenvalues

$$\lambda_k = \frac{S_{kk}^2}{n-1}$$

quantify the variance explained by each component.

In this framework, the eigenvectors define weighted combinations of genomic features (e.g., SNPs or genes) that represent independent axes of variation in the data. The sample scores project individuals onto these axes, making it possible to visualize how genetic variation is distributed across populations or conditions. Scree plots and cumulative variance curves are then used to assess how much of the overall variation is captured by the leading components.

1.3 Development and Validation of PCA Models

In the analysis, validation of PCA was performed primarily through explained variance diagnostics. Scree plots and cumulative variance curves were used to determine how many principal components captured meaningful structure in the data. The first few components explained most of the variation, and these were selected for further analysis and visualization.

Individuals were then projected into the reduced PC space, and the resulting coordinates were used for clustering with k-means. This allowed the identification of sample groupings and patterns of genetic similarity. Validation in this context focused on the interpretability and stability of these clusters across different numbers of principal components.

1.4 Operational Insights

The results of PCA provided a reduced-dimensional representation of the genomic data, which facilitated visualization of individual variation. By plotting the first two or three principal components, patterns of similarity among individuals could be observed, such as separation into distinct clusters or gradients of variation.

K-means clustering was then applied to the PC scores to group individuals based on their genetic similarity. The alignment of clusters with the PCA plots allowed for a clearer interpretation of group structure, confirming whether the reduced dimensions captured meaningful biological signal. These clusters, together with the PCA projections, served as the basis for identifying patterns in the data that would have been obscured in the original high-dimensional space.

1.5 Summary

This PCA-based framework addresses the statistical challenges of high-dimensional genomic data by reducing dimensionality and transforming correlated predictors into orthogonal components. The resulting principal components capture the major axes of variation in the dataset and make it possible to visualize patterns of similarity and difference among individuals. By applying k-means clustering to the PC scores, groups of genetically similar individuals can be identified, providing insights that would be difficult to detect in the original high-dimensional space. Together, PCA and clustering offer a practical and interpretable approach for exploring structure in genomic datasets while maintaining reproducibility through standardized preprocessing and projection procedures.

References

[1] Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10), R106.

[2] Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: a

review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1), 27–46.

[3] Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences USA*, 95(25), 14863–14868.

[4] Hall, P., Marron, J. S., & Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B*, 67(3), 427–444.

[5] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065), 20150202.

[6] Marchenko, V. A., & Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4), 457–483.

[7] Novembre, J., & Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, 40(5), 646–649.

[8] Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, 2(12), e190.

[9] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8), 904–909.

[10] Slatkin, M. (2008). Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6), 477–485.

[11] Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose. *The American Statistician*, 70(2), 129–133.