

Principal Component Analyses in Anthropological Genetics

PDF (Size: 180KB) PP. 9-14 DOI: 10.4236/aa.2011.12002

Author(s)

Xingdong Chen, Chao Chen, Li Jin

ABSTRACT

Principal component analyses (PCA) is a statistical method for exploring and making sense of datasets with a large number of measurements (which can be thought of as dimensions) by reducing the dimensions to the few principal components (PCs) that explain the main patterns. Thus, the first PC is the mathematical combination of measurements that accounts for the largest amount of variability in the data. Here, we gave an interpretation about the principle of PCA and its original mathematical algorithm, singular variable decomposition (SVD). PCA can be used in study of gene expression; also PCA has a population genetics interpretation and can be used to identify differences in ancestry among populations and samples, through there are some limitations due to the dynamics of microevolution and historical processes, with advent of molecular techniques, PCA on Y chromosome, mtDNA, and nuclear DNA gave us more accurate interpretations than on classical markers. Furthermore, we list some new extensions and limits of PCA.

KEYWORDS

Principal Component Analysis, Singular Value Decomposition, Human Genetics

Cite this paper

Chen, X., Chen, C. & Jin, L. (2011). Principal Component Analyses in Anthropological Genetics. *Advances in Anthropology*, 1, 9-14. doi: 10.4236/aa.2011.12002.

References

- [1] Alter, O., Brown, P. O., & Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97, 10101. doi:10.1073/pnas.97.18.10101
- [2] Alter, O., Brown, P. O., & Botstein, D. (2003). Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proceedings of the National Academy of Sciences*, 100, 3351. doi:10.1073/pnas.0530258100
- [3] Alter, O., & Golub, G. H. (2006). Singular value decomposition of genome-scale mRNA lengths distribution reveals asymmetry in RNA gel electrophoresis band broadening. *Proceedings of the National Academy of Sciences*, 103, 11828. doi:10.1073/pnas.0604756103
- [4] Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM Review*, 37, 573-595. doi:10.1137/1037127
- [5] Biswas, S., Storey, J., & Akey, J. (2008). Mapping gene expression quantitative trait loci by singular value decomposition and independent component analysis. *BMC Bioinformatics*, 9, 244. doi:10.1186/1471-2105-9-244
- [6] Cavalli-Sforza, L. L., Menozzi, P., & Piazza, A. (1994). *The history and geography of human genes*. Princeton, NJ: Princeton University Press.
- [7] Chen, L., Hodgson, K. O., & Doniach, S. (1996). A lysozyme folding intermediate revealed by solution X-ray scattering. *Journal of Molecular Biology*, 261, 658-671. doi:10.1006/jmbi.1996.0491
- [8] Clayton, D. G., Walker, N. M., Smyth, D. J., Pask, R., Cooper, J. D., Maier, L. M., Smink, L. J., Lam, A. C.,

- [Open Special Issues](#)
- [Published Special Issues](#)
- [Special Issues Guideline](#)

[AA Subscription](#)[Most popular papers in AA](#)[About AA News](#)[Frequently Asked Questions](#)[Recommend to Peers](#)[Recommend to Library](#)[Contact Us](#)

Downloads: 29,494

Visits: 136,920

[Sponsors, Associates, and Links >>](#)

Ovington, N. R., & Stevens, H. E. (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genetics*, 37, 1243-1246. doi:10.1038/ng1653

- [9] Fesel, C., & Coutinho, A. (1998). Dynamics of serum IgM autoreactive repertoires following immunization: strain specificity, inheritance and association with autoimmune disease susceptibility. *European Journal of Immunology*, 28, 3616-3629. doi:10.1002/(SICI)1521-4141(199811)28:11<3616::AID-IMMU3616>3.0.CO;2-B
- [10] Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Hausler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16, 906. doi:10.1093/bioinformatics/16.10.906
- [11] Handley, L. J. L., Manica, A., Goudet, J., & Balloux, F. (2007). Going the distance: Human population genetics in a clinal world. *TRENDS in Genetics*, 23, 432-439. doi:10.1093/bioinformatics/16.10.906
- [12] Holter, N. S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J. R., & Fedoroff, N. V. (2000). Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *Proceedings of the National Academy of Sciences*, 97, 8409. doi:10.1073/pnas.150242097
- [13] Hyv Rinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, 13, 411-430. doi:10.1016/S0893-6080(00)00026-5
- [14] Khan, J., Wei, J. S., Ringnér, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., & Peterson, C. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7, 673-679. doi:10.1038/89044
- [15] Lee, A. B., Luca, D., Klei, L., Devlin, B., & Roeder, K. (2010). Discovering genetic ancestry using spectral graph theory. *Genetic Epidemiology*, 34, 51-59.
- [16] Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., & Cavalli-Sforza, L. L. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319, 1100. doi:10.1126/science.1153717
- [17] Luca, D., Ringquist, S., Klei, L., Lee, A. B., Gieger, C., & Wichmann, H. (2008). On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. *The American Journal of Human Genetics*, 82, 453-463. doi:10.1016/j.ajhg.2007.11.003
- [18] Mellars, P. (2006). Going east: New genetic and archaeological perspectives on the modern human colonization of Eurasia. *Science*, 313, 796. doi:10.1016/j.ajhg.2007.11.003
- [19] Menozzi, P., Piazza, A., & Cavalli-Sforza, L. (1978). Synthetic maps of human gene frequencies in Europeans. *Science*, 201, 786. doi:10.1126/science.356262
- [20] Novembre, J., & Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, 40, 646-649. doi:10.1038/ng.139
- [21] Omberg, L., Golub, G. H., & Alter, O. (2007). A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies. *Proceedings of the National Academy of Sciences*, 104, 18371. doi:10.1073/pnas.0709146104
- [22] Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, 2, e190. doi:10.1371/journal.pgen.0020190
- [23] Pearson, K. LIII. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, 2, 559-572. doi:10.1080/14786440109462720
- [24] Pinhasi, R., Fort, J., & Ammerman, A. J. (2005). Tracing the origin and spread of agriculture in Europe. *PLoS Biology*, 3, e410. doi:10.1371/journal.pbio.0030410
- [25] 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*, 38, 904-909.
- [26] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81, 559-575. doi:10.1086/519795

- [27] Quackenbush, J. (2001). Computational analysis of microarray data. *Nature Reviews Genetics*, 2, 418-427. doi:10.1038/35076576
- [28] Raychaudhuri, S., Stuart, J. M., & Altman, R. B. (2000). Principal components analysis to summarize microarray experiments. Application to Sporulation Time Series, 455.
- [29] Reich, D., Price, A. L., & Patterson, N. (2008). Principal component analysis of genetic data. *Nature Genetics*, 40, 491-491. doi:10.1038/ng0508-491
- [30] Richards, J. A., & Jia, X. (2006). *Remote sensing digital image analysis: An introduction*. Berlin: Springer Verlag.
- [31] Romo, T. D., Clarage, J. B., Sorensen, D. C., & Phillips Jr, G. N. (1995). Automatic identification of discrete substates in proteins: Singular value decomposition analysis of time—Averaged crystal-