Project ideas

PCA and population migrations

DNA sequence data are used to learn about human population movements over time. Researchers identify many SNPs (single nucleotide polymorphism) on the sequence, each can be A, C, G, or T (4 choices). The SNP sequence is normalized into a vector. The vectors of many individuals are organized into a matrix and again normalized. The first k principal components of this matrix form a point in $\mathbb{R}^k$. A number of individuals make a point cloud. These points are seen to form clusters corresponding to different populations. These are often visualized by making scatterplots of the first two principal components. Learn about this. Download the SNP data and make cornerplots to try to visualize clusters in more than 2D.

Explore the Randomly Pivoted Cholesky (RPCholesky) of Tropp and collaborators

.

Evaluation of evidence integrals

The evidence integral is is the normalization integral of a Bayesiann posterior. They are used for model selection and for other purposes. Explore nested sampling, thermodynamic integration approaches to the evidence integral. Compare the value of the evidence integral to approximations using Laplace approximation, called the Bayesian Information Criterion.