Andreas Noack

Faster than FastPCA: High performance principal components analysis of genomics data

32 Vassar St
Rm 32-206
MIT CSAIL
Cambridge
MA 02139-4307
noack@mit.edu
Jiahao Chen
Jake Bolewski
Alan Edelman
Nikolaos Patsopoulos

Data for genome-wide association studies (GWAS) involve extremely large matrices with small nonnegative integer entries representing deviations from a reference genome. Principal components analysis (PCA) is traditionally used to identify clusters reflecting subpopulation structure in genomics populations.

We present implementations of high performance principal components analysis using Golub-Kahan-Lanczos (GKL) iterative bidiagonalization routines written in pure Julia. We demonstrate out of core computation on data stored on external files and databases, showing how Julia's user-extensible type system and generic functions with multimethods (multiple dispatch) allow for flexible and convenient definitions of new types reflecting new data structures, such as matrices stored in HDF5 files or in an array database. Operator overloading also allows for a natural expression of elementary operations such as matrix-vector products that exploit detailed knowledge of underlying storage formats and layouts for improved performance.

We compare our implementation of principal components analysis with existing tools designed for GWAS data, such as EIGENSOFT, FlashPCA and FastPCA. We show that in many cases, iterative bidiagonalization outperforms other methods implemented in these existing tools, such as randomized subspace iteration. We explore how convergence of GKL iterates and auxiliary quantities used in practical roundoff control strategies, such as the ω-recurrence coefficients in partial reorthogonalization reflect underlying structure latent in genomics data matrices.