Fast, Exact Bootstrap Principal Component Analysis for p>1 million

Aaron Fisher,[1] Brian Caffo,[1] Brian Schwartz,[2] & Vadim Zipunnikov[1]

1. Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
2. Department of Environmental Health Sciences, Johns Hopkins Bloomberg School of Public Health

Summary

Problem: Bootstrapping can be used to estimate uncertainty of Principal Component Analysis (PCA) due to sampling variability, but the procedure is computationally demanding for high dimensional (HD) data.

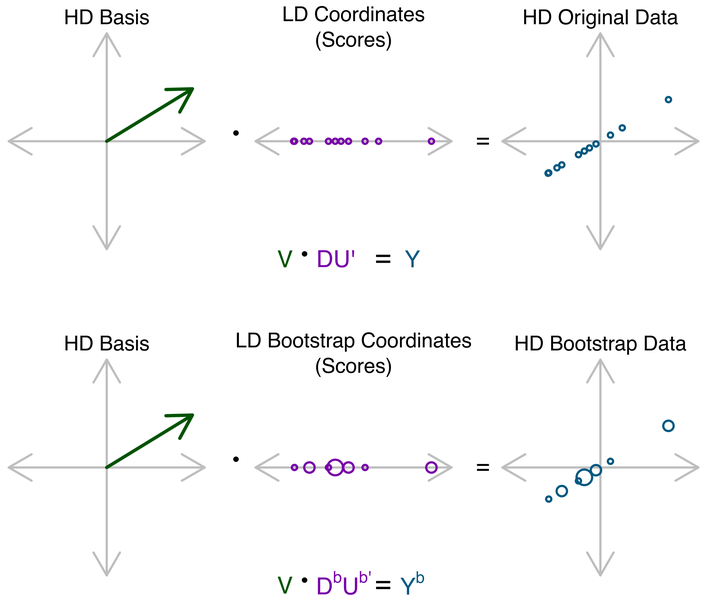
Solution: Speed improvements come from the fact that all bootstrap principal components (PCs) lie within the same, low-dimensional (LD) subspace.

Results: We calculate the 3 leading PCs of a brain MRI dataset (≈ 3 million voxels, 352 subjects), in 1000 bootstrap samples. *Computation time was reduced from 4 days to 120 minutes,* using a standard laptop.

Intuition

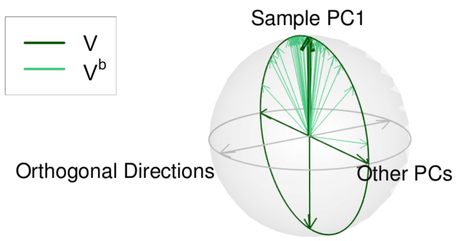
Basic Notation

- n = number of subjects
- p = number of measurements per subject (p>>n)
- Y = a (p by n) centered data matrix
- VDU' = svd(Y)
- b=1,2,…B is the bootstrap index



The n-dimensional span of V (or Y) contains:

- All sample points
- *All bootstrap sample points*
- *All bootstrap PCs*



Efficient PCA Calculations (via SVD)

Bootstrap PCs can be represented as rotations of the original sample PCs (V^b=VA^b)

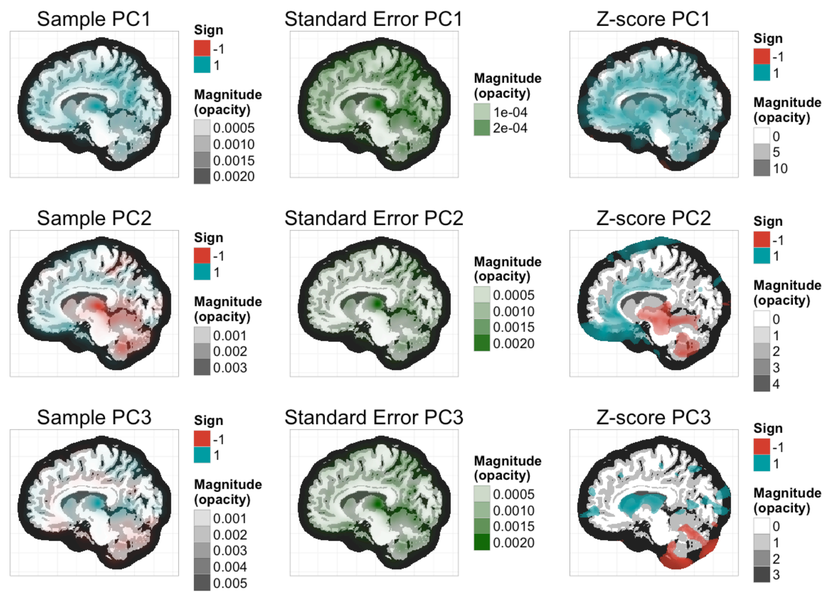
$$\begin{aligned}\mathbf{Y}^b &= \mathbf{Y}\mathbf{P}^b && \text{where } \mathbf{P}^b \text{ is a resampling matrix}\\ &= \mathbf{V}\mathbf{D}\mathbf{U}'\mathbf{P}^b && (\mathbf{D}\mathbf{U}'\mathbf{P}^b) \text{ are the resampled scores}\\ &= \mathbf{V}(\mathbf{A}^b\mathbf{D}^b\mathbf{U}^{b'}) && \text{where } \mathbf{A}^b\mathbf{D}^b\mathbf{U}^{b'} := svd(\mathbf{D}\mathbf{U}'\mathbf{P}^b)\\ &= (\mathbf{V}\mathbf{A}^b)\mathbf{D}^b(\mathbf{U}^b)'\\ &= svd(\mathbf{Y}^b)\end{aligned}$$

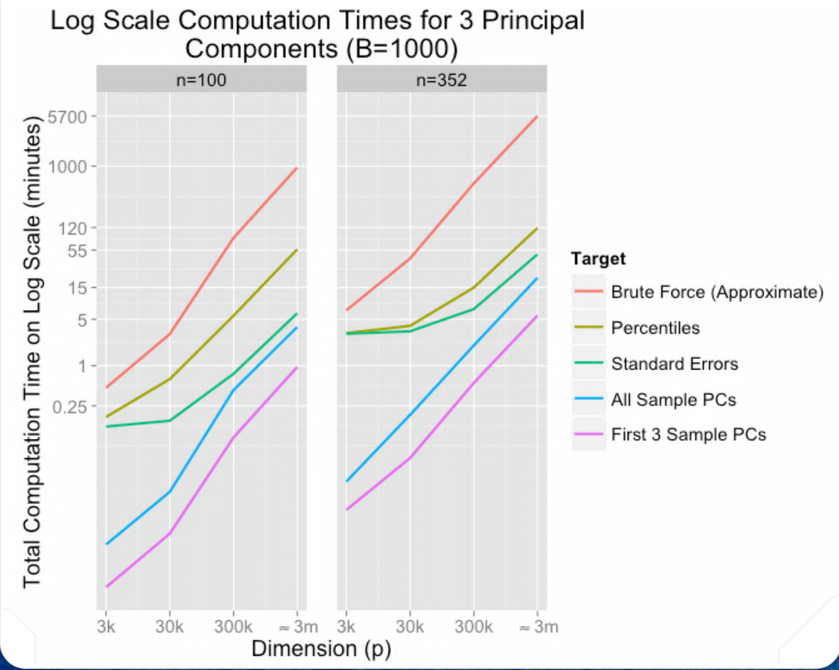
Many statistics and CIs only require the low dimensional matrices A^b:

- $E_b(V^b)$, which can be found via $E_b(VA^b)=VE_b(A^b)$
- Pointwise variances ($Var_b(V^b)$) from analogous simplifications
- Moment-based pointwise CIs
- Confidence cones created from angles between V and V^b

Results & Computation Times

- Dataset of Voxel Based Morphometry images (RAVENS)
- p=2,979,666; n=352





References & Acknowledgements

See paper for details on:

- Adjusting for axis reflections of bootstrap PCs
- Constructing confidence regions
- Block matrix algebra for p>10,000
- Adjusting for when the svd does not converge
- Electroencephalogram data example
- Simulations

arXiv: http://arxiv.org/abs/1405.0922
bootSVD package: https://github.com/aaronjfisher/bootSVD
ggBrain package: https://github.com/aaronjfisher/ggBrain

Aaron Fisher
aaronjfisher.com
fisher@jhu.edu
@prfFarnsworth (twitter)

Poster design inspired by Michael Barton & Alyssa Frazee
Texture from texturecrate (deep-lagoon-grunge)