Empirical Bayes PCA in high dimensions

Xinyi Zhong

Department of Statistics and Data Science

Yale University

Berkeley CDAR seminar

Sept. 27th, 2022

Joint work with

Adviser: Dr. Zhou Fan
Department of Statistics and Data Science
Yale University

Chang Su
Department of Biostatistics
Yale University

Motivation

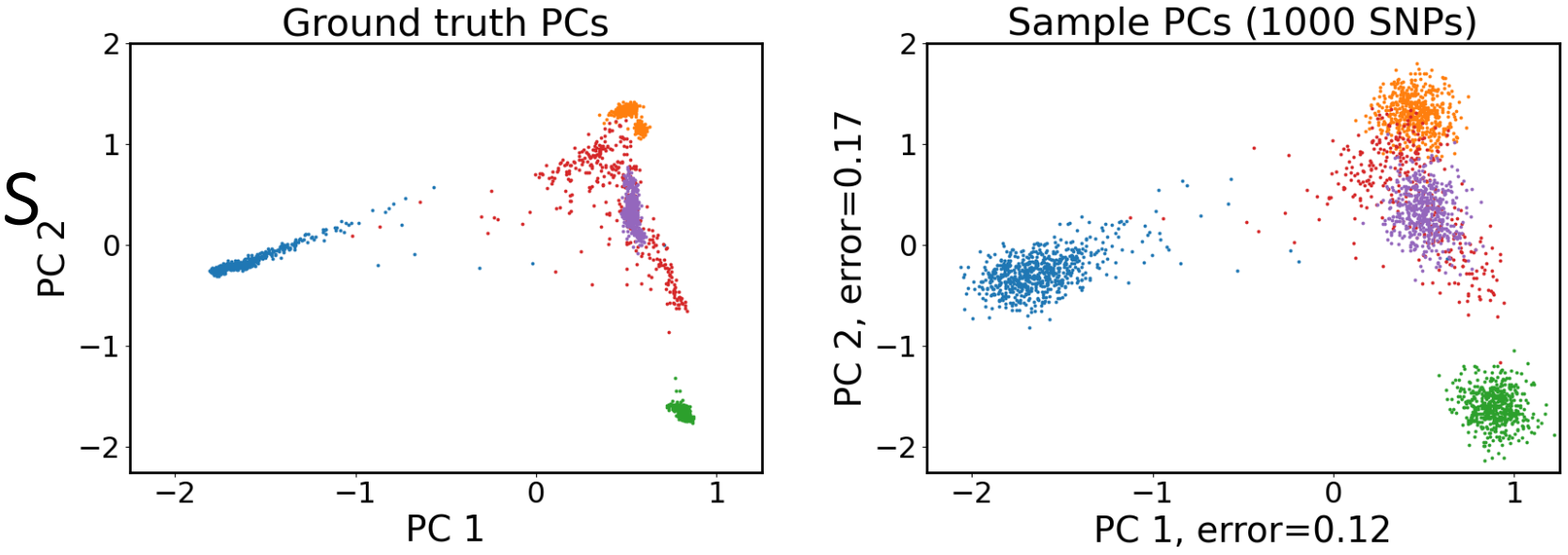
- Model: $\mathbf{X} = \mathbf{U}\,\mathbf{V}^{\top} + \mathbf{W}$
- Goal: estimate $\mathbf{U}\mathbf{V}^{\top}$

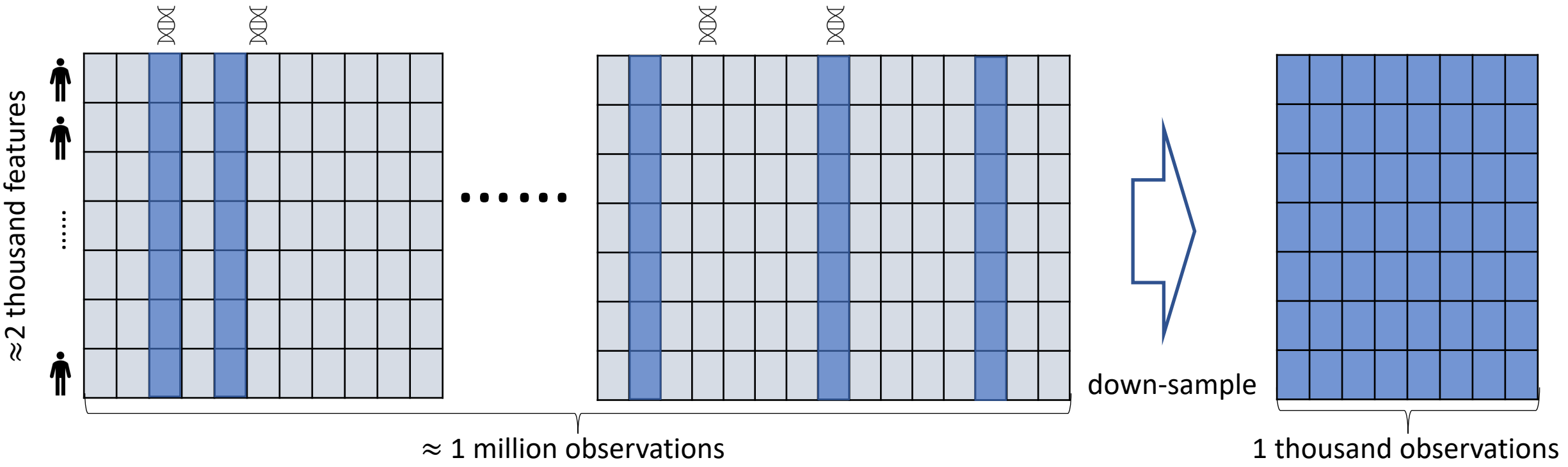
- Naïve PCA
  - High dimensional noise
  - Doesn’t utilize PC structure
- Bayes estimate
  - Intensive computation
  - Unavailable prior information

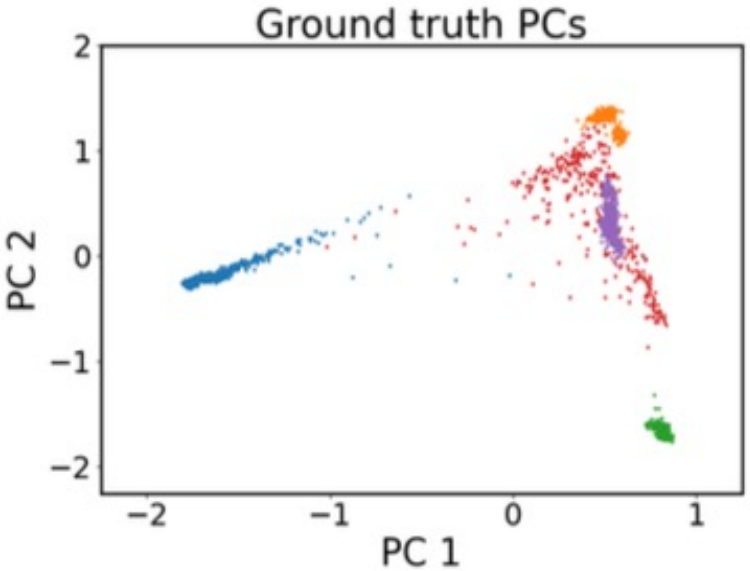
Empirical Bayes PCA (EBPCA) uses less computational cost to achieve more accurate estimate.

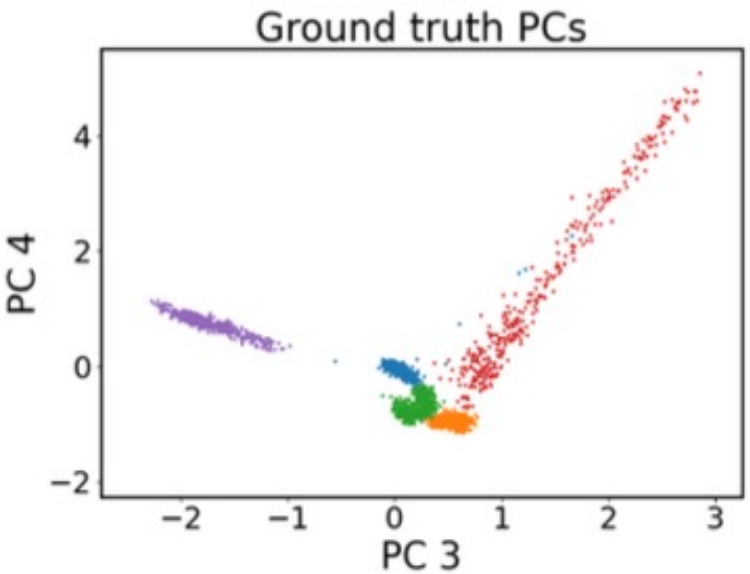
- polynomial time
- Bayes optimal

1000Genomes









African Caucasian East-Asian Hispanic South Asian

Outline

- Method and results
- Relation to other methods
- Extensions of the noise model

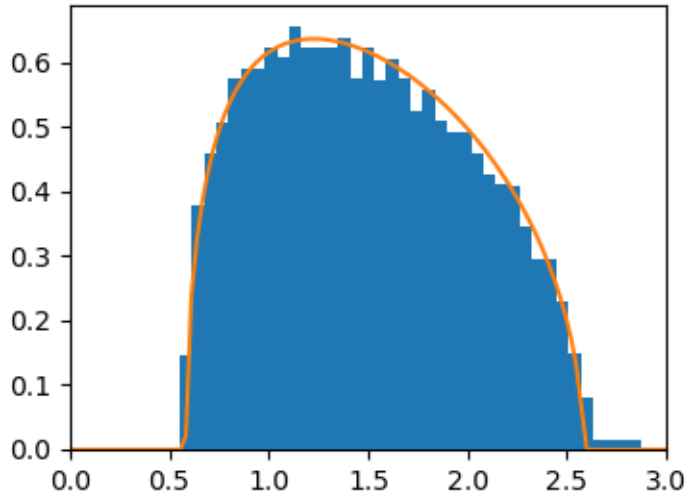
The EB-PCA method

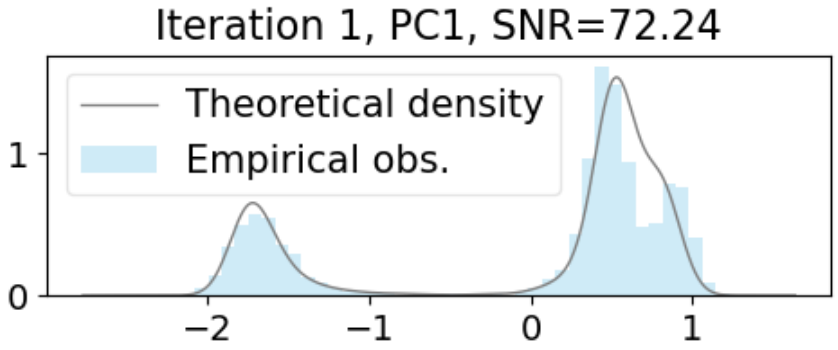
There are three statistical ideas underlying EB-PCA:

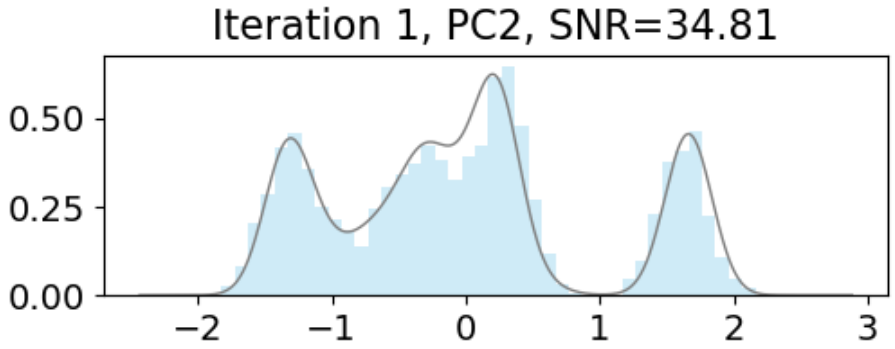
- Random matrix theory for the sample PCs
- Empirical Bayes estimation for normal means
- Bayesian inference using Approximate Message Passing (AMP)

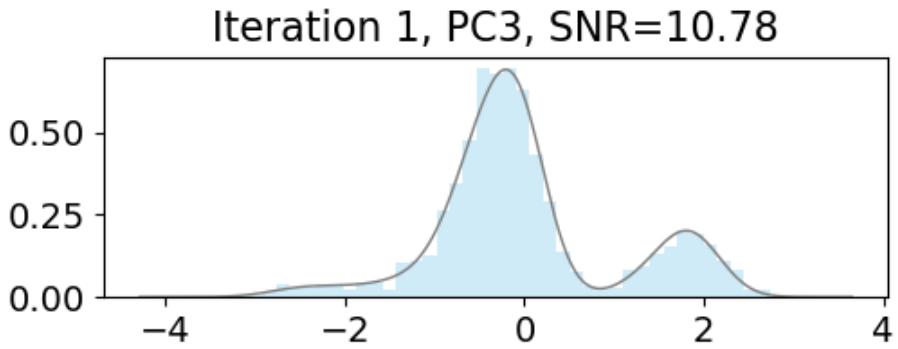
Rank-one signal-plus-noise model

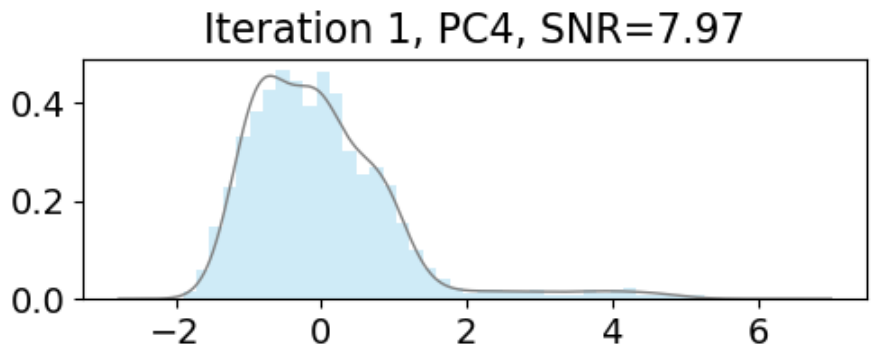
$$\mathbf{X} = \frac{s}{n} * \mathbf{u}\boldsymbol{v}^{\top} + \mathbf{W} \approx \frac{\lambda}{n} * \boldsymbol{f}\boldsymbol{g}^{\top}$$





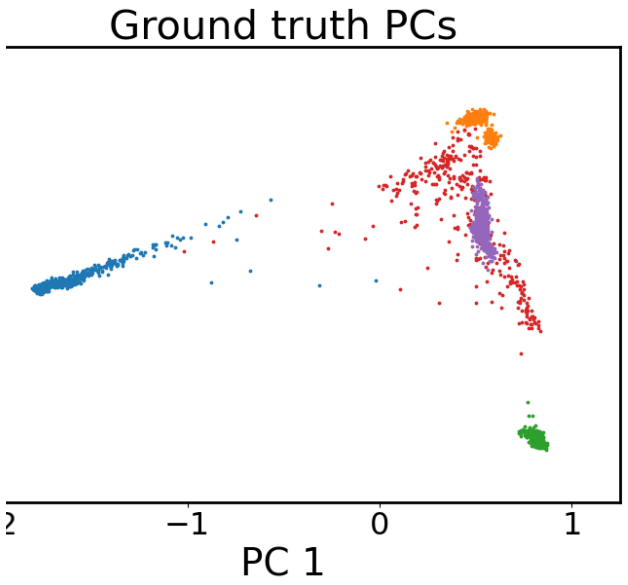


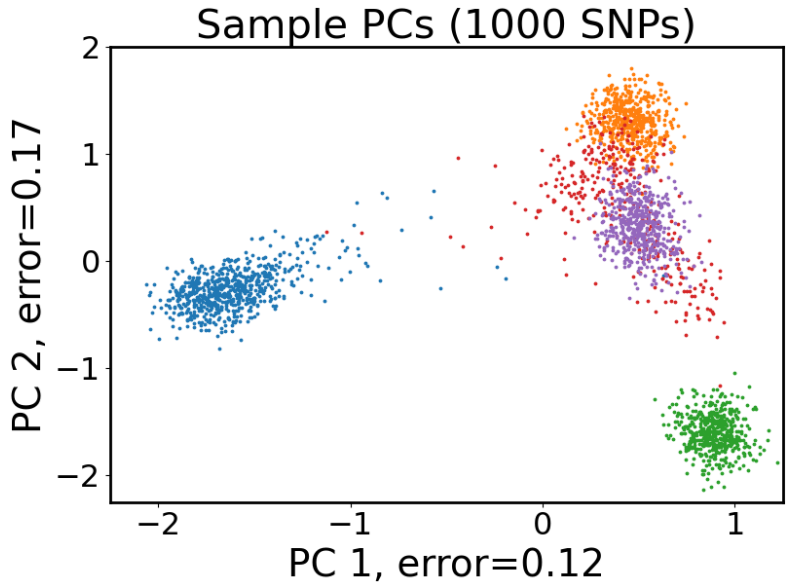




asymptotics for sample PCs.

$$\mathbf{g} \sim N(\mu_* * \boldsymbol{v}, \sigma_*^2)$$



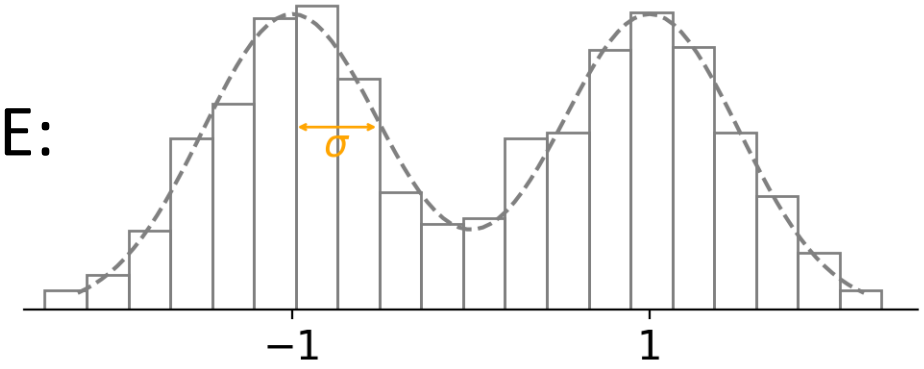


Empirical Bayes in compound decision

$$\mathrm{g_j} \sim N\left(\mu_* * \boldsymbol{v_j}, \sigma_*^2\right) \text{ for all } j \in [n] \neq \mathbf{g} \sim N(\mu_* * \boldsymbol{v}, \sigma_*^2)$$

Nonparametric empirical

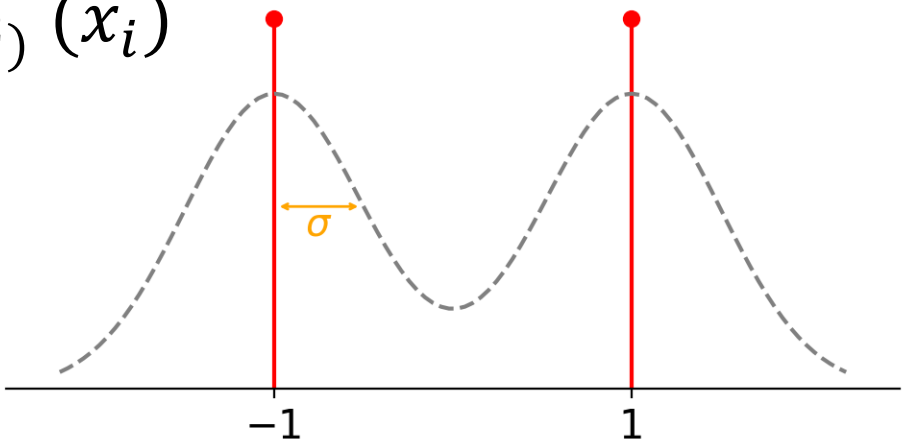
We estimate π through nonparametric MLE:



Model: $\boldsymbol{x} \sim N(\boldsymbol{\theta}, \sigma^2),\ \boldsymbol{\theta} \sim \pi$

Inference: $\hat{\pi} = \text{argmax}_{\pi} \sum_{i=1}^{n} \log f_{\pi * N(0,\sigma^2)}(x_i)$

Denoise: $\widehat{\boldsymbol{\theta}} = \mathrm{E}[\boldsymbol{\theta}|\boldsymbol{x}] \coloneqq \theta(\boldsymbol{\theta}|\mu, \sigma^2, \pi)$



Empirical Bayes in compound decision

Given $\mathbf{g} \sim N(\mu_* * \boldsymbol{v}, \sigma_*^2)$

1. Infer $v_j \sim \pi$
2. Denoise g by taking $\widehat{\boldsymbol{v}} = \mathrm{E}[\boldsymbol{v}|\boldsymbol{g}] \coloneqq \theta(\boldsymbol{g}|\mu_*, \sigma_*^2, \pi)$

Iterative refinement via AMP

As $\widehat{\boldsymbol{v}} \prec \boldsymbol{g}$ for $\boldsymbol{v}$, we expect $\boldsymbol{Y}\widehat{\boldsymbol{v}} \prec \boldsymbol{Y}\boldsymbol{g} \propto \boldsymbol{f}$ for $\boldsymbol{u}$.

$$\boldsymbol{v}_t = \theta(\boldsymbol{g}_t | \mu_*, \sigma_*^2, \pi)$$

$$\boldsymbol{f}_t = \boldsymbol{Y}\boldsymbol{v}_t$$

$$\boldsymbol{u}_t = \theta(\boldsymbol{f}_t | \bar{\mu}_*, \bar{\sigma}_*^2, \bar{\pi})$$

$$\boldsymbol{g}_{t+1} = \boldsymbol{Y}^\top \boldsymbol{u}_t$$

Iterative refinement via AMP

As $\widehat{\boldsymbol{v}} \prec \boldsymbol{g}$ for $\boldsymbol{v}$, we expect $\boldsymbol{Y}\widehat{\boldsymbol{v}} \prec \boldsymbol{Y}\boldsymbol{g} \propto \boldsymbol{f}$ for $\boldsymbol{u}$.

$$\boldsymbol{v}_t = \theta(\boldsymbol{g}_t | \mu_*, \sigma_*^2, \pi)$$

$$\boldsymbol{f}_t = \boldsymbol{Y}\boldsymbol{v}_t - b_t \boldsymbol{u}_{t-1}$$

$$\boldsymbol{u}_t = \theta(\boldsymbol{f}_t | \bar{\mu}_*, \bar{\sigma}_*^2, \bar{\pi})$$

$$\boldsymbol{g}_{t+1} = \boldsymbol{Y}^\top \boldsymbol{u}_t - \bar{b}_t \boldsymbol{v}_t$$

Theoretical guarantee of EB-PCA (informal)

Suppose $u_1, \dots, u_m \sim \bar{\pi}_*$ and $v_1, \dots, v_m \sim \pi_*$ for some priors $\pi_*, \bar{\pi}_*$, and the noise $\boldsymbol{W}$ has i.i.d. entries.

As $n, d \to \infty$, EB-PCA consistently estimates $\pi_*, \bar{\pi}_*$. Furthermore, for every iterate $\boldsymbol{f}_t$ and $\boldsymbol{g}_t$ of EB-PCA,

$$\mathbf{f}_{\mathrm{t}} \sim N(\bar{\mu}_t * \boldsymbol{u}, \bar{\sigma}_t^2), \quad \mathbf{g}_{\mathrm{t}} \sim N(\mu_t * \boldsymbol{v}, \sigma_t^2),$$

where $\bar{\mu}_t, \bar{\sigma}_t^2, \mu_t$ and σ_t^2 are deterministic sequence which can be prescribed by AMP state evolution.

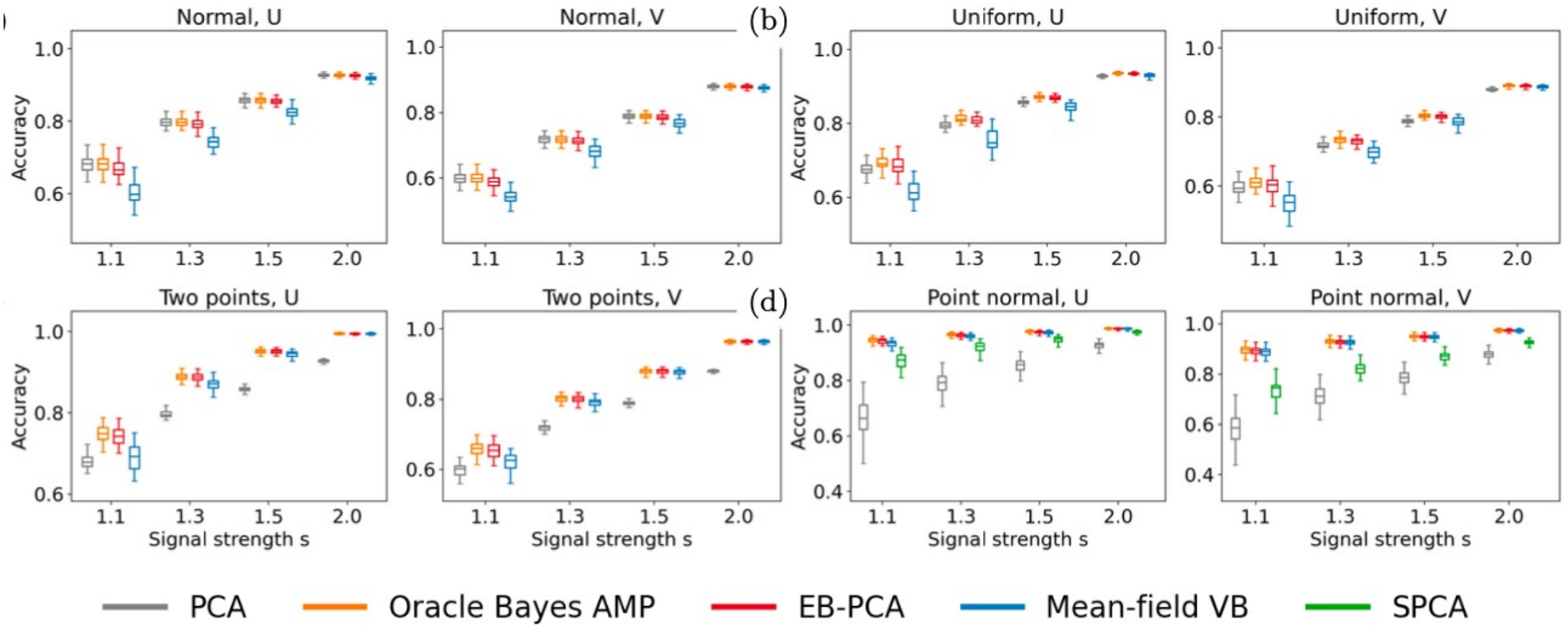
EB-PCA method

1. Get samples PCs and estimate their alignments to the true PCs.
2. Use (NP)MLE to estimate the prior distribution of the true PCs.
3. Follow the AMP algorithm to get improved estimates iteratively.

Comparison

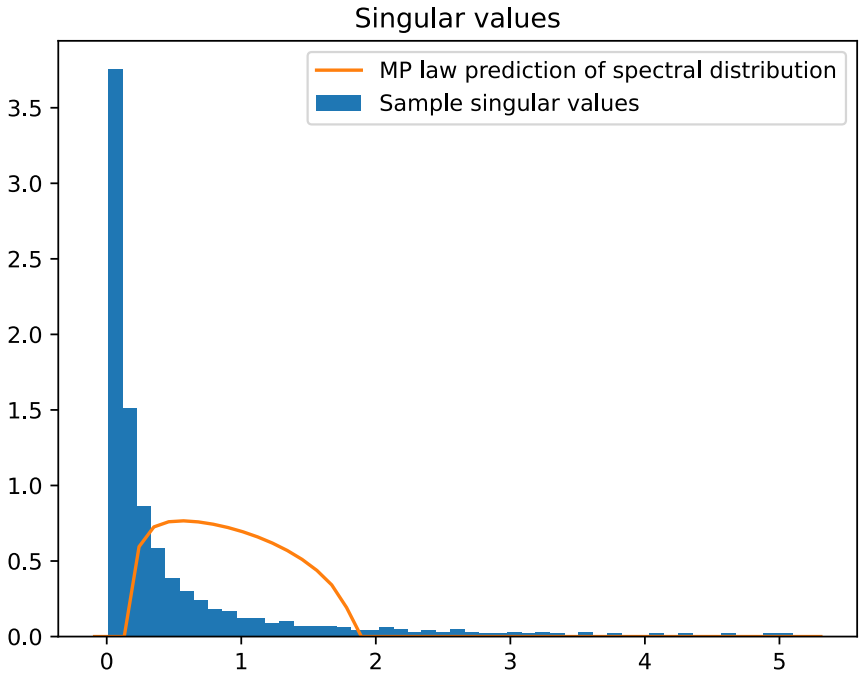
- Standard PCA
- Sparse PCA
- Mean-field Variational Bayes [Wang, Stephens 2021]
- James Stein denoisers
- …

Simulation Studies

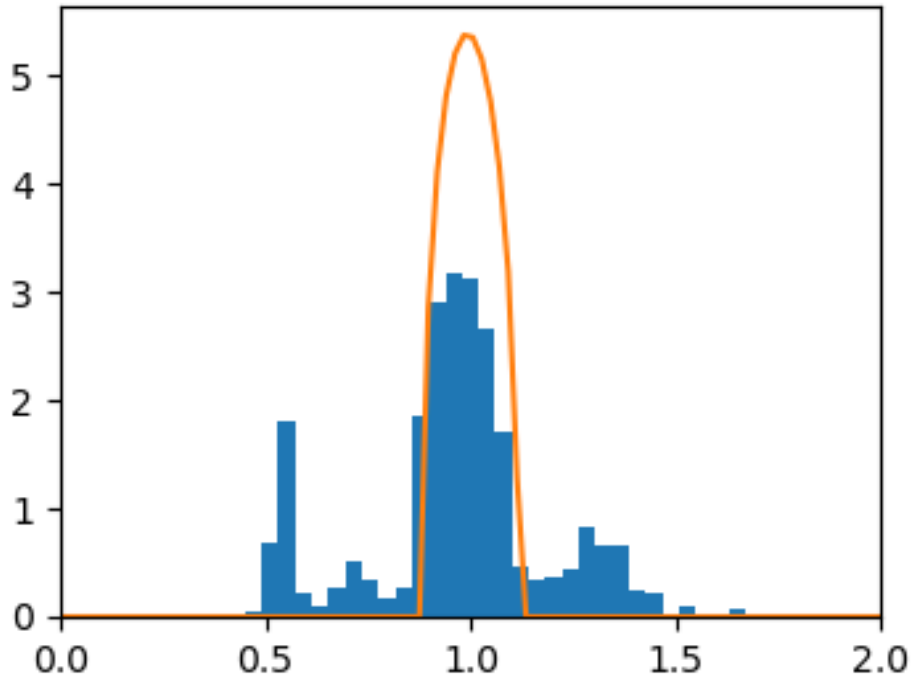


Rotationally invariant ensemble

Stock returns



HapMap3



Universality results

- Heavy-tailed entries
- Heterogeneous entries