Trace and Schatten norm estimation

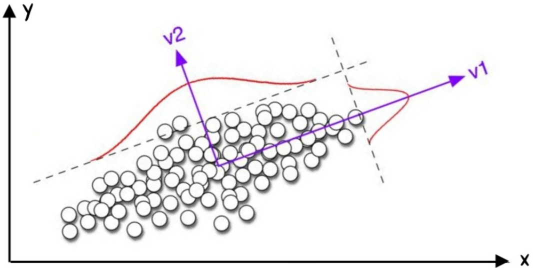
using randomized sampling

Zijian Wang

Graduate Seminar on Efficient Simulation
Randomized Numerical Linear Algebra
May 2, 2022

Motivation

- Numerical linear algebra plays a fundamental role in applied mathematics



https://www.analyticsvidhya.com/blog/2016/07/making-predictions-test-data-principal-component-analysis/

- Classical methods are often infeasible for large-scale problems
- Randomized algorithms can provide efficient approximations

Outline

1. Preliminaries
2. Trace estimation
   - Estimator based on randomized sampling
   - A priori error estimates
   - A posteriori error estimates
   - Application to Frobenius and Schatten 4-norm
3. Schatten norm estimation
   - Approach from classical statistics
   - Computationally efficient method
4. Summary and outlook

Review of matrix decomposition

- Eigendecomposition:

$$\boldsymbol{A} \in \mathbb{R}^{n \times n} \text{ PSD} \quad \Longrightarrow \quad \boldsymbol{A} = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^{\mathsf{T}}$$

  with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$.

- Singular value decomposition:

$$\boldsymbol{B} \in \mathbb{R}^{m \times n} \quad \Longrightarrow \quad \boldsymbol{B} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathsf{T}}$$

  with $r = \text{rank}(\boldsymbol{B})$ and singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$.

  Recall that σ_i^2 represent the non-zero eigenvalues of $\boldsymbol{B}^{\mathsf{T}}\boldsymbol{B}$.

Schatten norms

Definition

For a matrix $\boldsymbol{B} \in \mathbb{R}^{m \times n}$ with singular values $\sigma_1(\boldsymbol{B}) \geq \cdots \geq \sigma_r(\boldsymbol{B}) > 0$, we define its *Schatten p-norm* as:

$$\|\boldsymbol{B}\|_p = \begin{cases} \left(\sum_{i=1}^r \sigma_i(\boldsymbol{B})^p \right)^{\frac{1}{p}}, & p \in [1, \infty) \\ \sigma_1(\boldsymbol{B}), & p = \infty \end{cases}$$

Important properties:

- l^p-norm of the vector of singular values $\implies$ Hölder, monotonicity, etc.
- Orthogonal invariance: $\|\boldsymbol{UBV}\|_p = \|\boldsymbol{B}\|_p$ for orthogonal matrices $\boldsymbol{U}, \boldsymbol{V}$

Schatten norms

Definition

For a matrix $\boldsymbol{B} \in \mathbb{R}^{m\times n}$ with singular values $\sigma_1(\boldsymbol{B}) \geq \cdots \geq \sigma_r(\boldsymbol{B}) > 0$, we define its *Schatten p-norm* as:

$$\|\boldsymbol{B}\|_p = \begin{cases} \left(\sum_{i=1}^{r} \sigma_i(\boldsymbol{B})^p\right)^{\frac{1}{p}}, & p \in [1, \infty) \\ \sigma_1(\boldsymbol{B}), & p = \infty \end{cases}$$

Special cases:

- Frobenius norm: $\|\boldsymbol{B}\|_2 = \|\boldsymbol{B}\|_\mathsf{F} = \left(\sum_{i,j} (\boldsymbol{B})_{ij}^2\right)^{\frac{1}{2}}$
- Spectral norm: $\|\boldsymbol{B}\|_\infty = \|\boldsymbol{B}\|_\mathsf{s} = \sigma_1(\boldsymbol{B})$
- PSD matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$: $\|\boldsymbol{A}\|_1 = \mathrm{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} (\boldsymbol{A})_{ii}$

Outline

1. Preliminaries
2. Trace estimation
   - Estimator based on randomized sampling
   - A priori error estimates
   - A posteriori error estimates
   - Application to Frobenius and Schatten 4-norm
3. Schatten norm estimation
   - Approach from classical statistics
   - Computationally efficient method
4. Summary and outlook

Problem setting and first approach

- Given: Non-zero PSD matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ via $\boldsymbol{u} \mapsto \boldsymbol{A}\boldsymbol{u}$
- Goal: Estimate tr$(\boldsymbol{A})$ without computing $\boldsymbol{\delta}_i^\mathsf{T}(\boldsymbol{A}\boldsymbol{\delta}_i)$ for $i = 1, \ldots, n$
- Idea: Take a random test vector $\boldsymbol{\omega} \in \mathbb{R}^n$ with $\mathbb{E}[\boldsymbol{\omega}\boldsymbol{\omega}^\mathsf{T}] = \boldsymbol{I}_n$.
  Then, the random variable $X = \boldsymbol{\omega}^\mathsf{T}(\boldsymbol{A}\boldsymbol{\omega})$ satisfies:

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}[\operatorname{tr}(\boldsymbol{\omega}^\mathsf{T}\boldsymbol{A}\boldsymbol{\omega})] = \mathbb{E}[\operatorname{tr}(\boldsymbol{A}\boldsymbol{\omega}\boldsymbol{\omega}^\mathsf{T})] \\
&= \operatorname{tr}(\mathbb{E}[\boldsymbol{A}\boldsymbol{\omega}\boldsymbol{\omega}^\mathsf{T}]) = \operatorname{tr}(\boldsymbol{A}\,\mathbb{E}[\boldsymbol{\omega}\boldsymbol{\omega}^\mathsf{T}]) = \operatorname{tr}(\boldsymbol{A})
\end{aligned}$$

  $\implies X$ is an unbiased estimator of tr$(\boldsymbol{A})$
- Problem: Var$[X]$ possibly large
  $\implies$ Reduce variance by averaging independent copies of X

A consistent estimator of the trace

- Fix sample size $k \in \mathbb{N}$, $k \ll n$.
- For $i = 1, \ldots, k$, sample a test vector $\boldsymbol{\omega}_i \sim \boldsymbol{\omega}$ i.i.d. and compute:

$$X_i = \boldsymbol{\omega}_i^{\mathsf{T}}(\boldsymbol{A}\boldsymbol{\omega}_i), \quad \bar{X}_k = \frac{1}{k}\sum_{i=1}^{k} X_i$$

- This leads to an unbiased, consistent estimator of $\mathrm{tr}(\boldsymbol{A})$:

$$\mathbb{E}[\bar{X}_k] = \mathrm{tr}(\boldsymbol{A}), \quad \mathrm{Var}[\bar{X}_k] = \frac{1}{k}\mathrm{Var}[X]$$

- Computational complexity:
  - Simulate k independent copies of $\boldsymbol{\omega}$
  - Perform k matrix-vector multiplications with $\boldsymbol{A}$
  - $O(kn)$ additional arithmetic

Sampling distribution of the estimator

- By the central limit theorem, we have:

$$\begin{aligned} & \sqrt{k}\big(\bar{X}_k - \mathsf{tr}(\boldsymbol{A})\big) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \mathsf{Var}[X]\big) \quad \text{as } k \to \infty \\ \Longrightarrow \quad & \bar{X}_k \sim \mathcal{N}\big(\mathsf{tr}(\boldsymbol{A}), k^{-1}\mathsf{Var}[X]\big) \quad \text{for large } k \end{aligned}$$

- Curse of Monte Carlo: Fluctuations of the order $k^{-\frac{1}{2}}\sqrt{\mathsf{Var}[X]}$

Distribution of the test vector

Lemma

If the coordinates $(\boldsymbol{\omega})_i$ of the test vector $\boldsymbol{\omega} \in \mathbb{R}^n$ are independent samples from a standardized random variable Z, then $\mathbb{E}[\boldsymbol{\omega}\boldsymbol{\omega}^\mathsf{T}] = \boldsymbol{I}_n$ and one single copy of the trace estimator $X = \boldsymbol{\omega}^\mathsf{T}\boldsymbol{A}\boldsymbol{\omega}$ has the variance:

$$\mathrm{Var}[X] = 2\sum_{i\neq j}(\boldsymbol{A})_{ij}^2 + \left(\mathbb{E}[Z^4] - 1\right)\sum_i (\boldsymbol{A})_{ii}^2$$

Important examples:

- Girard estimator $\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$: $\mathbb{E}[Z^4] = 3$

$$\implies \mathrm{Var}[X] = 2\sum_{i,j}(\boldsymbol{A})_{ij}^2 = 2\|\boldsymbol{A}\|_\mathsf{F}^2 \leq 2\|\boldsymbol{A}\|_\mathsf{s}\,\mathrm{tr}(\boldsymbol{A})$$

Distribution of the test vector

Lemma

If the coordinates $(\boldsymbol{\omega})_i$ of the test vector $\boldsymbol{\omega} \in \mathbb{R}^n$ are independent samples from a standardized random variable Z, then $\mathbb{E}[\boldsymbol{\omega}\boldsymbol{\omega}^\top] = \boldsymbol{I}_n$ and one single copy of the trace estimator $X = \boldsymbol{\omega}^\top \boldsymbol{A} \boldsymbol{\omega}$ has the variance:

$$\mathrm{Var}[X] = 2 \sum_{i \neq j} (\boldsymbol{A})_{ij}^2 + \left(\mathbb{E}[Z^4] - 1\right) \sum_i (\boldsymbol{A})_{ii}^2$$

Important examples:

- Hutchinson estimator $\boldsymbol{\omega} \sim \mathcal{U}\{-1, 1\}^n$: $\mathbb{E}[Z^4] = 1$

$$\implies \mathrm{Var}[X] = 2 \sum_{i \neq j} (\boldsymbol{A})_{ij}^2 < 2\|\boldsymbol{A}\|_{\mathsf{F}}^2 \leq 2\|\boldsymbol{A}\|_{\mathsf{s}} \, \mathrm{tr}(\boldsymbol{A})$$

Outline

A priori error estimate

- Chebyshev's inequality implies for $t > 0$:

$$\mathbb{P}\big(|\bar{X}_k - \mathrm{tr}(\boldsymbol{A})| \geq t \cdot \mathrm{tr}(\boldsymbol{A})\big) \leq \frac{\mathrm{Var}[\bar{X}_k]}{t^2\,\mathrm{tr}(\boldsymbol{A})^2} = \frac{\mathrm{Var}[X]}{kt^2\,\mathrm{tr}(\boldsymbol{A})^2}$$

- In case of $\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$:

$$\mathbb{P}\big(|\bar{X}_k - \mathrm{tr}(\boldsymbol{A})| \geq t \cdot \mathrm{tr}(\boldsymbol{A})\big) \leq \frac{2\|\boldsymbol{A}\|_{\mathsf{s}}}{kt^2\,\mathrm{tr}(\boldsymbol{A})} = \frac{2}{kt^2\,\mathrm{intdim}(\boldsymbol{A})}$$

- Here, we interpret

$$\mathrm{intdim}(\boldsymbol{A}) = \frac{\mathrm{tr}(\boldsymbol{A})}{\|\boldsymbol{A}\|_{\mathsf{s}}} = \frac{\sum_{i=1}^n \lambda_i(\boldsymbol{A})}{\lambda_1(\boldsymbol{A})} \in [1, \mathrm{rank}(\boldsymbol{A})]$$

as a continuous measure of the rank.

A sharper error estimate

In some recent work, an exponential probability bound was established for Girard's trace estimator.

Proposition

Consider the trace estimator obtained from a test vector $\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$*. For* $\tau > 1$*, there holds:*

$$\mathbb{P}(\bar{X}_k \geq \tau\, \mathsf{tr}(\boldsymbol{A})) \leq \exp\left(-\frac{1}{2} k\, \mathsf{intdim}(\boldsymbol{A})(\tau^{\frac{1}{2}} - 1)^2\right),$$

$$\mathbb{P}(\bar{X}_k \leq \tau^{-1}\, \mathsf{tr}(\boldsymbol{A})) \leq \exp\left(-\frac{1}{4} k\, \mathsf{intdim}(\boldsymbol{A})(\tau^{-1} - 1)^2\right)$$

Outline

Confidence intervals based on t-distribution

- Idea: Make use of the samples $X_i = \boldsymbol{\omega}_i^\top \boldsymbol{A} \boldsymbol{\omega}_i$
- Compute the sample variance:

$$S_k^2 = \frac{1}{k-1} \sum_{i=1}^{k} (X_i - \bar{X}_k)^2, \quad \mathbb{E}[S_k^2] = \mathsf{Var}[X]$$

- For moderate sample sizes (e.g. $k \geq 30$):

$$\frac{\bar{X}_k - \mathrm{tr}(\boldsymbol{A})}{S_k / \sqrt{k}} \sim t(k-1)$$

- Confidence interval at level $1 - 2\alpha$ (e.g. $\alpha \geq 0.025$):

$$\mathrm{tr}(\boldsymbol{A}) \in \left[\bar{X}_k \pm t_{1-\alpha, k-1} \frac{S_k}{\sqrt{k}}\right]$$

Confidence intervals based on bootstrapping

- Idea: Resample from $\mathcal{X} = (X_1, \ldots, X_k)$ to gain more information about the sampling distribution
- For $b = 1, \ldots, B$:
  - Draw a bootstrap replicate $(X_1^*, \ldots, X_k^*)$ uniformly from $\mathcal{X}$ with replacement
  - Compute the average $\bar{X}_k^* = \frac{1}{k} \sum_{i=1}^{k} X_i^*$ and the error estimate $e_b^* = \bar{X}_k - \bar{X}_k^*$
- Report quantiles q_α and $q_{1-\alpha}$ of the error distribution $(e_1^*, \ldots, e_B^*)$ and the $(1 - 2\alpha)$ confidence interval $[\bar{X}_k + q_\alpha, \bar{X}_k + q_{1-\alpha}]$
- Typical values: $k \geq 30$, $B \geq 1000$, $\alpha \geq 0.025$

Recap - Trace estimation

- Construction of an unbiased and consistent trace estimator, based on randomized sampling
- Runtime advantageous for $k \ll n$
- Distribution of the test vector (normal, Rademacher)
  $\Longrightarrow$ Influence on the variance of the resulting estimator
- A priori error estimates using concentration inequalities
  $\Longrightarrow$ Need for a posteriori bounds in practice
- Confidence intervals based on t-distribution or bootstrapping

Outline

Application to Frobenius and Schatten 4-norm

- Given: Matrix $\boldsymbol{B} \in \mathbb{R}^{m \times n}$ accessed via $\boldsymbol{u} \mapsto \boldsymbol{B}\boldsymbol{u}$
- Draw i.i.d. test vectors $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$ and consider:

$$X_i = \boldsymbol{\omega}_i^\mathsf{T} \boldsymbol{B}^\mathsf{T} \boldsymbol{B} \boldsymbol{\omega}_i, \quad \bar{X}_k = \frac{1}{k} \sum_{i=1}^{k} X_i$$

- Recall that $\bar{X}_k$ is our trace estimator for $\boldsymbol{A} = \boldsymbol{B}^\mathsf{T}\boldsymbol{B}$:

$$\mathbb{E}[\bar{X}_k] = \mathrm{tr}(\boldsymbol{B}^\mathsf{T}\boldsymbol{B}) = \|\boldsymbol{B}\|_\mathsf{F}^2$$

- And the sample variance S_k^2 provides a way to estimate $\|\boldsymbol{B}\|_4$:

$$\mathbb{E}[S_k^2] = \mathrm{Var}[X] = 2\|\boldsymbol{B}^\mathsf{T}\boldsymbol{B}\|_\mathsf{F}^2 = 2\|\boldsymbol{B}\|_4^4$$

Outline

Approach from classical statistics

- Given: Matrix $\boldsymbol{B} \in \mathbb{R}^{m \times n}$ accessed via $\boldsymbol{u} \mapsto \boldsymbol{B}\boldsymbol{u}$
- Goal: Estimate Schatten $2p$-norm $\|\boldsymbol{B}\|_{2p}$ for natural numbers $p \geq 3$

- Sample i.i.d. isotropic test vectors $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_k \in \mathbb{R}^n$ to extract linear information $\boldsymbol{Y}_i = \boldsymbol{B}\boldsymbol{\omega}_i$

  $\Longrightarrow$ Sample matrix $\boldsymbol{Y} = [\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_k] \in \mathbb{R}^{m \times k}$

- Define $\boldsymbol{X} = \boldsymbol{Y}^\mathsf{T}\boldsymbol{Y} \in \mathbb{R}^{k \times k}$ and check that $(\boldsymbol{X})_{ij} = \boldsymbol{\omega}_i^\mathsf{T}\boldsymbol{B}^\mathsf{T}\boldsymbol{B}\boldsymbol{\omega}_j$

Approach from classical statistics

- For any $1 \le i_1, \dots, i_p \le k$, we have:

$$\begin{aligned}
&(\boldsymbol{X})_{i_1 i_2} (\boldsymbol{X})_{i_2 i_3} \cdots (\boldsymbol{X})_{i_p i_1} \\
&= \operatorname{tr}\big(\boldsymbol{\omega}_{i_1}^\top \boldsymbol{B}^\top \boldsymbol{B} \boldsymbol{\omega}_{i_2} \cdot \boldsymbol{\omega}_{i_2}^\top \boldsymbol{B}^\top \boldsymbol{B} \boldsymbol{\omega}_{i_3} \cdots \boldsymbol{\omega}_{i_p}^\top \boldsymbol{B}^\top \boldsymbol{B} \boldsymbol{\omega}_{i_1}\big) \\
&= \operatorname{tr}\big(\boldsymbol{\omega}_{i_1} \boldsymbol{\omega}_{i_1}^\top \boldsymbol{B}^\top \boldsymbol{B} \cdot \boldsymbol{\omega}_{i_2} \boldsymbol{\omega}_{i_2}^\top \boldsymbol{B}^\top \boldsymbol{B} \cdots \boldsymbol{\omega}_{i_p} \boldsymbol{\omega}_{i_p}^\top \boldsymbol{B}^\top \boldsymbol{B}\big)
\end{aligned}$$

- If $i_1, \dots, i_p$ are distinct, use independence and isotropy to compute:

$$\begin{aligned}
&\mathbb{E}\big[(\boldsymbol{X})_{i_1 i_2} (\boldsymbol{X})_{i_2 i_3} \cdots (\boldsymbol{X})_{i_p i_1}\big] \\
&= \operatorname{tr}\big(\mathbb{E}\big[\boldsymbol{\omega}_{i_1} \boldsymbol{\omega}_{i_1}^\top \boldsymbol{B}^\top \boldsymbol{B} \cdot \boldsymbol{\omega}_{i_2} \boldsymbol{\omega}_{i_2}^\top \boldsymbol{B}^\top \boldsymbol{B} \cdots \boldsymbol{\omega}_{i_p} \boldsymbol{\omega}_{i_p}^\top \boldsymbol{B}^\top \boldsymbol{B}\big]\big) \\
&= \operatorname{tr}\big(\mathbb{E}\big[\boldsymbol{\omega}_{i_1} \boldsymbol{\omega}_{i_1}^\top \boldsymbol{B}^\top \boldsymbol{B}\big] \cdot \mathbb{E}\big[\boldsymbol{\omega}_{i_2} \boldsymbol{\omega}_{i_2}^\top \boldsymbol{B}^\top \boldsymbol{B}\big] \cdots \mathbb{E}\big[\boldsymbol{\omega}_{i_p} \boldsymbol{\omega}_{i_p}^\top \boldsymbol{B}^\top \boldsymbol{B}\big]\big) \\
&= \operatorname{tr}\big((\boldsymbol{B}^\top \boldsymbol{B})^p\big) = \|\boldsymbol{B}\|_{2p}^{2p}
\end{aligned}$$

Approach from classical statistics

- Average over all sequences of distinct indices to obtain an unbiased estimator of $\|\boldsymbol{B}\|_{2p}^{2p}$:

$$U_p = \frac{(k-p)!}{k!} \sum_{1 \le i_1, \ldots, i_p \le k}^{\circ} (\boldsymbol{X})_{i_1 i_2} (\boldsymbol{X})_{i_2 i_3} \cdots (\boldsymbol{X})_{i_p i_1}$$

- Computationally demanding:
  - Almost k^p summands
  - Only feasible for small values of p (e.g. $p = 4$ or $p = 5$)
- High variability (according to theory of U-statistics):

$$k \,\mathsf{Var}[U_p] \to 2p^2 \|\boldsymbol{B}\|_{4p}^{4p} \quad \text{as } k \to \infty$$

Outline

Computationally efficient method

- Recall that for distinct indices $i_1, \ldots, i_p$:

$$\mathbb{E}\left[(\boldsymbol{X})_{i_1 i_2}(\boldsymbol{X})_{i_2 i_3}\cdots(\boldsymbol{X})_{i_p i_1}\right] = \|\boldsymbol{B}\|_{2p}^{2p}$$

- Average only over increasing sequences of indices:

$$V_p = \binom{k}{p}^{-1} \sum_{1 \le i_1 < \cdots < i_p \le k} (\boldsymbol{X})_{i_1 i_2}(\boldsymbol{X})_{i_2 i_3}\cdots(\boldsymbol{X})_{i_p i_1}$$

- Equivalent formulation:

$$V_p = \binom{k}{p}^{-1} \mathrm{tr}\left[\boldsymbol{X}_{\mathsf{U}}^{p-1}\boldsymbol{X}\right]$$

where $\boldsymbol{X}_{\mathsf{U}}$ denotes the strict upper triangular part of $\boldsymbol{X}$

Computationally efficient method

- Estimator V_p suffers from even higher variance than U_p
- But substantially cheaper in computation:
  - Matrix exponentiation via repeated squaring
  - Method is feasible for larger values of p
  - Employ more samples to account for the high variance
- Apply resampling methods to construct confidence intervals

Outline

Summary - Schatten norm estimation

- Estimate Frobenius norm and Schatten 4-norm for free using techniques from trace estimation
- Unbiased estimators of $\|\boldsymbol{B}\|_{2p}^{2p}$ based on randomized sampling
- Estimator U_p: High complexity and high variability
- Reduction to estimator V_p whose computation is rather cheap

Outlook

- Parts of the algorithms are *parallelizable*
- Flexibility in designing the test vector distribution:

  Sample from an *optimal measurement system* to minimize the variance and the number of required random bits

References

- P. G. Martinsson and J. A. Tropp (2020), 'Randomized numerical linear algebra: Foundations and algorithms', Acta Numerica 29, 403–572.
- D. A. Girard (1989), 'A fast "Monte Carlo cross-validation" procedure for large least squares problems with noisy data', Numer. Math. 56, 1–23.
- S. Gratton and D. Titley-Peloquin (2018), 'Improved bounds for small-sample estimation', SIAM J. Matrix Anal. Appl. 39, 922–931.
- W. Kong and G. Valiant (2017), 'Spectrum estimation from samples', Ann. Statist. 45, 2218–2247.
- J. Gedicke (2021), Lecture notes: Introduction to Basic Principles of Numerics, University of Bonn.
- http://math.uchicago.edu/ may/REU2016/REUPapers/Mazeika.pdf (last accessed: March 26, 2022)
- https://math.ntnu.edu.tw/ jschen/Papers/schatten-p-norm-JNCA.pdf (last accessed: March 26, 2022)