Is global moment matching the optimal scheme for spectrum estimation from samples?

Tiancheng Yu† Jiantao Jiao*, Yanjun Han* and Tsachy Weissman*

†Tsinghua University *Stanford University

Spectrum Estimation: See the shape of data!

- *What is*:
  - "spectrum": the sorted eigenvalues of any symmetric matrix, which are assumed to be not larger than **b**.
  - "spectrum estimation": estimate the eigenvalues of a given d-dimensional symmetric matrix under normalized sorted loss

$$L\left(\lambda, \hat{\lambda}\right) = \frac{1}{d}\sum_{i=1}^{d}\left|\hat{\lambda}_i - \lambda_i\right| \tag{1}$$

  where λ_i is the i-th largest eigenvalue of the underlying matrix and $\hat{\lambda}_i$ is the corresponding estimator.
  - "from samples": we do not have access to the matrix itself, but only know it is the covariance matrix of the given samples of size **n**.
- *Why*: In high-dimensional data analysis, it is extremely important to reduce the dimension of the observation to compute faster and generalize better. To this end, we have to estimate the spectrum of the covariance matrix and the underlying distribution, which determines the "shape" of the distribution.

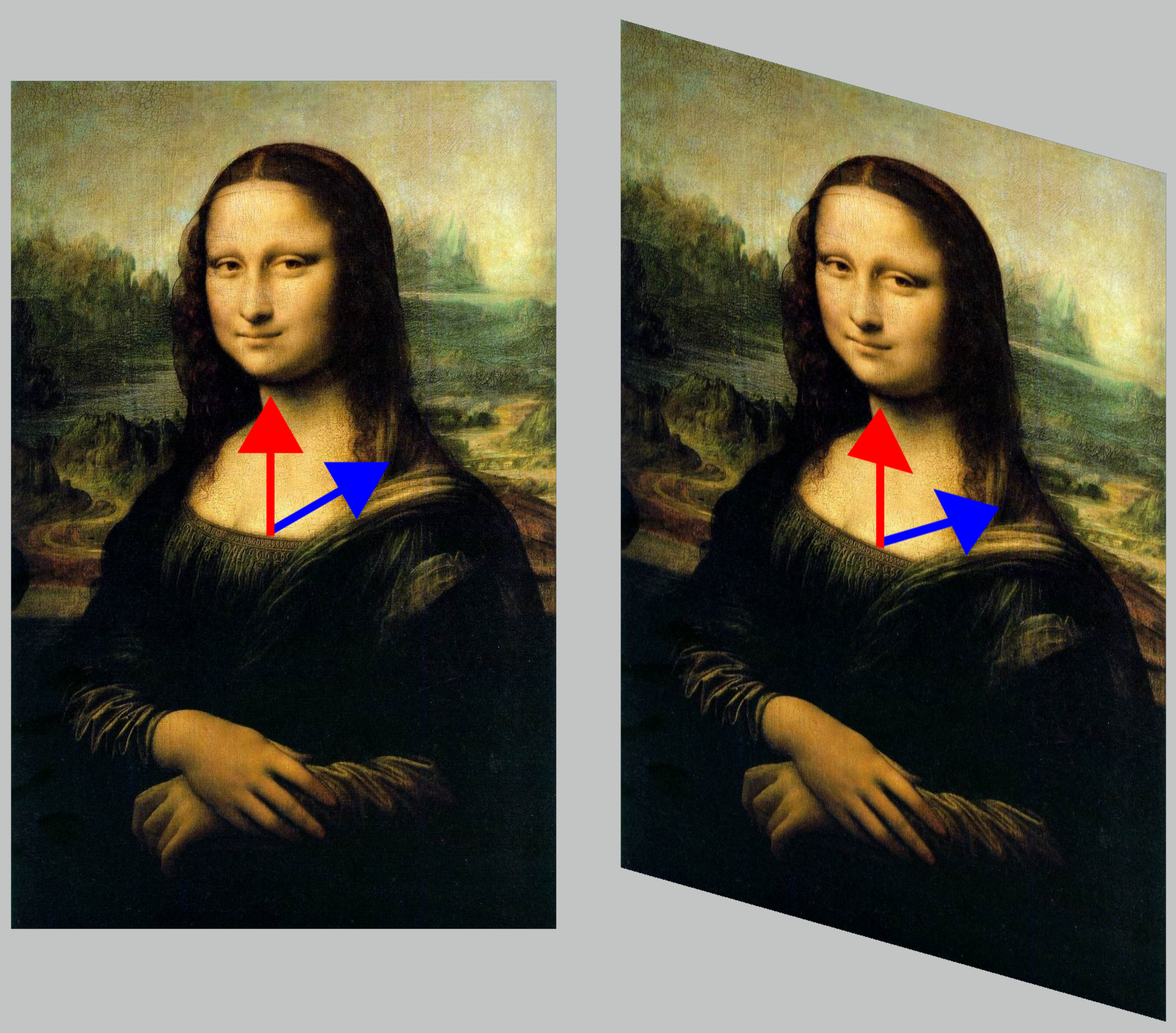


Figure: Eigenvalue determines the "shape" of the underlying transform. Here we do not change the length of eigenvalues, so only the direction changes.

Moment matching: Infer mazy distribution from simple features!

- *What is*
  - "moment": the power sum of eigenvalues in the following form

$$S_k = \sum_{i=1}^{d}\lambda_i^k \tag{2}$$

  - "moment matching": In [1], a nice methodology called moment matching is proposed. They first estimate the moments of order k for $k = 1, 2, \cdot, L$ and then search for a set of eigenvalues with similar moments.
- *Why*:
  - *Why moment matching can be done*: Although finding unbiased estimators for the eigenvalues are difficult, finding the unbiased estimator for the moment is easier.
  - *Why moment matching works*: The normalized sorted loss is equivalent to the Wasserstein distance, which is small of the when the first L order matched [2].
- *How to do better*: Since the moment matching in [1] is done in the whole interval, we call it "global moment matching", which induces larger approximation error. If we can concentrate the eigenvalues in smaller intervals and do moment matching locally (referred to as "local moment matching" below), the approximation error can be mitigated.

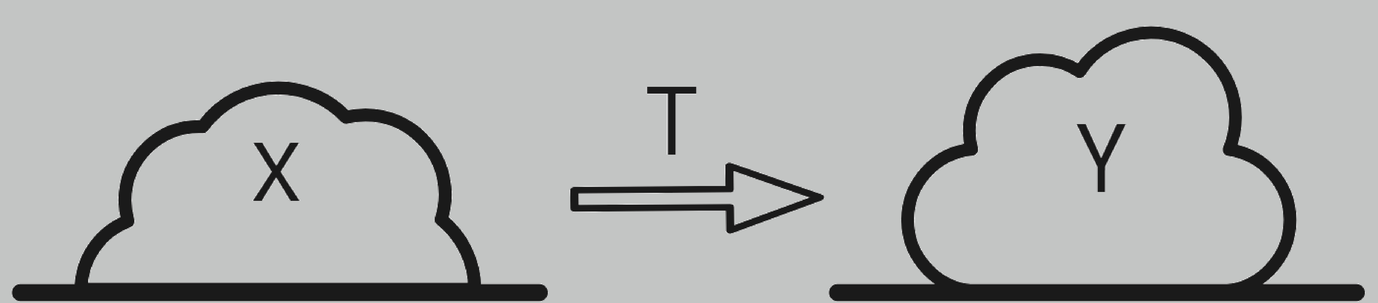


Figure: Wasserstein distance argument: moment matching implies similar distribution!

Independent case: global moment matching fails!

We first study a special case, i.e., when the different coordinates of the observations are independent. In this case, the local moment matching can be done and we prove the minimax optimal rate is

$$\inf_{\hat{\lambda}}\sup_{\lambda}\mathbb{E}\left\{\sum_{i=1}^{d}\left|\hat{\lambda}_i - \lambda_i\right|\right\} \asymp b\frac{d}{\sqrt{n\ln n}} \tag{3}$$

when $d \gtrsim \sqrt{n\ln n}$. However, the minimax rate of global moment matching is only

$$\sup_{\lambda}\mathbb{E}\left\{\sum_{i=1}^{d}\left|\hat{\lambda}_i - \lambda_i\right|\right\} \asymp b\frac{d}{\ln n} \tag{4}$$

In this case, the global moment matching fails. Indeed, its worst case performance is worse than the naive plug-in estimator. However, we proved in such case, the local moment matching method can achieve the minimax lower bound.

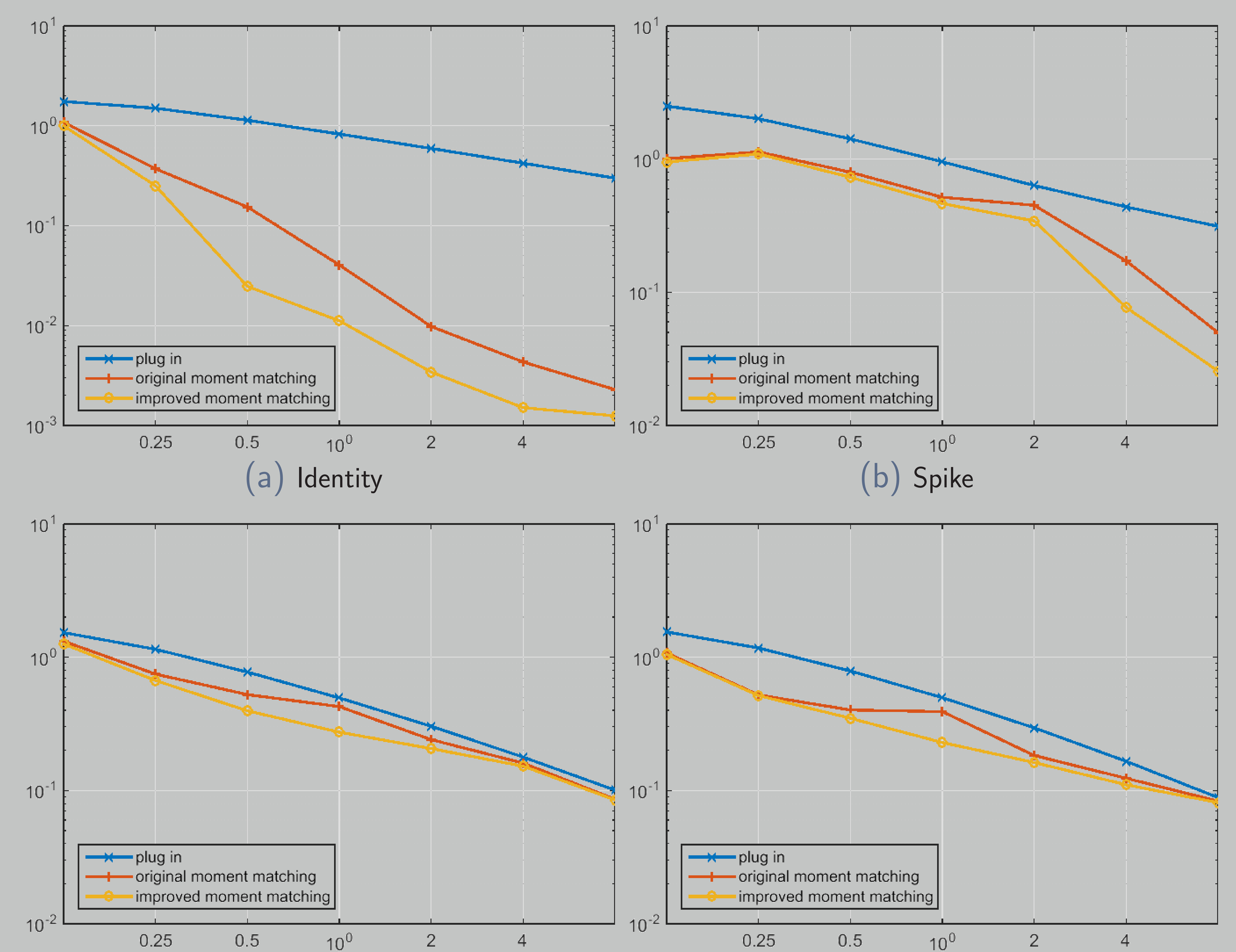
Our approach: improved upper bound and empirical performance!

We also improve the original global moment matching and achieved a better upper bound.

$$\mathbb{E}\left\{\sum_{i=1}^{d}\left|\hat{\lambda}_i - \lambda_i\right|\right\} \lesssim b\left[\left(\frac{Ckd}{n}\right)^{k/2} + \frac{d}{k}\right] \tag{5}$$

where we match the first **k** moments and **C** is a universal constant. We also find a lower bound and adaptive estimation.

To show we achieve better empirical performance, we compare the performance of the plug-in (SVD) estimator, original and improved moment matching in four different scenarios. The horizontal axis is the sample-dimension ratio n/d and the vertical axis represent the normalized sorted loss between the estimated and true spectrum.



(a) Identity (b) Spike (c) Uniform (d) Teoplitz

Figure: The moment matching method improve the performance significantly with non-sparse spectrum but marginally with sparse spectrum. This is because in sparse case, the spectrum of the sample covariance matrix concentrates to that of the population covariance matrix quickly.

Can we do better? Hopefully no!

Our intuition based on previous functional estimation problem and the independent case is: we should do moment matching on the smallest interval we can concentrate. Since in general, it is impossible to concentrate, we conjecture the global moment matching method can achieve the minimax lower bound (still under construction).

References

[1] Spectrum estimation from samples, kong, weihao and valiant, gregory, 2016.

[2] *An introduction to the approximation of functions, Rivlin, Theodore J, 2003.*

http://yutc.me

yutc@stanford.edu