27 January 2025 · CC-BY 4.0

Qeios

Peer Review

# Review of: "Distributional Matrix Completion via Nearest Neighbors in the Wasserstein Space"

**Bohan Zhou**[1]

[1] University of California, Santa Barbara, United States

The authors address the matrix completion problem in a novel setting where each matrix entry is a vector constructed from independent and identically distributed draws from a distribution. This is a natural extension of the classical matrix completion problem where entries are scalars. Each row corresponds to a user (e.g., a school in the figure), and each column corresponds to an item (e.g., the score distribution for a subject). To impute missing empirical distributions (e.g., the score distribution of English for School A), the authors propose using the Wasserstein barycenter of the nearest neighbors' empirical distributions (e.g., English score distributions from nearby schools). Here, a school is considered a neighbor if the distance between School X and School A is within a threshold, and the distances are calculated as the average of Wasserstein distances between the score distributions of available subjects from School X and School A.

Authors present two main theoretical results:

- Theorem 1 as an asymptotic approximation between the estimated missing empirical distribution and the true distribution in terms of Wasserstein distance, under the Lipschitz latent factor in Assumption 1, missing-completely-at-random in Assumption 2, and Regularity in Assumption 3.
- Theorem 2 as an asymptotic convergence in distribution between the quantiles for the estimated missing empirical distribution and the true distribution, under the similar assumptions.

Authors provide numerical examples in 1D, along with accompanying GitHub code. Given the computational complexity of the Wasserstein distances, which are required twice in the authors' method, they rely on closed-form solutions for both the Wasserstein distance and the Wasserstein barycenter in 1D using quantiles. In general, computing a Wasserstein barycenter involves

doi.org/10.32388/73HJ02

determining an optimal combination of samples from each marginal distribution and locating the support points via the Euclidean barycenter map. However, in 1D, the optimal combination is directly determined by the quantiles of the marginal distributions, simplifying the computation. This may also prevent an easy extension to multi-dimensions.

I think the theoretical results are complete and the accompanying remarks are insightful, though I personally didn't track the up-to-date asymptotic analysis results very closely in such a field.

**Scientific Questions:**

1. How realistic is Assumption 1 in the context of the motivating example? Can the authors provide a detailed explanation of such an assumption for the example in Figure 2?
2. What is the motivation for assumption (vi) in the regularity assumption 3?
3. What roles do correlated columns play in the analysis? Since one motivation is to study the influence of digital resources, could the authors clarify how column correlations are addressed?

**Suggestions:**

1. If I didn't overlook anything, I think the Wasserstein barycenter with weights by the average distance between rows appears to be no more difficult than the current version with equal weights. If so, could the authors discuss the theoretical consequences?
2. Including real-world applications could strengthen the justification for using Wasserstein distance as the central measure in this problem. This would help contextualize its necessity. For example, people may satisfy the L2 distance for a standardized test like the SAT.

## Declarations

**Potential competing interests:** No potential competing interests to declare.

doi.org/10.32388/73HJ02