Research Article

Distributional Matrix Completion via Nearest Neighbors in the Wasserstein Space

Jacob Feitelberg[1], Kyuseong Choi[2], Anish Agarwal[1], Raaz Dwivedi[2]

1. Columbia University, United States; 2. Cornell Tech, New York, United States

We introduce the problem of distributional matrix completion: Given a sparsely observed matrix of empirical distributions, we seek to impute the true distributions associated with both observed and unobserved matrix entries. This is a generalization of traditional matrix completion where the observations per matrix entry are scalar valued. To do so, we utilize tools from optimal transport to generalize the nearest neighbors method to the distributional setting. Under a suitable latent factor model on probability distributions, we establish that our method recovers the distributions in the Wasserstein norm. We demonstrate through simulations that our method is able to (i) provide better distributional estimates for an entry compared to using observed samples for that entry alone, (ii) yield accurate estimates of distributional quantities such as standard deviation and value-at-risk, and (iii) inherently support heteroscedastic noise. We also prove novel asymptotic results for Wasserstein barycenters over one-dimensional distributions.

Corresponding authors: Jacob Feitelberg, jef2182@columbia.edu; Kyuseong Choi, kc728@cornell.edu

1. Introduction

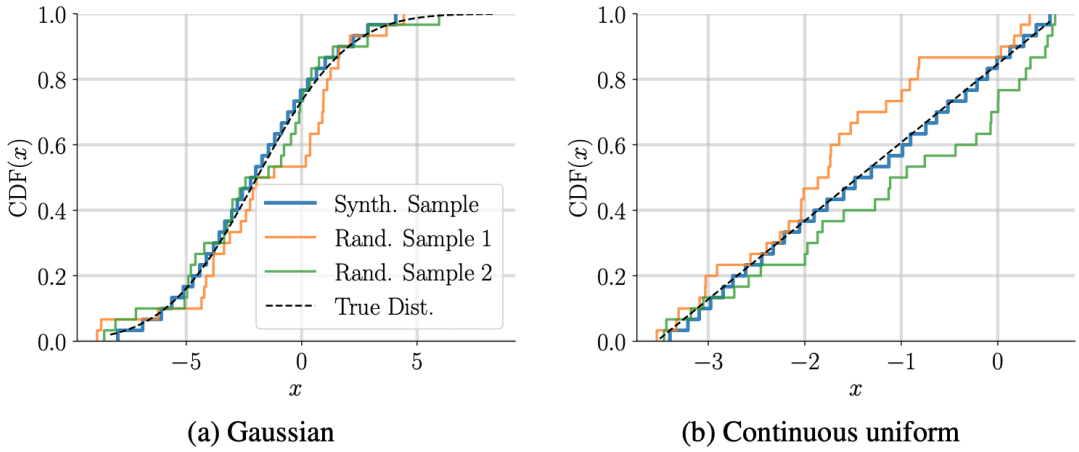
Matrix completion is the broad problem of imputing missing entries in a matrix. Algorithms for this problem have found widespread use in recommendation systems[1][2][3] used at companies such as Netflix, Amazon, and Meta, system identification[4], traffic sensing[5][6][7], device location sensing[8][9], and patient-level predictions in healthcare[10][11]. While the theory and practice of matrix completion is thoroughly researched, there has been little to no work on matrix completion over distributions of numbers. We refer to this new problem as *distributional matrix completion*.

 doi.org/10.32388/DX6MDB

To motivate this problem, consider the following example: a state education-board is trying to determine if using digital resources improves student test scores. They collect test-score and digital resource use data from schools going back 5 years for English, math, and science classes. The board would like to estimate the impact of digital resources on test scores at the classroom level. Since data is collected over 5 years, the same school could have test-score data for English, for instance, both using and not using digital resources. However, not every school will have tried each combination in the past 5 years. This impedes the board's ability to determine the impact of digital resources at the school level. However, since they have data from many schools across the state, they can use data imputation methods such as matrix completion to estimate the missing test scores.

Matrix completion methods have only been developed for scalar values, though, whereas test scores are inherently a distribution of values because there is more than one student in a class. One way to get around this problem is to consider *average* test scores per class. However, if later on, the board wants to know if average test scores went up because the highest-scoring students tested better or if the all students did better simultaneously, then we must move beyond averages (scalars) into the space of distributions. The goal of our paper is to explore how we can better exploit repeated measurements to (i) learn the underlying distributions associated with each matrix entry and (ii) better predict commonly used quantities such as median, variance, and value-at-risk. Distributional matrix completion's difficulty stems from two information losses: we only observe a subset of the distributions and for the distributions we do observe, we only have access to a finite number of samples, not the true distributions. Additionally, distributional objects exist in infinite-dimensional spaces, adding to the difficulty of extending formal matrix completion setups.

In this paper, we propose both a formal setup for distributional matrix completion and an estimation method to recover the unobserved true distribution per matrix entry. Utilizing tools from optimal transport, our method is able to generate synthetic distributions that closely approximate the respective true distributions. Further, perhaps surprisingly, the estimates consistently recover the true distributions more accurately compared to using just the empirical distribution for an observed matrix entry—see Figure 1. This allows for more accurate estimation of downstream distributional quantities such as variance or value-at-risk compared to simply using the observed empirical distribution.



(a) Gaussian (b) Continuous uniform

Figure 1. CDFs of random empirical distributions vs. our synthetic distribution from matrix completion. We simulate two distributional matrix completion examples, one with all empirical Gausian distributions and one with empirical continuous uniform distributions. In both cases, only the matrix entry we seek to estimate is unobserved. Our method's synthetic distributions, shown in the thick blue lines, provide much better estimates of the true distribution's CDF, shown in black dotted lines, than observed matrix entries alone. See Sec. 5 for more details on our simulated tests.

1.1. Related work

Our work seeks to bridge two disparate topics, matrix completion and optimal transport, in order to provide a way to estimate unobserved probability distributions. Here, we provide a brief overview of the relevant literature in both areas.

Matrix completion. There are many algorithms for matrix completion that mainly fall into two categories: empirical risk minimization (ERM) and matching. Empirical risk minimization (ERM) methods seek to minimize both the distance between estimated matrix entries and observed matrix entries along with a regularization term[12][13]. The regularizer seeks to prioritize less complex matrices and is sometimes replaced with a hard constraint, such as the matrix being low-rank[14].

Matching methods, or nearest neighbor methods, are popular for large-scale recommendation systems due to their simplicity and scalability[15][16]. These algorithms estimate a missing entry by finding "similar" rows (users) or columns (items) and then use their average as the estimate for a missing entry. These algorithms not only work well in practice, but have been shown to work under suitable latent-

factor models[17][18]. To implement matching methods, we only need to generalize a notion of similarity and averaging between matrix entries. We show how we do this in Sec. 3.

The matrix completion literature has also grown to include noisy matrix completion[19][17], panel data settings for causal inference[18][13], and even matrix completion over more exotic spaces such as finite fields[20]. These areas, though, still assume the matrix has scalar values.

Optimal transport. Optimal transport (OT), a field initially developed to solve logistics problems[21], provides efficient computational and theoretical tools to compute the distance between two probability distributions and the average between a collection of distributions. OT has found widespread use from areas such as shape and image registration[22] to systems control[23]. Most work in OT has focused on efficiently finding the map to transport one distribution to another at the lowest cost for a given cost function[24][25]. In this paper, we focus on the 2-Wasserstein distance, which is the cost of transporting one distribution to another when the cost function is the Euclidean distance. The Wasserstein distance also lends itself well to calculating distributional barycenters (i.e. averages)[26][27][28]. Wasserstein barycenters preserve the geometric properties of the input distributions. For instance, the barycenter of multiple Gaussian distributions is also Gaussian, unlike a mixture of Gaussians[29]. Thus, in our setting, if every observed distribution is Gaussian, we would expect the estimated distribtuions to also be Gaussian and not a mixture of Gaussians. While Wasserstein barycenters have a closed-form solution for one-dimensional distributions, they have been proven to be NP-Hard to calculate in high dimensions[30].

In classical optimal transport literature, it is assumed that distributions are known completely, instead of sampled. The statistics of OT, an area that has become increasingly popular, focuses on comparing how empirical distributios generated from samples differ from their respective true distributions[29][31][32]. This literature works with empirical distributions as objects as opposed to using a noise model. Because we only have access to empirical distributions, of particular use for this paper are the sample-size convergence rates of empirical distributions to true distributions proven in[33].

Matrix completion over distribution functions. We are aware of only one previous use of optimal transport for estimating unobserved distributions in the matrix setting:[34] present a method for learning counterfactual distributions in a synthetic control setting[35] using a quantile-based generalization of the synthetic control method. They perform a regression over quantile functions to construct their estimate, whereas we use a nearest neighbors method.[34] also assumes a linear mapping between latent factors and distributions whereas we allow our mapping to be nonlinear. They make

similar assumptions to ours on their distributions for their theoretical results such as each distribution's density being differentiable and lower bounded by a positive constant. However, we also assume several stronger regularity conditions in order to provide faster rates of error decay.

1.2. Organization and notation

In Sec. 2, we propose a Lipschitz latent function model for matrix completion over the space of one-dimensional probability distributions. In Sec. 3, we provide a distributional nearest neighbor estimation method to solve our new problem utilizing the geometry of the 2-Wasserstein space. In Sec. 4, we provide asymptotic error bounds for our estimator and establish its consistency as the number of rows and columns of the matrix, and the number of samples for a given observed matrix entry grow. We also establish the asymptotic distribution for the error of the estimand. In Sec. 5, we use simulations to empirically verify our theoretical error decay rates. We also demonstrate our method's accuracy in estimating distributional quantities such as mean, standard deviation, and quantiles. Finally, in Sec. 6, we conclude and discuss future research directions.

Notation. We refer to a probability measure μ's cumulative distribution function as F_μ, its quantile function as F_μ^{-1}, and its density function as f_μ. We write $X_n = \mathcal{O}_p(a_b)$ when X_n/a_b is bounded in probability. We write $X_n = o_p(a_b)$ when X_n/a_b converges to 0 in probability. We denote the set $\{1, \ldots, m\}$ as $[m]$. We denote the (i,j)-th entry of a matrix M as M_{ij}. We abuse notation slightly by referring to a set of samples and its respective empirical distribution with the same symbol when there is no ambiguity as to which object we are referring to. We abbreviate the term *almost surely* to *a.s.* and write $\overset{d}{=}$ to denote equality in distribution. Finally, in regards to our algorithm, we use the words *user* and *row* interchangeably. We do the same with the words *item* and *column*. This originates from the literature for recommendation systems which typically have users along rows and items along columns.

2. Setup and data-generating process

In this section, we describe the general setup of distributional matrix completion. We then review some necessary background on optimal transport, which is used in our error calculations and estimation procedure. We then propose a data-generating process (DGP) that allows us to provide an error bound and asymptotic distribution for our method in Sec. 4.

doi.org/10.32388/DX6MDB

2.1. Problem setup and generic nearest neighbors

In our setup, we analyze a partially observed $N+1$ by $M+1$ matrix, denoted Y, where each observed matrix entry contains an array of scalars. Within each matrix entry, Y_{ij}, the scalars are assumed to be drawn independently and identically (i.i.d.) according to some law μ_{ij}. For each matrix entry in a column j, we assume that number of samples is n_j. However, our algorithm and theoretical guarantees can easily generalize to unequally sized arrays within columns. If a matrix entry (i, j) is observed, then we denote the samples in that matrix entry as $\{y_{ij,k}\}_{k=1}^{n_j} \sim \mu_{ij}$. Note that for observed entries, we only have access to empirical data, not the true distribution. Note that we simply add 1 to the matrix size to simplify the notation in our main theorem. Using the observed empirical distributions, our goal is to estimate the true distributions, μ_{ij}, of the unobserved and observed matrix entries.

Missingness. Let A be an $N+1$ by $M+1$ binary matrix representing which entries are observed and which are not. Then, we have

$$\text{for} \quad i \in [N+1], j \in [M+1]: \qquad Y_{ij} = \begin{cases} [y_{ij,1}, \ldots, y_{ij,n_j}] & \text{if} A_{ij} = 1 \\ \text{unknown} & \text{if} A_{ij} = 0 \end{cases}$$

where $\{y_{ij,k}\}_{k=1}^{n_j} \sim \mu_{ij}$.

To explain this formal notation with the motivation example from Sec. 1, we present a matrix of histograms in Fig. 2. In this example, the input Y is a partially observed 3 by 6 matrix. There are four unobserved distributions. Our goal is to estimate the true distributions (gray) for both observed and unobserved entries using the observed empirical distributions (blue). This will allow us to better analyze the effect of digital resources in different class types for each school.

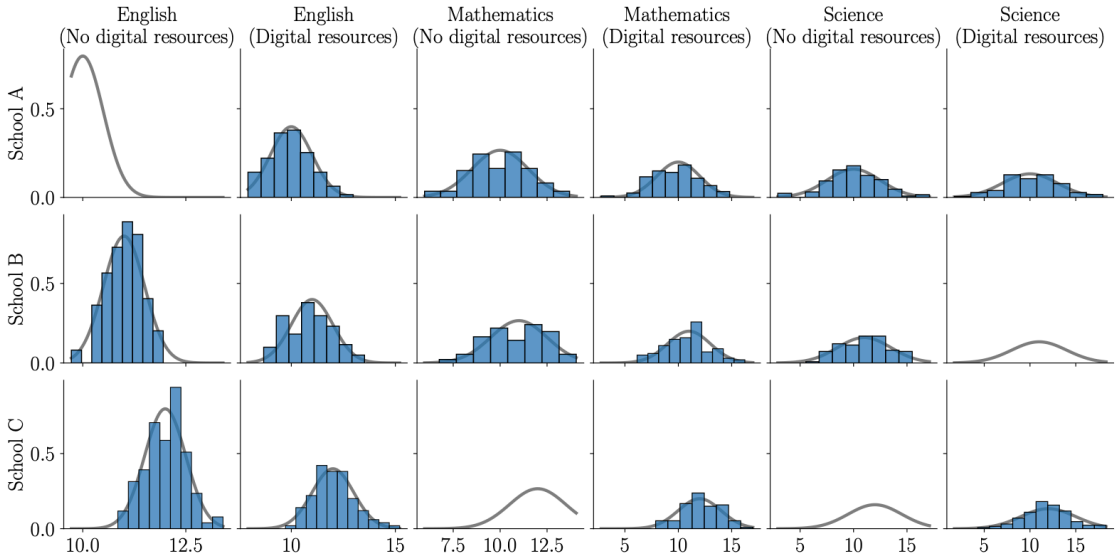


Figure 2. Simulated motivating example of a student test scores from multiple schools in different classes with and without digital resources. We visualize the motivating example from Sec. 1 with a matrix of empirical histograms with 4 unobserved entries shown without histograms. Each observed entry has 100 independent samples. The true distribution's probability density functions are shown in gray curves. Our goal is to not only estimate the unobserved distributions, but also the true distributions for the observed entries.

Generic user-user scalar nearest neighbors algorithm. The nearest neighbors algorithm requires a distance threshold, η, to define a neighborhood. The nearest neighbors algorithm then proceeds in two main steps impute a matrix entry (i, j):

(1) **Find the set of nearest neighbors for row** i**.** For each other row $u \neq i$ where $A_{uj} = 1$, define the columns that are observed in both rows i and u as

$$\mathcal{C}_{iu} \triangleq \{v \in [M+1] \setminus \{j\} : A_{iv} = 1, A_{uv} = 1\}.$$

Then, find the average distance between the row i and row u as:

$$\rho_{iu} \triangleq \begin{cases} |\mathcal{C}_{iu}|^{-1} \sum_{v \in \mathcal{C}_{iu}} (Y_{iv} - Y_{uv})^2 & \text{if} \sim |\mathcal{C}_{iu}| \geq 1 \\ \infty & \text{if} \sim |\mathcal{C}_{iu}| = 0. \end{cases}$$

Finally, define row i's η-nearest neighbors as

$$\mathcal{N}_{i,\eta} \triangleq \{u \in [N+1] : A_{uj} = 1, \rho_{iu} \leq \eta\}.$$

(2) **Find the average of the nearest neighbors in column** j**.** If $\mathcal{N}_{i,\eta} > 0$, then estimate the missing entry at (i, j) as

$$\hat{Y}_{ij} = \frac{1}{\left|\mathcal{N}_{i,\eta}\right|} \sum_{u \in \mathcal{N}_{i,\eta}} Y_{uj}.$$

If we found no nearest neighbors, then return that we could not find any neighborhood for row i.

Step 1 finds rows who have matrix values close to the observed entries in the same row we are trying to estimate in. It then defines the nearest neighbors as the rows with distance below η. Step 2 returns the average of the nearest neighbors, $\mathcal{N}_{i,\eta}$, in column j. To generalize this algorithm to the distributional setting, we need to provide distributional versions of 3 and 5.

Note that in this paper, we analyze user-user nearest neighbors. However, our work can be easily extended to user-item or item-item nearest neighbors. In item-item nearest neighbors, distances are calculated between columns and averages are taken over rows. In user-item nearest neighbors, distance and averages are taken both over rows and columns. See[18] for an example of how a user-item nearest neighbors algorithm is used to construct a doubly-robust estimator.

To generalize the nearest neighbors method from scalar matrix completion to the distributional setting, we require a notion of *distance* and *average* in the probability distribution space. While there are many distributional distances such as total variation and Kullback–Leibler divergence, we utilize the Wasserstein distance and barycenter from optimal transport for this new setting. It remains an interesting line of future research to explore the statistical and computational properties of other distributional distances.

2.2. Wasserstein distance background

The Wasserstein distance is a natural choice for distributional nearest neighbors because (i) it satisfies the properties of a metric, (ii) it has a closed-form solution in the one-dimensional case, and (iii) it behaves well when distributions do not share supports. For instance, for total variation, denoted TV, $TV(U(0, 1), U(2, 3)) = TV(U(0, 1), U(4, 5)) = 0$ because these continuous uniform distributions do not share supports. The 2-Wasserstein distance, denoted W_2, however, has $W_2(U(0, 1), U(2, 3)) < W_2(U(0, 1), U(4, 5))$. For other examples of the Wasserstein distance's useful geometric properties, see[36].

In the one-dimensional setting, the 2-Wasserstein distance can be written as an L^2 norm between quantile functions: For two probability measures on R, μ and ν with finite second moment, we have[29], Eq. 2]

$$W_2(\mu, \nu) = \left(\int_0^1 |F_\mu^{-1}(x) - F_\nu^{-1}(x)|^2 dx\right)^{1/2} = \| F_\mu^{-1} - F_\nu^{-1} \|_{L^2(0,1)}$$

where F_{μ}^{-1} and F_{ν}^{-1} are the respective quantile functions of μ and ν. For empirical distributions with the same number of samples, n, we have a simpler formula[33, Lemma 4.2]:

$$W_2(\mu_n, \nu_n) = \left(\frac{1}{n} \sum_{i=1}^{n} \left(X^{(i)} - Y^{(i)} \right)^2 \right)^{1/2}$$

where $X^{(i)}$ and $Y^{(i)}$ are the i-th order statistic of their respective empirical distributions μ_n and ν_n. Thus, the Wasserstein distance can be calculated in $\mathcal{O}(n\log(n))$ time between any two empirical distributions with the same number of samples, n, with the asymptotic runtime being dominated by the sorting operation. Note that we interchangeably use "Wasserstein" and "2-Wasserstein" as we only consider the 2-Wasserstein distance in this work. We denote $W_2(\mathrm{R})$, the *Wasserstein space*, as the space of one-dimensional probability distributions on R with finite second moment equipped with the 2-Wasserstein metric.

2.3. Wasserstein barycenter background

Let $\mu_1, \ldots, \mu_N \in W_2(\mathrm{R})$ be probability distributions. The Wasserstein barycenter is defined as the probability distribution μ that minimizes $\sum_{i=1}^{N} W_2^2(\mu, \mu_i)$, similar to how an average over scalars minimizes the squared distance to each scalar. The Wasserstein barycenter also has a simple closed-form solution as the measure with quantile function[29, Eq. 8]:

$$F_{\mu}^{-1} = \frac{1}{N} \sum_{i=1}^{N} F_{\mu_i}^{-1}.$$

When each distribution μ_j is an empirical distribution derived from order statistics $\{X_{\mu_j}^{i}\}_{i=1}^{n}$, then the Wasserstein barycenter's distribution is an empirical distribution derived from order statistics given by[37, Section 2.4]:

$$X_{\mu}^{(i)} = \frac{1}{N} \sum_{j=1}^{N} X_{\mu_j}^{(i)}.$$

To calculate the k-th order statistic of the Wasserstein barycenter, we first sort each distribution's data into respective order statistics, and then average the k-th order statistics of the input distributions. Ordering each entry's samples takes $\mathcal{O}(N \cdot n\log(n))$ time and calculating the order statistic average takes $\mathcal{O}(Nn)$ time. So, the runtime to calculate the barycenter is $\mathcal{O}(N \cdot n\log(n))$. The Wasserstein barycenter has several desirable properties: (i) it is computationally fast to calculate for empirical distributions, (ii) it behaves well with location-scale distributions such as Gaussian and continuous uniform, and (iii) it has a

closed-form solution in the one-dimensional case which facilitates theoretical analysis. As an example of the Wasserstein barycenter's geometric properties, it is shown in[29] that the barycenter of multiple Gaussian distributions is also Gaussian.

2.4. Data-generating process

Under no assumptions about the matrix, matrix completion is ill-posed since there are too many valid ways to impute missing entries. Therefore, matrix entries are often assumed to share some latent structure which reduces the degrees of freedom. One popular way to encode this latent structure is by assuming the matrix is low rank[14]. A more general framework to encode latent structure is given in[38] and contains low-rank matrices as a special case in the scalar setting[[17], Sec. D]. We utilize the more general framework here.

Assumption 1 (Lipschitz latent factor model on Wasserstein space). *Let the following latent structure hold: (i) There exists latent bounded metric spaces* $\left(\mathcal{H}_{row}, d_{row}(\cdot\,,\cdot)\right)$ *and* $\left(\mathcal{H}_{col}, d_{col}(\cdot\,,\cdot)\right)$ *for rows and columns, respectively, (ii) each row* i *has a latent vector* $x_{row}^{(i)} \in \mathcal{H}_{row}$*, each column* j *has a latent vector* $x_{col}^{(j)} \in \mathcal{H}_{col}$*, and (iii) there exists a function* $f\colon \mathcal{H}_{row} \times \mathcal{H}_{col} \to W_2(\mathrm{R})$ *such that for* $i \in [N+1], j \in [M+1], \mu_{ij} = f\left(x_{row}^{(i)}, x_{col}^{(j)}\right)$*. We assume that* f *is* L*-Lipschitz with respect to its row argument: For all* $x_1, x_2 \in \mathcal{H}_{row}, y \in \mathcal{H}_{col}$*, we have*

$$W_2\Big(f(x_1, y), f(x_2, y)\Big) \leq L d_{row}(x_1, x_2).$$

This assumptions means that the true distributions vary smoothly in the Wasserstein space as we vary the respective latent row vectors. Thus, rows which are close together in the latent row space will have similar distributions within the same column. In the scalar case, this latent-factor model reduces the degrees of freedom from the full $(N+1) \times (M+1)$ to $(N+1) + (M+1)$, thus allowing us to impute missing entries in the partially observed matrix.

Since our method is a user-user nearest neighbors algorithm, we only require the latent function, f, to be Lipschitz with respect to the row argument. However, our model can be easily extended to user-item and item-item nearest neighbors by restricting f to be Lipschitz with respect to the column argument as well.

In the distributional setting, this model allows us to induce a low-rank structure on, for example, the means of each distribution: Let $\mathcal{H}_{row} = [0, 1]^d$ and $\mathcal{H}_{col} = [0, 1]^d$ with their respective metrics being the Euclidean metric. Let each distribution be from the same location-scale family (such as Gaussian or continuous uniform) with the same scale $\sigma^2 > 0$. Then, we can induce a low-rank structure on the means of the distributions if the quantile function for each distribution is given by: $F^{-1}_{\mu_{ij}} = \sigma^2 F^{-1} + \langle x_{row}^{(i)}, x_{row}^{(j)} \rangle$

where F^{-1} is the common quantile function from the chosen location-scale family (e.g. Φ^{-1} in the Gaussian case where Φ is the standard Gaussian CDF).

A simpler example where both the location and scales change between distributions is given by: Let $\mathcal{H}_{row} \subset \mathrm{R}$ determine the mean of distributions, and $\mathcal{H}_{col} \subset \mathrm{R}_{>0}$ determine the variance of distributions. Here the quantile function for a measure μ_{ij} is given by $F^{-1}_{\mu_{ij}} = F^{-1} x^{(j)}_{col} + x^{(i)}_{row}$. Both of these examples can be shown to respect the Lipschitz condition. Next, we make an assumption about the missingness structure.

Assumption 2 (MCAR). *We assume the missing-completely-at-random (MCAR) case where each matrix entry's missingness* $A_{ij} \sim Bernoulli(p)$*, is i.i.d. across matrix entries, and is independent of the latent factors of the rows and columns.*

MCAR is a standard missingness pattern where the missingness is independent of both observed and unobserved factors. In the education example, this means that the chance of having data for a specific class for both using and not using digital resources is independent of the school, the class type, and the other data in the matrix. While MCAR can be relaxed to the missing not-at-random (MNAR) case, the theoretical analysis becomes more complex and missingness structure is not the main focus of this paper. We leave it as future work to provide theoretical bounds for the MNAR setting.

3. Estimation method

In this section, we propose a generalization of scalar nearest neighbors to the distributional setting. Nearest neighbors scales well to very large datasets often encountered in recommendation systems and panel-data settings, making it very popular in practice. On top of that, nearest neighbors only requires a notion of *distance* and *average* to be implemented, which makes it a suitable choice for our setting with complex infinite-dimensional objects. Singular-value based methods do not generalize as easily to distributional matrix completion because there is no similar notion of singular values for our setting. It remains interesting future work to explore if it possible to generalize singular-value based methods to the distributional setting.

In Secs. 2.2 and 2.3, we motivated why the 2-Wasserstein distance and barycenter are natural choices for comparing and averaging one-dimensional distributions. Thus, in the generic nearest neighbors algorithm, we can simply replace the squared difference in (1) with the Wasserstein distance and the scalar average in (2) with the Wasserstein barycenter.

3.1. Nearest neighbors method

The inputs to our method are a data matrix, Y, a masking matrix A, and a distance threshold parameter, $\eta \geq 0$. For each entry (i,j), we calculate $\hat{\mu}_{ij}$ as an estimate of μ_{ij}. We propose a nearest neighbors (NN) method, denoted Dist-NN, for our setting below:

Dist-NN(Y, A, i, j, η)**:**

(1) **Find the set of nearest neighbors for row** i**.** For each other row $u \neq i$ where $A_{uj} = 1$, define the columns that are observed in both rows i and u as

$$\mathcal{C}_{iu} \triangleq \{v \in [M+1] \setminus \{j\} : A_{iv} = 1, A_{uv} = 1\}.$$

Then, find the average distance between the row i and row u as:

$$\rho_{iu} \triangleq \begin{cases} |\mathcal{C}_{iu}|^{-1} \sum_{v \in \mathcal{C}_{iu}} W_2^2(Y_{iv}, Y_{uv}) & \text{if} \sim |\mathcal{C}_{iu}| \geq 1 \\ \infty & \text{if} \sim |\mathcal{C}_{iu}| = 0. \end{cases}$$

Finally, define row i's η-nearest neighbors as

$$\mathcal{N}_{i,\eta} \triangleq \{u \in [N+1] : A_{uj} = 1, \rho_{iu} \leq \eta\}.$$

(2) **Find the Wasserstein barycenter of the nearest neighbors in column** j**.** If $\mathcal{N}_{i,\eta} > 0$, then estimate the quantile function of μ_{ij} as:

$$F^{-1}_{\hat{\mu}_{ij}} = \frac{1}{\left|\mathcal{N}_{i,\eta}\right|} \sum_{u \in \mathcal{N}_{i,\eta}} F^{-1}_{Y_{uj}}.$$

If we found no nearest neighbors, then return that we could not find any neighborhood for row i.

In step 1, we calculate pairwise distances between row i and every other row that is observed in column j to estimate row i's neighbors, which we denote $\mathcal{N}_{i,\eta}$. Once we have row i's neighbors, in step 2, we find the Wasserstein barycenter of the observed distributions in column j for the nearest neighbors. This barycenter can be calculated using its quantile function, which has a closed-form solution from (5).

4. Main results

In this section, we present our main results showing that under certain regularity conditions on the probability distributions, our nearest neighbors method produces estimates close to their respective true probability distributions with high probability.

doi.org/10.32388/DX6MDB

4.1. Error decay rate

To state our main results, we require several regularity conditions on the underlying distributions.

Assumption 3 (Regularity conditions). *We say that a measure ν with distribution function F and density f is regular if (i) F is twice-differentiable and continuous on (a, b) where $-\infty < a < b < \infty$, (ii) there exists a universal $C > 0$ such that for all $x \in (a, b), f(x) \geq C$, (iii) ν has a finite second moment, (iv) $(F^{-1})'$ is L'-Lipschitz, (v) f is non-decreasing in a right-neighborhood of a and non-increasing in a left-neighborhood of b, and (vi)*

$$\sup_{x \in (a,b)} \frac{F(x)(1-F(x))}{f^2(x)} \left| f'(x) \right| < 2.$$

All continuous uniform distributions automatically satisfy this regularity condition because their probability density functions are constant on their respective supports. Note that the 2 in (10) is merely to simplify the analysis and can be raised to some $\gamma < \infty$ at the cost of worsening some parts of our error decay rate. Also note that since we assume that our densities are uniformly lower bounded and compact, Gaussian distributions do not satisfy Assum. 3. Truncated Gaussian distributions satisfy these regularity conditions. However, in our simulations in Sec. 5, we find with Gaussian distributions empirical error rates are close to our theoretical guarantees. Note that this is a common issue with analyzing the asymptotic behavior of the Wasserstein distance. See[[39], Remark 1] for a detailed discussion of this problem.

We are now ready to provide our first main result—scaling of the error for the Dist-NN estimate (proven in App. A).

Theorem 1 (Rate of error decay for $\hat{\mu}_{ij}$). *Let Assums. 1, 2, and 3 hold. Let $|\mathcal{N}_{i,\eta}|$ be the number of neighbors for row i with distance threshold η. Let N, M, and $\{n_v\}_{v \neq j}$ be fixed. Then, conditioned on $\left|\mathcal{N}_{i,\eta}\right| \geq 1$ and $\left|\mathcal{C}_{iu}\right| \geq \frac{1}{2}Mp^2$ for all $u \in [N+1]$, we have that*

$$W_2^2(\hat{\mu}_{ij}, \mu_{ij}) = \mathcal{O}_p\left(\frac{1}{n_j \left|\mathcal{N}_{i,\eta}\right|} + \frac{\log^2 n_j}{n_j^2} + \eta + \sqrt{\frac{\log(2N)}{Mp^2}} \right) \quad \text{as} \quad n_j \to \infty.$$

Each estimated distribution, $\hat{\mu}_{ij}$, is close to its respective true distribution, μ_{ij}, asymptotically. The first two terms go to 0 as $n_j \to \infty$. The third term goes to 0 as $\eta \to 0$ and the final term goes to 0 as the number of columns, $M \to \infty$. However, there is a tradeoff between optimizing the hyperparameter η and increasing the number of neighbors. Comparing this bound to the rate of convergence of the nearest neighbors algorithm for scalar matrix completion from[17] or[18], we see one similar component: the dependence on

the number of columns is also on the order of $\mathcal{O}((Mp^2)^{-1/2})$. However, there is one major difference between noisy matrix completion and distributional matrix completion:

Remark 1 (Number of neighbors can be small as long as n_j is large). The number of neighbors and number of rows does not have to increase to infinity for our error to go to 0. This is because as n_j increases, we get a better estimate of the true distributions in the j-th column, and these true distributions can be used to exactly construct μ_{ij}. Note that this $\mathcal{O}((n_j|\mathcal{N}_{i,\eta}|)^{-1})$ relationship between the 2-Wasserstein distance and a barycenter approximation from empirical distributions exists only in the asymptotic sense. In expectation, for general one-dimensiopnal distributions, this term is replaced with $\mathcal{O}(n_j^{-1} + |\mathcal{N}_{i,\eta}|^{-1})$[29, Thm. 3.2]. However, in App. D, we show that a similar rate holds for continuous uniform distributions in expectation.

Remark 2 (Asymptotic result for Wasserstein barycenters). To prove Thm. 1, we had to prove an asymptotic error bound on the Wasserstein barycenter between empirical distributions. To the best of our knowledge, this is the first asymptotic bound for 1-dimensional Wasserstein barycenters between empirical distributions. We separate out this result into its own proposition in Prop. 1 in the appendix.

Remark 3 ($n_j \to \infty$ with $\{n_v\}_{v \neq j}$ fixed). This bound is asymptotic in n_j while keeping the number of samples in the other columns fixed. We require this because in the proof, we require the neighborhood set, $\mathcal{N}_{i,\eta}$, to vary independently from the variable we send to infinity, n_j. Note that we calculate Wasserstein distances and barycenters between distributions within the same columns. So, keeping the number of samples fixed within a column allows us to utilize (6) to calculate both Wasserstein distances and barycenters.

Remark 4 (Error does not decay to zero with matrix size). If n_j is finite, then the error will be bounded away from 0 because the Wasserstein barycenter of empirical distributions is a quantizations of the Wasserstein barycenter of the true barycenter distribution. This is empirically shown in our simulations and follows from the definition of the Wasserstein barycenter of empirical distributions shown in (4). This is captured in the $\mathcal{O}(\log^2 n_j / n_j^2)$ term, which stems from the uniform error between an empirical quantile function and its respective true quantile function.

To prove Thm. 1, we break up the bound into “variance” and “bias” terms and independently bound each term. Note that these terms are not in fact the variance and bias of our estimate. We merely label these terms as such to denote which where the error in these terms is derived from: The variance term is the sampling error in each matrix entry from observing a finite number of samples. The bias term captures

how close the nearest neighbors barycenter is to the true distribution we are trying to estimate. This bias is due to the fact that we can only provide an approximation of the distance between row latent factors from a finite number of columns.

4.2. Asymptotic distribution for quantiles

We have the following asymptotic distribution for any quantile of our estimand (proven in App. B):

Theorem 2 (Asymptotic distribution of $F^{-1}_{\hat{\mu}_{ij}}(t)$). Let Assums. 1, 2, and 3 hold. Let the sequence $\{n_{j,M}, N_M, \eta_M, \mathcal{N}_{i,\eta_M}\}_{M=1}^{\infty}$ satisfy

$$\sqrt{\left|\mathcal{N}_{i,\eta_M}\right|}\frac{\log n_{j,M}}{\sqrt{n_{j,M}}} = o_p(1), n_{j,M}\left|\mathcal{N}_{i,\eta_M}\right|\left(\eta_M + \frac{\log(2N_M)}{Mp^2}\right) = o_p(1), \text{ and } \quad \left|\mathcal{N}_{i,\eta_M}\right| \geq 1.$$

Then, we have that for almost all $t \in (0, 1)$

$$\frac{\sqrt{n_{j,M}\left|\mathcal{N}_{i,\eta_M}\right|}}{\sigma_{\mathcal{N}_{i,\eta_M}}(t)}\left(F^{-1}_{\hat{\mu}_{ij}}(t) - F^{-1}_{\mu_{ij}}(t)\right) \xrightarrow{d} \mathcal{N}(0, 1) \quad \text{as} \quad M \to \infty, \text{ where}$$

$$\sigma^2_{\mathcal{N}_{i,\eta_M}}(t) \triangleq \frac{1}{\left|\mathcal{N}_{i,\eta}\right|}\sum_{u \in \mathcal{N}_{i,\eta_M}} \frac{t - t^2}{f^2_{\mu_{uj}}(F^{-1}_{\mu_{uj}}(t))}.$$

This asymptotic Gaussian distribution allows us to provide approximate confidence bands for the quantile function of our estimate. Note that this is a pointwise, and not necessarily uniform, convergence in distribution. We provide a proof of this theorem in . However, using Bonferroni's correction[40], we can use this pointwide result to provide uniform confidence bands for the quantile function calculated for a finite number of points. We show how to do this in Sec. 5.

Remark 5 (Number of neighbors can be finite). This convergence in distribution occurs even when the number of neighbors is finite. This is possible because as $n_j \to \infty$, we get closer to observing the true distributions in column j. With the true distributions, if our "bias" goes to 0, then we can recover our the unobserved distribution exactly.

While this confidence interval requires access to the true distributions of row i's neighborhood to calculate $\sigma_{\mathcal{N}_{i,\eta}}(t)$, we show in Sec. 5 that a bootstrap estimate provides a reasonable approximation. Kernel density estimates (KDE's) can also be used to estimate $\sigma_{\mathcal{N}_{i,\eta}}(t)$. See[39, Eq. 9] for an example of using KDE's to estimate quantities like $\sigma_{\mathcal{N}_{i,\eta}}(t)$.

4.3. Consequences for location-scale distributional matrix completion

Here, we introduce the location-scale category, which demonstrates how one can interpret distributional matrix completion as a generalization of noisy scalar matrix completion, and we provide asymptotic guarantees for examples that fall under this category.

4.3.1. Location-scale family: A generalization of (scalar) noisy matrix completion

In noisy scalar matrix completion[19][17], an observed scalar matrix entry is assumed to have an underlying true value, Y_{ij}, corrupted by by i.i.d. noise, ε_{ij} usually assumed to be symmetric and Sub-Gaussian. Then, the observed value is generated by sampling a single sample from the noise distribution with its center at Y_{ij}. Thus, noisy matrix completion is simply distributional matrix completion with the number of samples per matrix entry set to 1, and each distribution being from the same family. Noisy matrix completion, similar to distributional matrix completion, encounters difficulties from two sources of information loss because the underlying matrix is both masked and corrupted with noise. In order for the noise to be i.i.d., the distribution family must be of the location-scale type: distributions that depend solely on scale and location parameters. This family includes the Gaussian, continuous uniform, discrete uniform, and Student's t distributions.

If X is distributed according to a location-scale distribution, then for $\sigma > 0, \mu \in \mathrm{R}$, $Y \triangleq \sigma X + \mu$ is in the same distribution family as X and $F_Y^{-1}(x) = \sigma F_X^{-1}(x) + \mu$. Thus, for the location-scale family, the quantile function of Y is linear in terms of the quantile function of X, facilitating analysis through the Wasserstein space. Specifically, as stated above in Secs. 2.2 and 3, the 2-Wasserstein distance is a norm between quantile functions and the Wasserstein barycenter is given by an average over quantile functions. Thus, for any location-scale family, the Wasserstein barycenter will remain in the same family of distributions.

Example 1 (Simple location-scale). Let the latent row space, $\mathcal{H}_{row}$, be a bounded subset of R and the latent column space, $\mathcal{H}_{col}$ be a bounded subset of $\mathrm{R}_{>0}$. Let $F^{-1}_{f(a,b)}(x) = bF_{\mu}^{-1}(x) + a$ for some chosen measure μ that comes from a location-scale family. Locations are drawn from $\mathcal{H}_{row}$ and scales are drawn from $\mathcal{H}_{col}$. Each row has a different location and each column has a different scale.

We can check that Ex. 1 falls under the Lipschitz DGP in Assum. 1: By simple algebra we have $W_2(f(a_1, b), f(a_2, b)) = |a_1 - a_2|$. Thus, the Lipschitz constant is 1 and $d_{row}(x, y) = |x - y|$. Applying this simple location-scale model to motivating education example from Sec. 1, we model the student test scores at each school as a location-scale type (Gaussian, uniform, etc.) where each school has a constant *mean*

(location) across classes, and each subject/digital resources use pair has a different *standard deviation* (scale).

Under this category of examples, we have the following cleaner error rates for our method.

4.3.2. Error rates for the location-scale example

In the following corollary, let $\tilde{\mathcal{O}}_p$ suppress logarithmic dependencies and constants besides (M, N, p, n_j, d). Then, we have the following bounds (proven in App. C).

Corollary 1 (Location-scale with bilinear factor model with homoscedasticity). *Let Assums. 1, 3, and 2 hold. Let d_{row} and d_{col} be the Euclidean measure and μ_{row} be the uniform measure. Let $N, M,$ and p be fixed. Let $n_v = n_{-j}$ for $v \neq j$. Both asymptotic guarantees below are conditioned on the number of neighbors being at least $\frac{1}{4}(Np)^{\frac{2}{d+2}}$ and the number of shared columns between row i and row u being at least $\frac{1}{2}Mp^2$ for $u \in [N+1]$. Then, we have*

(a) ***Homoscedastic bilinear factor model.*** *Let $\mathcal{H}_{row} = [0,1]^d$, $\mathcal{H}_{col} = [0,1]^d$ and $F^{-1}_{f(x,y)} = \sigma F^{-1}_{\mu} + \langle x, y \rangle$ for some constant scale $\sigma > 0$ and location-scale distribution μ. Then, we have*

$$W_2^2(\hat{\mu}_{ij}, \mu_{ij}) = \tilde{\mathcal{O}}_p\left(\frac{1}{n_j(Np)^{\frac{2}{d+2}}} + \frac{\log^2 n_j}{n_j^2} + \frac{1}{p\sqrt{M}} + \frac{1}{n_{-j}}\right) \quad \text{as} \quad n_j \to \infty.$$

Furthermore, the number of neighbors is at least $\frac{1}{4}(Np)^{\frac{2}{d+2}}$ with probability at least $1 - 2\exp\left(-(Np)^{\frac{2}{d+2}}/16\right)$.

(b) ***Heteroscedastic independent location and scale model.*** *Let rows determine location and columns determine scale (or vice versa), $\mathcal{H}_{row} = [0,1]$, and $\mathcal{H}_{col} = [0,1]$. Then, we have*

$$W_2^2(\hat{\mu}_{ij}, \mu_{ij}) = \tilde{\mathcal{O}}_p\left(\frac{1}{n_j(Np)^{\frac{2}{3}}} + \frac{\log^2 n_j}{n_j^2} + \frac{1}{p\sqrt{M}} + \frac{1}{n_{-j}}\right) \quad \text{as} \quad n_j \to \infty.$$

Furthermore, the number of neighbors is at least $\frac{1}{4}(Np)^{\frac{2}{3}}$ with probability at least $1 - \exp\left(-Np^2/16\right)$.

Part (a) of corollary 1 corresponds to the setting of bilinear factor models with homoscedastic noise. This setting is commonly assumed in causal panel data settings with doubly robust estimators[18][13]. Part (b) in the corollary corresponds to Ex. 1. We examine this setting in our simulations because it is easier to interpret in the distributional setting.

5. Simulation results

We simulate the location-scale setting described in Ex. 1. The Python code to run our tests is available at our repository on GitHub. No commercial libraries are needed to reproduce our results. All of our tests were run on a MacBook Pro with an M2 chip and 32 GB of RAM. Thus, our tests should be easy to replicate. In these simulations, we let the sample size per matrix entry be n and thus equal across columns. In Fig. 3, we show error decay rates with respect to the number of samples, n, rows, $N+1$, and the product of n and the number of neighbors in the Gaussian location-scale case. We test the Gaussian case to connect with previous matrix completion literature, who primarily consider Sub-Gaussian noise. We also test our method on continuous uniform distributions in Figs. 1 and 4.

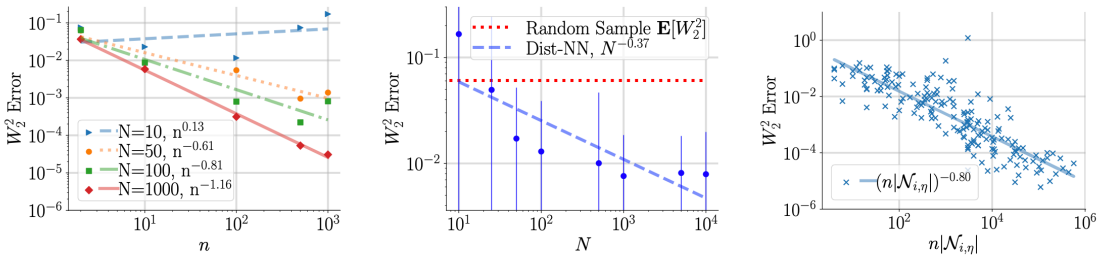
Each matrix entry is drawn from the same distribution family with only the location or scale being different between matrix entries like in Ex. 1. We draw locations from $Unif(-5, 5)$ and scales from $Unif(1, 5)$. Our experiments have one missing entry in the matrix. Nearest neighbors estimates one entry at a time. Thus, if we can estimate one matrix entry, then we can estimate all of them.

We use cross-validation on observed matrix entries to choose our threshold parameter η. Specifically, if we are trying to estimate μ_{ij}, then we loop over each observed entry in row i, hold it out, and run our method to estimate that left our entry. We then compare our estimate with the observed entry. We choose the η that minimizes the squared Wasserstein distance between our estimate and the observed entry. Since this problem is nonlinear and nonconvex, we use the Tree of Parzen Estimators (TPE) method[41] to choose η. This method is a Bayesian optimization method that uses a Gaussian process to model the objective function. We use the Python library hyperopt to run TPE and used the standard settings with a maximum of 50 iterations.

Error with respect to number of samples, n. As shown in Fig. 3, as the number of samples, n, and the number of rows, $N+1$, go up, our estimation error drops rapidly. For the error plot against the number of samples, we can see that our error decay rate improves from about $\mathcal{O}(n^{-1})$ to $\mathcal{O}(n^{-1.16})$ as the number of neighbors increases. This is supported by our theoretical result in Thm. 1 where as the number of neighbors increases, the dominant rate with respect to n becomes $\mathcal{O}(\log^2 n/n^2)$. We also see that the error rate power with respect to $n|\mathcal{N}_{i,\eta}|$ is around $\mathcal{O}(\left(n\left|\mathcal{N}_{i,\eta}\right|\right)^{-0.8})$, which is close to the asymptotic bound of $\mathcal{O}((n\left|\mathcal{N}_{i,\eta}\right|)^{-1})$.

Error with respect to number of rows, N. For the error plot against the number of rows, we see that the error also drops rapidly with the number of rows. Again, we manage to achieve a better error decay rate

than is predicted by our theoretical results. We also plot the expected error of an observed random sample in the dotted line to show that our method is able to produce an estimate that is far better than an observed random sample. Even for just 20 rows, our error is already significantly better than the expected error of a random sample. Thus, our synthetic sample is a much better estimate of the true distribution than an entry's random samples alone. We call this ability "denoising" because it mirrors the denoising ability of scalar nearest neighbors.



(a) Scaling with # samples n (b) Scaling with # rows N (c) Scaling with $n|\mathcal{N}_{i,\eta}|$

Figure 3. Scaling of error with sample size n, number of rows N, and effective sample size $n\left|\mathcal{N}_{i,\eta}\right|$. Every distribution in the matrix is a Gaussian distribution. Each row has an expected value sampled from $Unif(-5, 5)$. Each column has a standard deviation sampled from $Unif(1, 5)$. In plot (a), we set the number of columns to 30. We also require at least 2 nearest neighbors. In plot (b), we simulated a random sample 100 times to estimate the expected error of a random sample. In plot (c), we set the number of samples to 500 and the number of columns to 10. We also cut the plot off on the top at 0.4 so that the lower error samples can be better visualized. We simulated each setting 50 times. Each curve is fitted using least squares to the power function $f(x) = ax^b$.

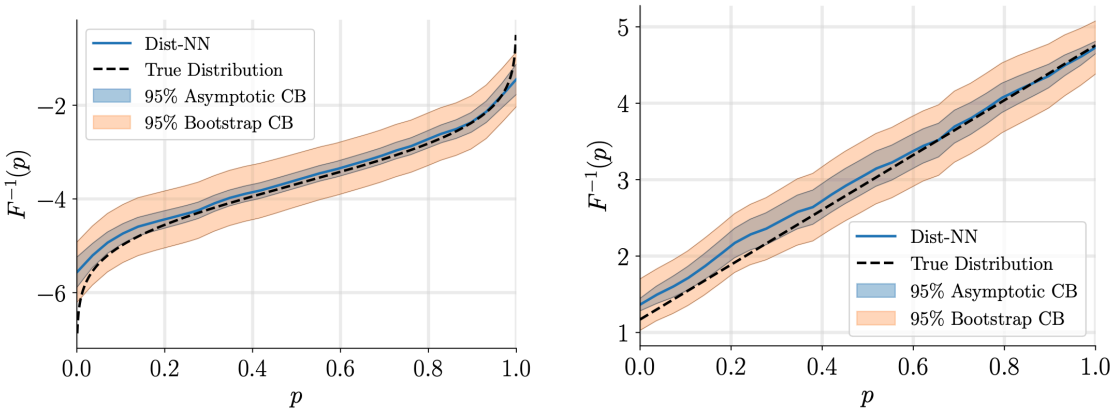
Confidence bands. Using the results of Thm. 2, we can provide confidence bands for the quantiles of our estimates. In this simulated setting, we have access to the true distributions of the matrix entries we take a barycenter over. In practice, the $\sigma_{i,\mathcal{N}_{i,\eta}}$ quantity would need to be estimated. However, we show that a bootstrap estimate of the confidence bands using the empirical samples alone provides a good estimate of the true confidence bands.

In Fig. 4, we plot both the true and bootstrap confidence bands in the Gaussian and uniform location-scale cases. For our bootstrap method, we resample from both the individual neighboring distributions and resample over the neighbor set itself. We resample over samples and neighbors 10 times each for these simulations. Clearly, the bootstrap confidence bands are more conservative than the asymptotic

doi.org/10.32388/DX6MDB

confidence bands. Note that we provide estimates for the Gaussian case even though the Gaussian distribution does not satisfy our assumption that the true distribution has a continuous quantile function on [0,1] because it is undefined at the boundary. This is why our estimate is poor around the boundary. However, the continuous uniform distribution satisfies our assumption. So, our estimate is much better around the boundary.

Note that we provide estimates of simultaneous confidence bands as opposed to pointwise confidence bands for the quantile function. I.e., our confidence regions provide 95% coverage at each of the n points simultaneously. We provide simultaneous confidence bands using Bonferroni's correction[40] by dividing the confidence level, $\alpha = 0.05$, by the number of confidence intervals we plot, n.



(a) Gaussian (b) Continuous uniform

Figure 4. Asymptotic and bootstrap simultaneous confidence bands for Gaussian and continuous uniform location-scale case. The bootstrap confidence bands are more conservative than the bands provided by our asymptotic result. However, the bootstrap estimate resamples the neighboring distributions as well whereas the asymptotic one does not, which could make the bootstrap confidence bands more accurate. Also note that for the Gaussian case, our estimate is worse around $p = 0$ and $p = 1$. This is expected, because our theoretical guarantees rely on the true distribution being supported on a compact interval.

Denoising. We show the method's denoising ability through empirical CDF's in Fig. 1. This ability of nearest neighbors means that information can be shared across rows to achieve empirical distributions that are much closer to their respective true distributions than a random sample. This feature of our method is beneficial for downstream analysis since distributional quantities such as mean, variance, and value-at-risk can be estimated with a much higher accuracy than an isolated observed set of samples.

Value-at-risk (VaR) is commonly used in financial modeling and is defined for $\alpha \in (0, 1)$ as $VaR_X(\alpha) = F^{-1}_{-X}(1 - \alpha)$ where X is a random variable and F^{-1}_{-X} is the quantile function of $-X$.

We show empirically that our method estimates distributional quantities well in Fig. 5 where, clearly, the synthetic sample produces estimates of the mean, standard deviation, value-at-risk, and median a lot closer to their respective true values than what an observed random sample alone provides. This shows that running our method on observed distributions can provide much better estimates of their true distributions than just using their random samples in isolation. We also tested our method against scalar nearest neighbors, where the respective distributional quantity is calculated for each observed distribution beforehand to create the scalar matrix, and then we run scalar nearest neighbors. For estimating these distributional quantities, our distributional nearest neighbors does just as well or better than scalar nearest neighbors. However, for scalar nearest neighbors, each distributional quantity must be calculated ahead of time before the estimation procedure, whereas new distributional quantities can be estimated from our method without rerunning the entire procedure.

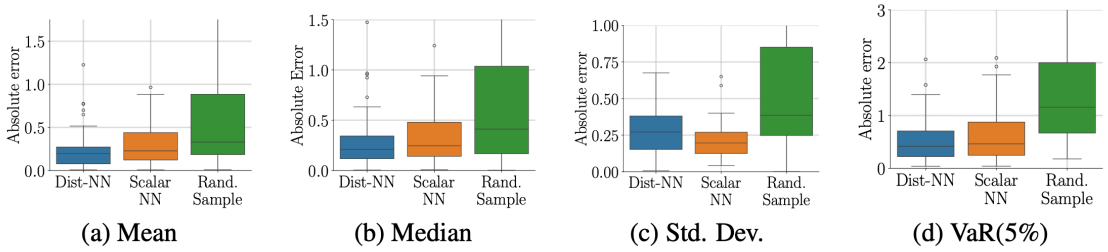


Figure 5. Error for distributional nearest neighbors (Dist-NN), scalar nearest neighbors, and a random sample for estimating means, standard deviations, value-at-risk numbers (VaR (5%)), and medians. We use the same Gaussian location-scale setup Fig. 3. Dist-NN is able to estimate all of these distributional quantities far better than a random sample and with better or about the same accuracy as scalar nearest neighbors. We cut off the y-axis to visualize the difference between Dist-NN and scalar nearest neighbors since both perform better than an observed random sample.

6. Discussion

In this paper, we proposed a new problem, distributional matrix completion, where matrix entries are one-dimensional empirical distributions. We also proposed a nearest neighbors method to solve this new problem using tools from optimal transport and proved theoretical asymptotic bounds and distributions

doi.org/10.32388/DX6MDB

for our estimate's error. Additionally, our simulations showcased the ability of our method to not only recover the unobserved distributions but also create synthetic distributions that are consistently closer to their true distributions than an observed random sample alone.

We also showed empirically that our method produces estimates for quantities such as means, standard deviations, and quantiles that are just as good or better than scalar nearest neighbors. Whereas scalar nearest neighbors requires recreating the entire matrix with the desired distributional quantities, our method allows calculation of new quantities without rerunning the whole estimation procedure. A key takeaway from these findings is that even for observed distributions, utilizing information from neighbors can provide far better estimates of distributional quantities of interest.

While our theoretical results are limited by the assumptions we require, our method is not limited by these assumptions and would, for instance, run in settings without i.i.d. sampling or with different numbers of samples in each matrix entry. In future work, we will seek to loosen these assumptions as much as possible. We also plan on extending our method to support higher dimensions. The main challenge with this is that the Wasserstein barycenter suffers from the curse of dimensionality[30]. This slowdown is evident even on small 2D distributions such as grayscale images. However, we believe there are interesting dimension-free results if one restricts the class of probability distributions (see[31] for one such result). Finally, we plan on extending our theoretical analysis to the missing-not-at-random (MNAR) setting where the missingness is not independent of the observed data. This will allow us to extend our method and theoretical results to settings in causal inference to generate counterfactual distributions.

Appendix

In this appendix, we provide the proofs of the main theorems and corollary.

Appendix A. Proof of Thm. 1: Asymptotic bound

First, we split up the error into bias and variance components by definition a new term: Let

$$\bar{\mu}_{ij} \triangleq \arg\min_{\mu} \sum_{u \in \mathcal{N}_{i,\eta}} W_2^2(\mu, \mu_{uj}). \tag{13}$$

From[29, Eq. 8], this minimum has a closed form solution in 1-dimension where $\bar{\mu}_{ij}$ is the measure with quantile function given by

	$\hat{F}^{-1}_{\mu_{ij}} = \frac{1}{\lvert\mathcal{N}_{i,\eta}\rvert} \sum_{u \in \mathcal{N}_{i,\eta}} F^{-1}_{\mu_{uj}}.$		

Then, we have

	$W_2^2\left(\hat{\mu}_{ij}, \mu_{ij}\right) \overset{(3)}{=} \lVert \hat{F}^{-1}_{\mu_{ij}} - F^{-1}_{\mu_{ij}} \rVert^2_{L^2(0,1)}$	$= \lVert \hat{F}^{-1}_{\mu_{ij}} - F^{-1}_{\mu_{ij}} + F^{-1}_{\mu_{ij}} - F^{-1}_{\mu_{ij}} \rVert^2_{L^2(0,1)}$		(14)
		$\overset{(a)}{\leq} \left(\lVert \hat{F}^{-1}_{\mu_{ij}} - F^{-1}_{\mu_{ij}} \rVert_{L^2(0,1)} + \lVert F^{-1}_{\mu_{ij}} - F^{-1}_{\mu_{ij}} \rVert_{L^2(0,1)} \right)^2$		
		$\overset{(b)}{\leq} 2 \left[\underbrace{W_2^2\left(\hat{\mu}_{ij}, \bar{\mu}_{ij}\right)}_{\triangleq \mathcal{V}} + \underbrace{W_2^2\left(\bar{\mu}_{ij}, \mu_{ij}\right)}_{\triangleq \mathcal{B}} \right],$		

where (a) follows from the Minkowski inequality[42, Thm. 198] and (b) follows from the Cauchy-Schwarz inequality[42, Thm. 7].

Next, we bound the terms $\mathcal{B}$ and $\mathcal{V}$. Notice that the term $\mathcal{V}$ is akin to a variance term measured as the Wasserstein distance between the barycenter $\hat{\mu}_{ij}$ obtained from the empirical distributions of the neighbors versus the barycenter $\bar{\mu}_{ij}$ constructed from their true distributions. On the other hand, $\mathcal{B}$ denotes a bias-like term (that would arise if we had infinite sample size in each cell), which measures the squared Wasserstein distance between the latter barycenter and the distribution μ_{ij} being estimated. To proceed further, we utilize two auxiliary results.

The next proposition, proven in Sec. A.1, characterizes the convergence of "empirical" barycenter to the "population" barycenter as a function of both the sample size and the number of distributions:

Proposition 1 (Convergence of the barycenter of empirical measures). *Consider a collection of measures* $\{v_j\}_{j=1}^k$ *each of which satisfies Assum. 3. For each* $j \in [k]$*, let* $\hat{v}_{j,n}$ *denote the empirical distribution obtained from* n *i.i.d. samples from* v_j*. Define the two barycenters*

	$\bar{v} \triangleq \arg\min_v \sum_{j=1}^k W_2^2(v, v_j) \quad \text{and} \quad \bar{v}_n \triangleq \arg\min_v \sum_{j=1}^k W_2^2(v, \hat{v}_{j,n}).$		

Then we have

$$W_2^2(\bar{\nu}_n, \bar{\nu}) = \mathcal{O}_p\left(\frac{1}{nk} + \frac{\log^2 n}{n^2}\right) \quad \text{as} \quad n \to \infty$$

uniformly with constants in $\mathcal{O}_p(\cdot)$ *that do not depend on* $\{\nu_j\}_{j=1}^k$ *or the number of distributions,* k.

Prop. 1 shows that the empirical barycenter $\bar{\nu}_n$ converges to the population barycenter $\bar{\nu}$ at the order of $\mathcal{O}(1/nk + \log^2 n/n^2)$. Next, we provide a lemma showing that we can apply Prop. 1.

Lemma 1. *Let* (i,j) *be the entry we seek to estimate. Let* $\bar{\mu}_{ij}$ *be defined as in (13),* $\hat{\mu}_{ij}$ *be the distribution with quantile function defined in (9), and* $\mathcal{N}_{i,\eta}$ *defined as in (8). Then, we have have that the measures in column* j *satisfy the conditions of Prop. 1, and we thus have*

$$W_2^2\left(\hat{\mu}_{ij}, \bar{\mu}_{ij}\right) = \mathcal{O}_p\left(\frac{1}{n_j|\mathcal{N}_{i,\eta}|} + \frac{\log^2 n_j}{n_j^2}\right) \quad \text{as} \quad n_j \to \infty, \tag{15}$$

Proof. Note that in this proof, we abuse notation slightly by letting Y_{uv} be the samples in matrix entry (u,v) regardless if (u,v) is observed or not. For $u \in [N+1]$, let I_u be the indicator random variable that $\rho_{iu} \le \eta$. Recall the definition of the row-wise distance from 12:

$$\rho_{iu} \triangleq \begin{cases} |\mathcal{C}_{iu}|^{-1} \sum_{v \in \mathcal{C}_{iu}} W_2^2(Y_{iv}, Y_{uv}) & \text{if} \sim |\mathcal{C}_{iu}| \ge 1 \\ \infty & \text{if} \sim |\mathcal{C}_{iu}| = 0, \end{cases} \quad \text{where}$$

$$\mathcal{C}_{iu} \triangleq \{v \in [M+1] \setminus \{j\} : A_{iv} = 1, A_{uv} = 1\}.$$

Then, from MCAR, A is independent from the samples in each entry and from the latent factors. In particular, $Y_{uj} \perp\!\!\!\perp A_{uj}$ for $u \in N+1$. Next, condition on the latent row and column factors $\mathcal{U}_{row}$ and $\mathcal{U}_{col}$. Thus, each distribution is now fixed. Then, $Y_{uj} \perp\!\!\!\perp I_u$ because the samples in column j are not used to calculate ρ_{iu}. A_{uj} is dependent on ρ_{iu} because $\mathcal{C}_{iu}$ is only defined if $A_{uj} = 1$. But, we already showed by MCAR

that $Y_{uj} \perp\!\!\!\perp A_{uj}$. Next, in column j, we assume that each observed matrix entry has n_j samples that are drawn i.i.d. from Sec. 2. Thus, we can apply Prop. 1 to the empirical distribution set $\{Y_{uj}\}_{u \in \mathcal{N}_{i,\eta}}$. Finally, since the bound in Prop. 1 is uniform with constants that do not depend on the distributions or $\left|\mathcal{N}_{i,\eta}\right|$, then we can remove the conditioning on $\mathcal{U}_{row}$ and $\mathcal{U}_{col}$. □

Next, we claim an asymptotic bound on the bias term (proven in Sec. A.3):

$$W_2^2\left(\bar{\mu}_{ij}, \mu_{ij}\right) = \mathcal{O}_p\left(\eta + \sqrt{\frac{\log(2N)}{Mp^2}}\right). \tag{16}$$

Putting together (15) and (16), we get

$$\begin{aligned} W_2^2(\hat{\mu}_{ij}, \mu_{ij}) &\overset{(14)}{\leq} 2\left(W_2^2\left(\hat{\mu}_{ij}, \bar{\mu}_{ij}\right) + W_2^2\left(\bar{\mu}_{ij}, \mu_{ij}\right)\right) \\ &= \mathcal{O}_p\left(\frac{1}{n_j\left|\mathcal{N}_{i,\eta}\right|} + \frac{\log^2 n_j}{n_j^2} + \eta + \sqrt{\frac{\log(2N)}{Mp^2}}\right) \quad \text{as} \quad n_j \to \infty. \end{aligned}$$

A.1. Proof of proposition 1: Convergence of the barycenter of empirical measures

From[[29], Eq. 8], the quantile functions of each barycenter has an explicit formula:

$$\bar{F}_v^{-1} = \frac{1}{k}\sum_{j=1}^{k} F_{v_j}^{-1} \quad \text{and} \quad \bar{F}_{\hat{v}}^{-1} = \frac{1}{k}\sum_{j=1}^{k} F_{\hat{v}_{j,n}}^{-1}.$$

So, we can write the Wasserstein distance between the two barycenters as:

$$\begin{aligned} W_2^2(\bar{v}, \bar{v}) &\overset{(3)}{=} \|\bar{F}_{\hat{v}}^{-1} - \bar{F}_v^{-1}\|^2 \\ &\overset{(5)}{=} \left\|\frac{1}{k}\sum_{j=1}^{k}\left(F_{\hat{v}_{j,n}}^{-1} - F_{v_j}^{-1}\right)\right\|_{L^2(0,1)}^2 \\ &= \frac{1}{n}\left\|\frac{1}{k}\sum_{j=1}^{k}\sqrt{n}\left(F_{\hat{v}_{j,n}}^{-1} - F_{v_j}^{-1}\right)\right\|_{L^2(0,1)}^2 \\ &\triangleq \frac{1}{n}\left\|\frac{1}{k}\sum_{j=1}^{k}\sqrt{n}\hat{q}_{v_j,n}\right\|_{L^2(0,1)}^2, \end{aligned} \tag{17}$$

where we have defined $\hat{q}_{v_j,n} \triangleq \sqrt{n}(F^{-1}_{\hat{v}_{j,n}} - F^{-1}_{v_j})$ in the last step. To complete the proof, we next derive the asymptotic distribution of $\hat{q}_{v_j,n}$. While it is well known that $\hat{q}_{v_j,n}$ converges to a weighted Brownian bridge in distribution[[43], Ch. 18],[1] the next lemma establishes a stronger result, namely a convergence in probability for the barycenters. See Sec. A.1.1 for the proof.

Lemma 2 (Approximation of barycenter). *Consider a collection of measures $\{v_j\}_{j=1}^k$ each of which satisfies Assum. 3. For each $j \in [k]$, let $\hat{v}_{j,n}$ denote the empirical distribution obtained from n i.i.d. samples from v_j. Then there exists sequences of standard Brownian bridges $(\mathrm{B}_{j,n})_{j=1}^k$ such that as $n \to \infty$, we have (uniformly)*

$$\frac{2}{n}\left\|\frac{1}{k}\sum_{j=1}^{k}\left[\sqrt{n}(F^{-1}_{\hat{v}_{j,n}} - F^{-1}_{v_j}) - \frac{\mathrm{B}_{j,n}}{f_{v_j}\circ F^{-1}_{v_j}}\right]\right\|^2_{L^2(0,1)} \overset{a.s.}{=} \mathcal{O}\left(\frac{\log^2 n}{n^2}\right).$$

Using the Brownian bridges appearing in Lem. 2, we obtain

$$\begin{aligned}
\bar{\bar{W}}_2^2(\hat{\boldsymbol{v}}, v) &\overset{(17)}{=} \frac{1}{n}\left\|\frac{1}{k}\sum_{j=1}^{k}\hat{q}_{v_j,n}\right\|^2_{L^2(0,1)} \\
&= \frac{1}{n}\left\|\frac{1}{k}\sum_{j=1}^{k}\left(\hat{q}_{v_j,n} - \frac{\mathrm{B}_{j,n}}{f_{v_j}\circ F^{-1}_{v_j}} + \frac{\mathrm{B}_{j,n}}{f_{v_j}\circ F^{-1}_{v_j}}\right)\right\|^2_{L^2(0,1)} \\
&\overset{(i)}{\le} \frac{2}{n}\left\|\frac{1}{k}\sum_{j=1}^{k}\left(\hat{q}_{v_j,n} - \frac{\mathrm{B}_{j,n}}{f_{v_j}\circ F^{-1}_{v_j}}\right)\right\|^2_{L^2(0,1)} + \frac{2}{n}\left\|\frac{1}{k}\sum_{j=1}^{k}\frac{\mathrm{B}_{j,n}}{f_{v_j}\circ F^{-1}_{v_j}}\right\|^2_{L^2(0,1)}.
\end{aligned}$$

where (i) follows from applying Minkowski's inequality[[42], Thm. 198] followed by the Cauchy-Schwarz inequality[[42], Thm. 7]. While Lem. 2 establishes that the first term on the RHS of the above display is bounded almost surely, the next result (proven in Sec. A.1.2) provides a tight control on the second term in probability:

Lemma 3 (Norm of weighted average of Brownian bridges). *Let $(\mathrm{B}_j)_{j=1}^k$ be sequences of standard Brownian bridges. Let $(w_j(\cdot))_{j=1}^k$ be positive L-Lipschitz-continuous functions. Then, we have*

$$\frac{1}{n}\left\|\frac{1}{k}\sum_{j=1}^{k} w_j \cdot \mathrm{B}_j\right\|^2_{L^2(0,1)} = \mathcal{O}_p\left(\frac{1}{nk}\right).$$

Putting together Lems. 2 and 3, we get our final result:

$$W_2^2(\bar{\nu}, \bar{\nu}) = \mathcal{O}_p\left(\frac{\log^2 n}{n^2} + \frac{1}{nk}\right).$$

A.1.1. Proof of Lem. 2: Approximation of barycenter

First, we will apply the following theorem (Note that we only restate the part of theorem on Brownian bridges and not Kiefer processes):

Theorem A (Thm. 6 in[44]). *Consider a measure ν satisfying Assum. 3 and let $\hat{\nu}_n$ denote the empirical distribution obtained from n i.i.d. samples from ν. Then there exists a sequence of Brownian bridges $\mathrm{B}_1, \ldots, \mathrm{B}_n$ such that*

$$\sup_{0<t<1}\left|f(F^{-1}(t))\sqrt{n}(F^{-1}_{\hat{\nu}_n}(t) - F^{-1}_{\nu}(t)) - \mathrm{B}_n(t)\right|$$

$$\overset{a.s.}{=} \begin{cases} \mathcal{O}\left(n^{-1/2}\log n\right) & \text{if} \quad \gamma < 2 \\ \mathcal{O}\left(n^{-1/2}(\log\log n)^{\gamma}(\log n)^{(1+\epsilon)(\gamma-1)}\right) & \text{if} \quad \gamma \geq 2, \end{cases}$$

where $\epsilon > 0$ is arbitrary.

From Assum. 3, we have $\gamma < 2$ for each distribution. So, we have that for each distribution ν_j and its respective approximation by a sequence of Brownian bridges, $\left(\mathrm{B}_{j,l}\right)_{l=1}^n$,

$$\sup_{t\in(0,1)}\left|f_{\nu_j}\left(F^{-1}_{\nu_j}(t)\right)\hat{q}_{\nu_j,n}(t) - \mathrm{B}_{j,n}(t)\right| \overset{a.s.}{=} \mathcal{O}\left(\frac{\log n}{\sqrt{n}}\right).$$

This result holds for each measure, but we need it to hold uniformly for any finite set of measures. Thus, we will unpack the proof of this theorem to show that it holds uniformly. Thm. A is proven by combining three theorems where u_n be the quantile process $\hat{q}_n$ for a $Unif(0, 1)$ random variable:

Theorem B (Thm. 1 in[44]). *If the uniform (0,1) random variables $U_1, U_2, \ldots$ are defined on a rich enough probability space, then one can define, for each n, a Brownian bridge $\{\mathrm{B}_n(y) : 0 \leq y \leq 1\}$ on the same probability*

space such that, for all z, we have

$$\mathrm{P}\left(\sup_{0\le y\le 1}\left|u_n(y)-\mathrm{B}_n(y)\right| > n^{-1/2}(A\log n+z)\right)\le Be^{-Cz}$$

for positive absolute constants $A, B,$ and C;

Theorem C (Thm. 2 in[44]). *With $\delta_n = 25n^{-1}\log\log n$ we have*

$$\limsup_{n\to\infty}\sup_{\delta_n\le y\le 1-\delta_n}(y(1-y)\log\log n)^{-1/2}\left|u_n(y)\right| \overset{a.s.}{\le} 4; \text{ and} \tag{19}$$

Theorem D(Thm. 3 in[44]). *Let $X_1, X_2, \ldots$ be i.i.d random variables with a continuous distribution function F which is also twice differentiable on (a, b) and $F' = f \neq 0$ on (a, b). Let the quantile processes $\hat{q}_n(y)$ and respective $u_n(y)$ be defined in terms of the order statistics $X_{k:n}$ and $U_{k:n} = F(X_{k:n})$. Assume that for some $\gamma > 0$,*

$$\sup_{a<x<b}F(x)(1-F(x))\left|\frac{f'(x)}{f^2(x)}\right|\le\gamma,$$

and f is nondecreasing (increasing) on an interval to the right of a (to the left of b). Then, with δ_n as in Thm. C

$$\sup_{0<y<1}\left|f(F^{-1}(y))q_n(y)-u_n(y)\right| \overset{a.s.}{=} \begin{cases}\mathcal{O}(n^{-1/2}\log\log n) & \text{if } \gamma<1\\ \mathcal{O}(n^{-1/2}(\log\log n)^2) & \text{if } \gamma=1\\ \mathcal{O}(n^{-1/2}(\log\log n)^{\gamma}(\log n)^{(1+\varepsilon)(\gamma-1)}) & \text{if } \gamma>1\end{cases}$$

where $\varepsilon > 0$ is arbitrary. The constants in the $\mathcal{O}(\cdot)$ are respectively, $2(\max(45, 25(2^{\gamma}/(1-\gamma)))) + 40\gamma 10^{\gamma}$ *if* $\gamma < 1$, *102 if* $\gamma = 1$, *and* $2\max(45, (2^{\gamma}/(\gamma-1))25^{\gamma})$ *if* $\gamma > 1$.

First, Thm. B can be easily extended to our setting using a simple union-bound:

$$\mathrm{P}\left(\sup_{i\in[k]}\sup_{0\le y\le 1}\left|u_{n,i}(y)-\mathrm{B}_{n,i}(y)\right| > n^{-1/2}(A\log n+z)\right)\le kBe^{-Cz}$$

So, as long as k does not grow faster than $\exp(z)$, this statement will still hold and we thus have

$$\sup_{i\in[k]}\sup_{0\le y\le 1}\left|u_{n,i}(y)-\mathrm{B}_{n,i}(y)\right| \overset{a.s.}{=} \mathcal{O}(n^{-1/2}\log n).$$

Next, for Thm. C, we have

$$\limsup_{n\to\infty}\sup_{i\in[k]}\sup_{\delta_n\le y\le 1-\delta_n}(y(1-y)\log\log n)^{-1/2}\left|u_{n,i}(y)\right| \tag{20}$$

$$= \sup_{i\in[k]}\limsup_{n\to\infty}\sup_{\delta_n\le y\le 1-\delta_n}(y(1-y)\log\log n)^{-1/2}\left|u_{n,i}(y)\right|$$

$$\overset{(19)}{\le} \sup_{i\in[k]} 4$$

$$= 4.$$

For Thm. D, we claim that we have the same asymptotic behavior in our case (proven at the end of this section)

$$\sup_{i\in[k]}\sup_{0<y<1}\left|f(F^{-1}(y))q_n(y)-u_n(y)\right| \tag{21}$$

$$\overset{a.s.}{=} \begin{cases} \mathcal{O}(n^{-1/2}\log\log n) & \text{if } \gamma<1 \\ \mathcal{O}(n^{-1/2}(\log\log n)^2) & \text{if } \gamma=1 \\ \mathcal{O}(n^{-1/2}(\log\log n)^{\gamma}(\log n)^{(1+\varepsilon)(\gamma-1)}) & \text{if } \gamma>1 \end{cases}$$

Putting these pieces together, we get

$$\sup_{i\in[k]}\sup_{0<y<1}\left|f_i(F_i^{-1}(y))\hat{q}_{n,i}(y)-\mathrm{B}_{n,i}(y)\right|$$

$$\overset{(b)}{\le} \sup_{i\in[k]}\sup_{0<y<1}\left|f_i(F_i^{-1}(y))\hat{q}_{n,i}(y)-u_{n,i}(y)\right| + \left|u_{n,i}(y)-\mathrm{B}_{n,i}(y)\right|$$

$$\overset{(c)}{=} \mathcal{O}(n^{-1/2}\log n + n^{-1/2}(\log\log n)^2)$$

$$= \mathcal{O}(n^{-1/2}\log n)$$

where (b) follows from the triangle inequality and (c) follows because $\gamma < 2$ for each measure and we can set $\varepsilon < 1/(\gamma - 1) - 1$ if $1 < \gamma < 2$ in (21).

We can now finish the main proof. From Assum. 3, there exists a positive constant C that lower bounds each density function. Thus, for n large enough (which we have shown exists for any admissible value of k), we have for some universal constant c

	$\sup_{j\in[k]}\sup_{0<y<1}\left\lvert f_j(F_j^{-1}(y))\hat{q}_{n,j}(y) - \mathrm{B}_{n,j}(y)\right\rvert$	$\overset{a.s}{\leq} c(n^{-1/2}\log n + kn^{-1/2}(\log\log n)^2)$		(22)
	$\sup_{j\in[k]}\sup_{0<y<1}\left\lvert\hat{q}_{n,j}(y) - \frac{\mathrm{B}_{n,j}(y)}{f_j(F_j^{-1}(y))}\right\rvert$	$\overset{a.s.}{\leq} \frac{c}{C}(n^{-1/2}\log n + kn^{-1/2}(\log\log n)^2)$		

Then, for any $y \in (0, 1)$, almost surely,

	$\left\lvert\frac{1}{k}\sum_{j=1}^{k}\left(\hat{q}_{v_j,n}(y) - \frac{\mathrm{B}_{n,j}(y)}{f_j(F_j^{-1}(y))}\right)\right\rvert$	$\overset{(c)}{\leq} \frac{1}{k}\sum_{j=1}^{k}\left\lvert\hat{q}_{v_j,n}(y) - \frac{\mathrm{B}_{n,j}(y)}{f_j(F_j^{-1}(y))}\right\rvert$	
		$\overset{(22)}{\leq} \frac{c}{C}n^{-1/2}\log n$	

where (c) follows the triangle inequality. Next, since this holds for all y, we can take the $L^2(0, 1)$ of both sides to get

	$\frac{2}{n}\left\lVert\frac{1}{k}\sum_{j=1}^{k}\left(\hat{q}_{v_j,n}(y) - \frac{\mathrm{B}_{n,j}(y)}{f_j(F_j^{-1}(y))}\right)\right\rVert^2_{L^2(0,1)} \leq \frac{2c}{n^2C}\log^2 n = \mathcal{O}\left(\frac{\log^2 n}{n^2}\right).$	

Proof of claim (21). We will only repeat the parts of the proof that differ in our case. Let $y \in ((l-1)/n, l/n]$ and ξ be between y and $U_{l:n} = y + \sqrt{n}u_n(y)$. Then, we have from[[44], Eq. 3]

	$\sup_{i\in[k]}\left\lvert f_i(F_i^{-1}(y))\hat{q}_{n,i}(y) - u_{n,i}(y)\right\rvert \leq \sup_{i\in[k]}\frac{1}{2}n^{-1/2}u^2_{n,i}(y)f_i(F_i^{-1}(y))\frac{\left\lvert f_i'(F_i^{-1}(\xi))\right\rvert}{f_i^3(F_i^{-1}(\xi))}.$	

Next, from (20) and since each term is nonnegative, we have for large enough n

$$\sup_{i\in[k]}\left|f_i(F_i^{-1}(y))\hat{q}_{n,i}(y)-u_{n,i}(y)\right|\le 8n^{-1/2}(\log\log n)y(1-y)\sup_{i\in[k]}f_i(F_i^{-1}(y))\frac{\left|f_i'(F_i^{-1}(\xi))\right|}{f_i^3(F_i^{-1}(\xi))}.$$

From the proof of[[44], Thm. 3], we have

$$\sup_{i\in[k]}\left|f_i(F_i^{-1}(y))\hat{q}_{n,i}(y)-u_{n,i}(y)\right|\le 8\gamma 5\cdot 10^{\gamma}n^{-1/2}(\log\log n).$$

Next, from[[44], Eq. 3.10], we have

$$\sup_{i\in[k]}\sup_{0\le y\le\delta_n}\left|u_n(y)\right|\overset{a.s.}{\le}45n^{-1/2}\log\log n.$$

From[[44], Eq. 3.13], if $U_{l:n}\ge y$, then

$$\sup_{i\in[k]}\left|f_i(F_i^{-1}(y))q_{n,i}(y)\right|\le u_n(y).$$

If $U_{k:n}<y$, then[[44], Eq. 3.14] establishes

$$\sup_{i\in[k]}\left|f_i(F_i^{-1}(y))q_{n,i}(y)\right|\le\begin{cases}\frac{2^{\gamma}}{1-\gamma}n^{1/2}y & \text{if}\quad \gamma<1\\ \frac{2^{\gamma}}{\gamma-1}n^{1/2}y^{\gamma}U_{l:n}^{-(\gamma-1)} & \text{if}\quad \gamma>1\\ 2n^{1/2}y\log(y/U_{l:n}) & \text{if}\quad \gamma=1\end{cases}$$

Next, from the end of proof of[[44], Thm. 3], we have our result.

A.1.2. Proof of Lem. 3: Norm of average of Brownian bridges

We seek to prove that Prop. 1 holds with constants that do not depend on k. This will allow us to remove the conditioning on the entire neighborhood set. We have

$$\frac{1}{n}\Big\|\frac{1}{k}\sum_{j=1}^{k} w_j \cdot \mathrm{B}_j\Big\|^2_{L^2(0,1)} = \frac{1}{nk}\Big\|\frac{1}{\sqrt{k}}\sum_{j=1}^{k} w_j \cdot \mathrm{B}_j\Big\|^2_{L^2(0,1)} \quad .$$

$$\widetilde{\mathrm{G}_k}$$

Clearly, G_k is a Gaussian process with $\mathrm{E}[\mathrm{G}_k(t)] = 0$ and $\mathrm{G}_k(0) = \mathrm{G}_k(1) = 0$. We also know that it has continuous sample paths and is thus bounded. However, we wish to show that it is uniformly bounded over all values of k. For $s, t \in [0, 1]$, we have

$$\begin{aligned} Cov(\mathrm{G}_k(s), \mathrm{G}_k(t)) &= \frac{1}{k} Cov\left(\sum_{j=1}^{k} w_j(s)\mathrm{B}_j(s), \sum_{j=1}^{k} w_j(t)\mathrm{B}_j(t)\right) \\ &\overset{(a)}{=} \frac{1}{k}\sum_{j=1}^{k} Cov(w_j(s)\mathrm{B}_j(s), w_j(t)\mathrm{B}_j(t)) \\ &= \frac{1}{k}\sum_{j=1}^{k} w_j(s)w_j(t)(\min(s,t) - st) \end{aligned}$$

where (a) follows from the independence on the Brownian bridges. Next, we have for all $j \in [k]$ and $t \in [0, 1]$, $w_j(t) < C'$ for some universal C'. Thus,

$$Var(\mathrm{G}_k(t)) = Cov(\mathrm{G}_k(t), \mathrm{G}_k(t)) = \frac{1}{k}\sum_{j=1}^{k} (w_j(t))^2(t - t^2) \le (t - t^2)C'^2 \le (0.5)C'^2. \tag{23}$$

Let $\sigma^2_{\mathrm{G}_k} \triangleq \sup_{t\in[0,1]} \mathrm{E}[(\mathrm{G}_k(t))^2]$. Since we have that almost surely, the paths of G_k are bounded, then by the Borell–TIS inequality[45, Thm. 2.1.1], we have that for $u > 0$,

$$\mathrm{P}\left(\sup_{t\in[0,1]} \mathrm{G}_k(t) - \mathrm{E}\left[\sup_{t\in[0,1]} \mathrm{G}_k(t)\right] > u\right) \le \exp(-u^2/(2\sigma^2_{\mathrm{G}_k})) \overset{(23)}{\le} \exp(-u^2/C'^2).$$

Rewriting, we get

$$\mathrm{P}\left(\sup_{t\in[0,1]} \mathrm{G}_k(t) \le u + \mathrm{E}\left[\sup_{t\in[0,1]} \mathrm{G}_k(t)\right]\right) \ge 1 - \exp(-u^2/C'^2).$$

The final step is to provide a uniform upper bound on $\mathrm{E}\left[\sup_{t\in[0,1]} \mathrm{G}_k(t)\right]$. From Dudley's theorem[45, Thm. 1.3.13], there exists a universal constant K such that

$$\mathrm{E}\left[\sup_{t\in[0,1]} \mathrm{G}_k(t)\right] \le K\int_0^{diam([0,1])/2} \sqrt{\log(\mathcal{N}([0,1],d,\varepsilon))}\,d\varepsilon \tag{24}$$

where $d(s,t) = \left(\mathrm{E}\left[(\mathrm{G}_k(s) - \mathrm{G}_k(t))^2\right]\right)^{1/2}$, $diam([0,1])$ is the maximum distance under d between two points in $[0,1]$, and $\mathcal{N}([0,1],d,\varepsilon)$ is the smallest number of balls of length ε that cover $[0,1]$ under d. Next, we claim that there is a constant $\tilde{C} > 0$ such that

$$(d(s,t))^2 \le \tilde{C}|s-t|. \tag{25}$$

We prove this claim in Sec. A.2. So, we can provide the following upper bounds:

$$diam([0,1]) = \max_{s,t\in[0,1]} d(s,t) \le \sqrt{\tilde{C}}, \text{ and}$$

$$\mathcal{N}([0,1],d,\varepsilon) \le \mathcal{N}([0,1],|\cdot|,\varepsilon^2/\tilde{C}) \overset{(a)}{\le} \begin{cases} \frac{3\tilde{C}}{\varepsilon^2} & \text{if} \quad \varepsilon \le \sqrt{\tilde{C}} \\ 1 & \text{if} \quad \epsilon > \sqrt{\tilde{C}} \end{cases}$$

where (a) follows from[46, Eq. 4.10]. Plugging this into (24), we get

$$\mathrm{E}\left[\sup_{t\in[0,1]} \mathrm{G}_k(t)\right] \leq K\int_0^{\sqrt{C/2}} \sqrt{\log\left(\frac{3\tilde{C}}{\varepsilon^2}\right)}\, d\varepsilon \triangleq \tilde{K} < \infty.$$

Going back to the probability term, we get

$$\mathrm{P}\left(\sup_{t\in[0,1]} \mathrm{G}_k(t) \leq u + \tilde{K}\right) \geq 1 - \exp(-u^2/C'^2).$$

Thus, we have that $\sup_{t\in[0,1]} \mathrm{G}_k(t) = \mathcal{O}_p(1)$ with constants that do not depend on k or the functions w_j. Thus, we have that

$$\|\mathrm{G}_k\|^2_{L^2(0,1)} \leq \left\|\sup_{t\in[0,1]} \mathrm{G}_k(t)\right\|^2_{L^2(0,1)} = \left(\sup_{t\in[0,1]} \mathrm{G}_k(t)\right)^2 = \mathcal{O}_p(1).$$

A.2. Proof of claim 83

First, let $t \geq s$. Since G_k has mean 0 at any time,

$(d(t,s))^2$	$= \mathrm{E}[(\mathrm{G}_k(t) - \mathrm{G}_k(s))^2]$	
	$= Var(\mathrm{G}_k(t) - \mathrm{G}_k(s))$	
	$= Var(\mathrm{G}_k(t)) + Var(\mathrm{G}_k(s)) - 2Cov(\mathrm{G}_k(t), \mathrm{G}_k(s))$	
	$= \frac{1}{k}\sum_{j=1}^{k}\left[w_j(t)^2(t-t^2) + w_j(s)^2(s-s^2) - 2w_j(t)w_j(s)(s-st)\right]$	

Now, we just consider one summand since the bound will apply for all summands:

$$w_j(t)^2(t-t^2) + w_j(s)^2(s-s^2) - 2w_j(t)w_j(s)(s-st)$$

$$= w_j(t)^2 t - w_j(t)^2 t^2 + w_j(s)^2 s - w_j(s)^2 s^2 - 2w_j(t)w_j(s)s + 2w_j(t)w_j(s)st$$

$$= w_j(s)^2 s - 2w_j(t)w_j(s)s + w_j(t)^2 s - w_j(t)^2 s + w_j(t)^2 t$$

$$-w_j(t)^2 t^2 - w_j(s)^2 s^2 + 2w_j(t)w_j(s)st$$

$$= w_j(s)^2 s - 2w_j(t)w_j(s)s + w_j(t)^2 s - w_j(t)^2 s + w_j(t)^2 t - (w_j(t)t - w_j(s)s)^2$$

$$\leq w_j(s)^2 s - 2w_j(t)w_j(s)s + w_j(t)^2 s - w_j(t)^2 s + w_j(t)^2 t$$

$$\leq s(w_j(t) - w_j(s))^2 + w_j(t)^2(t-s)$$

$$\leq (w_j(t) - w_j(s))^2 + w_j(t)^2(t-s)$$

$$\overset{(a)}{\leq} L^2(t-s)^2 + w_j(t)^2(t-s)$$

$$\overset{(b)}{\sim} C(t-s)$$

where (a) follows from the w_j functions being L-Lipschitz, and (b) follows from the fact that $t, s \in [0, 1]$ and there is a multiplicative constant that makes $(t-s)$ dominate $(t-s)^2$ with in $[0, 1]$.

A.3. Proof of (16): Asymptotic bias bound

For this proof, we are conditioned on the number of neighbors $\left|\mathcal{N}_{i,\eta}\right|$. We do not write this for notational brevity. First, we claim the following finite-sample error bound on the bias (proven after this proof): There exists universal constants c, c_f and K such that for any fixed $\delta \in (0, 1)$,

$$\mathrm{P}\left(2W_2^2\left(\tilde{\mu}_{ij}, \mu_{ij}\right) \leq 2c_f\left(\eta + K\sqrt{\frac{2}{cMp^2}\log\left(\frac{2N}{\delta}\right)}\right) \,\middle|\, \left|\mathcal{N}_{i,\eta}\right| \geq 1\right) \geq 1 - \delta. \tag{26}$$

Taking the complement, we have

$$\mathrm{P}(2W_2^2\left(\bar{\mu}_{ij}, \mu_{ij}\right) \geq 2c_f\left(\eta + 2K\sqrt{\frac{2}{cMp^2}\log\left(\frac{2N}{\delta}\right)}\right)) \leq \delta.$$

where c and c_f are constants. Note that each term in the bound is fixed except for δ. Thus, we satisfy the boundedness in probability definition. So, we have

$$W_2^2\left(\bar{\mu}_{ij}, \mu_{ij}\right) = \mathcal{O}_p\left(\eta + \sqrt{\frac{\log\ (2N)}{Mp^2}}\right).$$

We now return to our earlier claim (26)

Proof of claim (26). Similar to the arguments in[[18], Sec. 4.3], we have

$$\begin{aligned}
\bar{W}_2(\mu_{ij}, \mu_{ij}) &\overset{(3)}{=} \| F^{-1}_{\bar{\mu}_{ij}} - F^{-1}_{\mu_{ij}} \|_{L^2(0,1)} \\
&\overset{(17)}{=} \| \frac{1}{|\mathcal{N}_{i,\eta}|} \sum_{k \in \mathcal{N}_{i,\eta}} \left(F^{-1}_{\mu_{kj}} - F^{-1}_{\mu_{ij}}\right) \|_{L^2(0,1)} \\
&= \frac{1}{|\mathcal{N}_{i,\eta}|} \| \sum_{k \in \mathcal{N}_{i,\eta}} \left(F^{-1}_{\mu_{kj}} - F^{-1}_{\mu_{ij}}\right) \|_{L^2(0,1)} \\
&\overset{(a)}{\leq} \frac{1}{|\mathcal{N}_{i,\eta}|} \sum_{k \in \mathcal{N}_{i,\eta}} \| F^{-1}_{\mu_{kj}} - F_{\mu_{ij}} \|_{L^2(0,1)} \\
&\leq \frac{|\mathcal{N}_{i,\eta}|}{|\mathcal{N}_{i,\eta}|} \max_{k \in \mathcal{N}_{i,\eta}} \| F^{-1}_{\mu_{kj}} - F^{-1}_{\mu_{ij}} \|_{L^2(0,1)} \\
&\overset{(3)}{=} \max_{k \in \mathcal{N}_{i,\eta}} W_2(\mu_{kj}, \mu_{ij}).
\end{aligned}$$

where (a) follows from Minkowski's inequality[[42], Thm. 198]. By nonnegativity and squaring, we have

$$\bar{W}_2^2(\mu_{ij}, \mu_{ij}) \leq \max_{k \in \mathcal{N}_{i,\eta}} W_2^2(\mu_{kj}, \mu_{ij}).$$

Next, we claim that

$$\mathrm{E}[\rho_{iu} \left| x^{(i)}_{row}, x^{(u)}_{row}\right] = \mathrm{E}[W_2^2(Y_{ij}, Y_{uj}) \left| x^{(i)}_{row}, x^{(u)}_{row}\right] \tag{27}$$

which we prove at the end of this section. Since the latent spaces are bounded and the latent function f is Lipschitz, then the space of distributions is also bounded in Wasserstein distance. Since the space is

bounded in Wasserstein distance and since each distribution has finite support, then there exists a universal constant $y_{\max}$ such that $W_2^2(Y_{ij}, Y_{uj}) \leq y_{\max}$. Thus, we have[2]

$$\left\| W_2^2(Y_{ij}, Y_{uj}) \right\|_{\psi_2} \leq \frac{y_{\max}}{\sqrt{\ln(2)}} = K.$$

So, by the Hoeffding Inequality (Theorem 2.6.3 in[46]), we have that for any fixed row u,

$$\mathrm{P}\left(\left| \rho_{ik} - \mathrm{E}[W_2^2(Y_{ij}, Y_{uj})] \right| \geq t \,\middle|\, \mathcal{C}_{iu}, x_{row}^{(i)}, x_{row}^{(u)} \right) \leq 2\exp\left(-c \frac{t^2}{K^2} \left| \mathcal{C}_{iu} \right| \right).$$

So, by total probability, we have

$$\begin{aligned}
&\mathrm{P}\left(\left| \rho_{iu} - * W_2^2(Y_{ij}, Y_{uj}) \right| \geq t \,\middle|\, x_{row}^{(i)}, x_{row}^{(u)} \right) \\
&\leq \mathrm{P}\left(\left| \rho_{iu} - * W_2^2(Y_{ij}, Y_{uj}) \right| \geq t \,\middle|\, |\mathcal{C}_{ik}| \geq \frac{1}{2} Mp^2, x_{row}^{(i)}, x_{row}^{(u)} \right) \\
&+ \mathrm{P}\left(|\mathcal{C}_{ik}| \leq \frac{1}{2} Mp^2, x_{row}^{(i)}, x_{row}^{(u)} \right) \\
&\leq 2\exp\left(-c \frac{t^2}{2K^2} Mp^2 \right) + \exp\left(-\frac{1}{8} Mp^2 \right).
\end{aligned}$$

Taking a union bound, we can remove the conditioning to get

$$\mathrm{P}\left(\max_{k \neq i} \left| \rho_{ik} - * W_2^2(Y_{ij}, Y_{kj}) \right| \leq t \right) \geq 1 - 2N\exp\left(-c \frac{t^2}{2K^2} Mp^2 \right) - N\exp\left(-\frac{1}{8} Mp^2 \right). \tag{28}$$

Denote the event above as $\mathcal{E}$. Since the latent metric spaces $\mathcal{H}_{row}$ and $\mathcal{H}_{col}$ are bounded and since the latent function f is L-Lipschitz with respect to its row argument, then there exists a constant $c_f \geq 0$ where

$$W_2^2(\mu_{kj}, \mu_{ij}) \leq c_f x \sim \mathcal{H}_{col} \mathrm{E}[W_2^2(f(x_{row}^{(k)}, x), f(x_{row}^{(k)}, x))] \tag{29}$$

where $x \sim \mathcal{H}_{col}$ means that x is drawn from the distribution over the column latent space. Next, we have from[[47], Lem. 3] that for any distributions μ and ν with respective empirical distributions $\hat{\mu}$ and $\hat{\nu}$ derived from n samples each,

$$\mathrm{E}[W_2^2(\hat{\mu}, \hat{\nu})] \geq W_2^2(\mu, \nu) \tag{30}$$

where the expectation is taken over the randomness in the sampling. So, we have that on event $\mathcal{E}$,

$$\begin{aligned}
&\max_{k \in \mathcal{N}_{i,\eta}} W_2^2(\mu_{kj}, \mu_{ij}) \\
&\overset{(29)}{\leq} \max_{k \in \mathcal{N}_{i,\eta}} c_{f x \sim \mathcal{H}_{col}} * W_2^2(f(x_{row}^{(k)}, x), f(x_{row}^{(k)}, x)) \\
&\overset{(30)}{\leq} \max_{k \in \mathcal{N}_{i,\eta}} c_f \mathcal{H}_{col} * * W_2^2(Y_{kj}, Y_{ij}) \\
&\overset{(28)}{\leq} \max_{k \in \mathcal{N}_{i,\eta}} c_f(\rho_{ik} + t) \\
&\leq c_f(\eta + t).
\end{aligned}$$

Putting this together, we have

$$\mathrm{P}\left(W_2^2\left(\bar{\mu}_{ij}, \mu_{ij}\right) \leq c_f(\eta + t) \,\middle|\, |\mathcal{N}_{i,\eta}| \geq 1\right) \geq 1 - 2N\exp\left(-c\frac{t^2}{2K^2}Mp^2\right) - N\exp\left(-\frac{1}{8}Mp^2\right).$$

Finally, it can be easily verified that by setting

$$t = K\sqrt{\frac{2}{cMp^2}\log\left(\frac{2N}{\delta}\right)}$$

we get

$$\mathrm{P}(2W_2^2(\bar{\mu}_{ij}, \mu_{ij}) \leq 2c_f\left(\eta + K\sqrt{\frac{2}{cMp^2}\log\left(\frac{2N}{\delta}\right)}\right) \Big| \, |\mathcal{N}_{i,\eta}| \geq 1)$$

$$\geq 1 - \delta - N\exp\left(-\frac{1}{8}Mp^2\right).$$

Finally, since in the statement of Thm. 1, we condition on $\left|\mathcal{C}_{iu}\right| \geq \frac{1}{2}Mp^2$ for all $u \in [N+1]$, then we can remove the third term in the probability bound to get:

$$\mathrm{P}(2W_2^2(\bar{\mu}_{ij}, \mu_{ij}) \leq 2c_f\left(\eta + K\sqrt{\frac{2}{cMp^2}\log\left(\frac{2N}{\delta}\right)}\right) \Big| \, |\mathcal{N}_{i,\eta}| \geq 1) \geq 1 - \delta.$$

Proof of claim (27). Recall the definition of ρ_{iu} from (7):

$$\rho_{iu} \triangleq \begin{cases} \frac{1}{|\mathcal{C}_{iu}|}\sum_{j \in \mathcal{C}_{iu}} W_2^2(Y_{ij}, Y_{uj}) & \text{if} \, |\mathcal{C}_{iu}| \geq 1 \\ \infty & \text{if} \, |\mathcal{C}_{iu}| = 0. \end{cases}$$

Let S denote the randomness from sampling n points from each distribution. Next, we only need to consider the case of $\left|\mathcal{C}_{iu}\right| \geq 1$ because we give a high-probability bound on this quantity being large. Under this, we have

$$\mathrm{E}[\rho_{iu} \Big| \mathcal{C}_{iu}, x_{row}^{(i)}, x_{row}^{(u)}] = \mathrm{E}[\frac{1}{|\mathcal{C}_{iu}|}\sum_{j \in \mathcal{C}_{iu}} W_2^2(Y_{ij}, Y_{uj}) \Big| \mathcal{C}_{iu}, x_{row}^{(i)}, x_{row}^{(u)}]$$

$$= \frac{1}{\left|\mathcal{C}_{iu}\right|}\sum_{j \in \mathcal{C}_{iu}} \mathrm{E}[W_2^2(Y_{ij}, Y_{uj}) \Big| \mathcal{C}_{iu}, x_{row}^{(i)}, x_{row}^{(u)}]$$

$$= \frac{1}{|\mathcal{C}_{iu}|}\sum_{j \in \mathcal{C}_{iu}} \mathrm{E}[W_2^2(Y_{ij}, Y_{uj}) \Big| x_{row}^{(i)}, x_{row}^{(u)}]$$

where the last line follows from the independence of the missingness from the distributions. This expectation is taken over two sources of randomness: the distribution over $\mathcal{H}_{col}$ and the distribution over

the sampling from each distribution, S. Next, since each sample is drawn i.i.d. and since each column vector is also drawn i.i.d then we have that $* \, W_2^2(Y_{ij}, Y_{uj}) \left| x_{row}^{(i)}, x_{row}^{(u)} \right.$ is constant across column latent vectors. Thus, we have that

$$\mathrm{E}[\rho_{iu} \left| \mathcal{C}_{iu}, x_{row}^{(i)}, x_{row}^{(u)} \right.] = \mathrm{E}[W_2^2(Y_{ij}, Y_{uj}) \left| x_{row}^{(i)}, x_{row}^{(u)} \right.].$$

Since this holds for all $\mathcal{C}_{iu}$, we can remove the conditioning to get our claim:

$$\mathrm{E}[\rho_{iu} \left| x_{row}^{(i)}, x_{row}^{(u)} \right.] = \mathrm{E}[W_2^2(Y_{ij}, Y_{uj}) \left| x_{row}^{(i)}, x_{row}^{(u)} \right.].$$

Appendix B. Proof of Thm. 2: Asymptotic distribution of estimate

Let $t \in (0, 1)$. First, we can do a similar bias-variance decomposition as in (14):

$$F^{-1}_{\hat{\mu}_{ij}}(t) - F^{-1}_{\mu_{ij}}(t) = F^{-1}_{\hat{\mu}_{ij}}(t) - F^{-1}_{\mu_{ij}}(t) + F^{-1}_{\mu_{ij}}(t) - F^{-1}_{\mu_{ij}}(t).$$

We claim that

$$\frac{\sqrt{n_{j,M} \left| \mathcal{N}_{i,\eta_M} \right|}}{\sigma_{\mathcal{N}_{i,\eta_M}}(t)} \left(F^{-1}_{\hat{\mu}_{ij}}(t) - F^{-1}_{\mu_{ij}}(t) \right) \xrightarrow{d} \mathcal{N}(0, 1) \tag{31}$$

and

$$\frac{\sqrt{n_{j,M} \left| \mathcal{N}_{i,\eta_M} \right|}}{\sigma_{\mathcal{N}_{i,\eta_M}}(t)} \left(F^{-1}_{\mu_{ij}}(t) - F^{-1}_{\mu_{ij}}(t) \right) = o_p(1). \tag{32}$$

We use the following fact to prove both claims for random variables X_m:

	$$\text{if} \quad X_m = \mathcal{O}_p(a_m) \quad \text{and} \quad \lim_{m \to \infty} a_m = 0 \quad \text{then} \quad X_m = o_p(1).$$	

Putting together claims (31) and (32), we have

	$$\frac{\sqrt{n_{j,M}\left\lvert \mathcal{N}_{i,\eta_M} \right\rvert}}{\sigma_{\mathcal{N}_{i,\eta_M}}(t)} \left(F^{-1}_{\hat{\mu}_{ij}}(t) - F^{-1}_{\mu_{ij}}(t) \right) \xrightarrow{d} \mathcal{N}(0,1).$$	

Proof of claim (31). First, define $\hat{q}_{uj}(t) \triangleq \sqrt{n_{j,M}}\left(F^{-1}_{Y_{uj}}(t) - F^{-1}_{\mu_{uj}}(t) \right)$. Recall from the proof of Prop. 1 that we have that for each distribution μ_{uj} for $u \in \mathcal{N}_{i,\eta_M}$ and its respective approximation by a sequence of standard Brownian bridges, $\left(\mathrm{B}_{u,l} \right)_{l=1}^{n_{j,M}}$,

	$$\sup_{t \in (0,1)} \left\lvert \hat{q}_{uj}(t) - \frac{\mathrm{B}_{u,n_{j,M}}(t)}{f_{\mu_{uj}}\left(F^{-1}_{\mu_{uj}}(t) \right)} \right\rvert \overset{a.s.}{=} \mathcal{O}\left(\frac{\log n_{j,M}}{\sqrt{n_{j,M}}} \right).$$	

Next, we have

	$\sqrt{n_{j,M}\left\lvert \mathcal{N}_{i,\eta_M} \right\rvert}\left(F^{-1}_{\hat{\mu}_{ij}}(t) - F^{-1}_{\mu_{ij}}(t) \right)$	$\overset{(9)}{=} \sqrt{n_{j,M}\left\lvert \mathcal{N}_{i,\eta_M} \right\rvert}\left(\frac{1}{\left\lvert \mathcal{N}_{i,\eta_M} \right\rvert} \sum_{u \in \mathcal{N}_{i,\eta_M}} F^{-1}_{Y_{uj}}(t) - F^{-1}_{\mu_{uj}}(t) \right)$
		$= \frac{1}{\sqrt{\left\lvert \mathcal{N}_{i,\eta_M} \right\rvert}} \sum_{u \in \mathcal{N}_{i,\eta_M}} \hat{q}_{uj}(t)$

Now, we will proceed similar to the proof of Prop. 1 by adding and subtracting the Brownian bridge approximation:

doi.org/10.32388/DX6MDB

$$\frac{1}{\sqrt{\left|\mathcal{N}_{i,\eta_M}\right|}}\sum_{u\in\mathcal{N}_{i,\eta_M}}\hat{q}_{uj}(t)$$

$$=\frac{1}{\sqrt{\left|\mathcal{N}_{i,\eta_M}\right|}}\sum_{u\in\mathcal{N}_{i,\eta_M}}\left(\hat{q}_{uj}(t)-\frac{\mathrm{B}_{u,n_{j,M}}(t)}{f_{\mu_{uj}}\left(F^{-1}_{\mu_{uj}}(t)\right)}+\frac{\mathrm{B}_{u,n_{j,M}}(t)}{f_{\mu_{uj}}\left(F^{-1}_{\mu_{uj}}(t)\right)}\right)$$

$$=\frac{1}{\sqrt{\left|\mathcal{N}_{i,\eta_M}\right|}}\sum_{u\in\mathcal{N}_{i,\eta_M}}\left(\hat{q}_{uj}(t)-\frac{\mathrm{B}_{u,n_{j,M}}(t)}{f_{\mu_{uj}}\left(F^{-1}_{\mu_{uj}}(t)\right)}\right)+\frac{1}{\sqrt{\left|\mathcal{N}_{i,\eta_M}\right|}}\sum_{u\in\mathcal{N}_{i,\eta_M}}\frac{\mathrm{B}_{u,n_{j,M}}(t)}{f_{\mu_{uj}}\left(F^{-1}_{\mu_{uj}}(t)\right)}$$

Analyzing the first sum, we have from the following bound from applying proposition 1 just like in the proof of Lem. 2:

$$\frac{1}{\sqrt{\left|\mathcal{N}_{i,\eta_M}\right|}}\sum_{u\in\mathcal{N}_{i,\eta_M}}\left(\hat{q}_{uj}(t)-\frac{\mathrm{B}_{u,n_{j,M}}(t)}{f_{\mu_{uj}}\left(F^{-1}_{\mu_{uj}}(t)\right)}\right)$$

$$\leq\left|\frac{1}{\sqrt{\left|\mathcal{N}_{i,\eta_M}\right|}}\sum_{u\in\mathcal{N}_{i,\eta_M}}\left(\hat{q}_{uj}(t)-\frac{\mathrm{B}_{u,n_{j,M}}(t)}{f_{\mu_{uj}}\left(F^{-1}_{\mu_{uj}}(t)\right)}\right)\right|$$

$$\overset{(a)}{\leq}\frac{1}{\sqrt{\left|\mathcal{N}_{i,\eta_M}\right|}}\sum_{u\in\mathcal{N}_{i,\eta_M}}\left|\hat{q}_{uj}(t)-\frac{\mathrm{B}_{u,n_{j,M}}(t)}{f_{\mu_{uj}}\left(F^{-1}_{\mu_{uj}}(t)\right)}\right|$$

$$\overset{(18)}{=}\mathcal{O}\left(\sqrt{\left|\mathcal{N}_{i,\eta_M}\right|}\frac{\log n_{j,M}}{\sqrt{n}}\right)$$

$$\overset{11}{=}o_p(1)$$

where (a) follows from the triangle inequality. Analyzing the second sum, we have

$$\frac{1}{\sqrt{\left|\mathcal{N}_{i,\eta_M}\right|}}\sum_{u\in\mathcal{N}_{i,\eta_M}}\frac{\mathrm{B}_{u,n_{j,M}}(t)}{f_{\mu_{uj}}\left(F^{-1}_{\mu_{uj}}(t)\right)}\qquad\overset{d}{=}\frac{1}{\sqrt{\left|\mathcal{N}_{i,\eta_M}\right|}}\sum_{u\in\mathcal{N}_{i,\eta_M}}\frac{X_u}{f_{\mu_{uj}}\left(F^{-1}_{\mu_{uj}}(t)\right)}$$

where $\{X_u\}_{u\in\mathcal{N}_{i,\eta_M}}$ are i.i.d. $\mathcal{N}(0,t-t^2)$ random variables because $\mathrm{B}_{u,n_{j,M}}(t)\sim\mathcal{N}(0,t-t^2)$ from the definition of a standard Brownian bridge[[48], Prop. 8.1.1]. Next, the sum of independent mean-zero Gaussian

random variables is Gaussian with a variance equal to the sum of summand's variances and mean-zero. Thus, we have

$$\frac{1}{\sqrt{\left|\mathcal{N}_{i,\eta_M}\right|}}\sum_{u\in\mathcal{N}_{i,\eta_M}}\frac{X_u}{f_{\mu_{ij}}\left(F^{-1}_{\mu_{ij}}(t)\right)}\sim\mathcal{N}\left(0,\frac{1}{\left|\mathcal{N}_{i,\eta_M}\right|}\sum_{u\in\mathcal{N}_{i,\eta_M}}\frac{t-t^2}{f^2_{\mu_{ij}}\left(F^{-1}_{\mu_{ij}}(t)\right)}\right)\overset{(12)}{=}\mathcal{N}(0,\sigma^2_{\mathcal{N}_{i,\eta}}(t))$$

Thus, we have

$$\frac{1}{\sigma_{\mathcal{N}_{i,\eta}}(t)\sqrt{\left|\mathcal{N}_{i,\eta_M}\right|}}\sum_{u\in\mathcal{N}_{i,\eta_M}}\frac{\mathrm{B}_{u,n_{j,M}}(t)}{f_{\mu_{ij}}\left(F^{-1}_{\mu_{ij}}(t)\right)}\sim\mathcal{N}(0,1)$$

which comples the proof of claim (31).

Proof of claim (32). Here, we consider the sequence $\frac{\sqrt{n_{j,M}\left|\mathcal{N}_{i,\eta_M}\right|}}{\sigma_{\mathcal{N}_{i,\eta_M}}(t)}\left(F^{-1}_{\mu_{ij}}(t)-F^{-1}_{\mu_{ij}}(t)\right)$. From the proof of Thm. 1, we have:

$$\|F^{-1}_{\mu_{ij}}-F^{-1}_{\mu_{ij}}\|_{L^2(0,1)}\overset{(3)}{=}W_2\left(\bar{\mu}_{ij},\mu_{ij}\right)\overset{(16)}{=}\mathcal{O}_p\left(\left(\eta_M+\sqrt{\frac{\log\left(2N_n\right)}{M_n p^2}}\right)^{1/2}\right). \tag{33}$$

Now, let $a_M(t)=\frac{\sqrt{n_{j,M}\left|\mathcal{N}_{i,\eta_M}\right|}}{\sigma_{\mathcal{N}_{i,\eta_M}}(t)}$. Thus, we have

$\|a_M(t)\left(F^{-1}_{\mu_{ij}}-F^{-1}_{\mu_{ij}}\right)\|_{L^2(0,1)}$	$=a_M(t)\|F^{-1}_{\mu_{ij}}-F^{-1}_{\mu_{ij}}\|_{L^2(0,1)}$
	$\overset{(33)}{=}\mathcal{O}_p\left(a_M(t)\left(\eta_M+\sqrt{\frac{\log\left(2N_M\right)}{Mp^2}}\right)^{1/2}\right)$
	$\overset{(11)}{=}o_p(1)$

Next, from Assum. 3, each distribution has an L-Lipschitz quantile function and is thus continuously differentiable, which makes it Lipschitz as well. So, there exists a universal Lipschitz constant for the quantile functions since the union of all quantile functions have bounded range (because the union of the distributions has bounded support). Thus, $\{F^{-1}_{\mu_{ij}}\}_{n=1}^{\infty}$ is equicontinuous[[49], Examples 11.15] (For the definition of equicontinuous sets of functions see[[49], Ch. 11]. Next, from[[50], Lem. 3.2], we know that if a sequence of equicontinuous functions converges in $L^2(0,1)$ then it also converges uniformly. So, we have that for any $t \in (0,1)$, $a_M(t)\left(\tilde{F}^{-1}_{\mu_{ij}}(t) - F^{-1}_{\mu_{ij}}(t)\right) = o_p(1)$.

Appendix C. Corollaries

Let $\phi(x,r) = \tilde{x} \sim \mathcal{H}_{row} * {}_{v} \sim \mathcal{H}_{col} * W_2^2(f(x,v), \tilde{f}(x,v)) \leq r$. We require the following two lemmas to remove the conditioning. First, we have a bound on the probability of having no neighbors: Next, we have a high-probability lower bound on the number of neighbors (proven in Sec. C.2):

Lemma 4 (Lower bound on number of neighbors). *Let n_{-j} be the number of samples in each matrix entry not in column j. Let there exist constants c_1 and K such that $\eta' \geq \frac{6c_1}{n_{-j}}$ and $\eta \geq \eta' + K\sqrt{\frac{4\log(N)}{cMp^2}}$. Let $\mathcal{N}_{i,\eta}$ be the nearest neighbors for row i. Then we have*

$$\mathrm{P}\left(\left|\mathcal{N}_{i,\eta}\right| \geq \frac{1}{2}N\tilde{p}_{i,\eta'} \,\middle|\, x^{(i)}_{row}\right) \geq 1 - \exp\left(-\frac{N\tilde{p}_{i,\eta'}}{8}\right) \quad \text{where}$$

$$\tilde{p}_{i,\eta'} \triangleq \left(1 - \frac{1}{N^2} - \exp\left(-\frac{Mp^2}{8}\right)\right) \cdot p \cdot \phi\left(x^{(i)}_{row}, \frac{\eta'}{3} - \frac{6c_1}{n_{-j}}\right).$$

Next, we have a simplified lower bound for the previous lemma:

Lemma 5 (Simplified lower bound on number of neighbors). *We provide a simplified lower bound on $\frac{1}{2}N\hat{p}_{i,\eta'}$, to give a lower bound on the number of neighbors. For N, M large enough, we have*

	$\frac{1}{2}N\hat{p}_{i,\eta'}$	$\geq \frac{1}{4}Np \cdot \phi\left(x_{row}^{(i)}, \frac{\eta'}{3} - \frac{6c_1}{n_{-j}}\right).$	

Now, we are prepared to state and prove our corollary.

C.1. Latent factors drawn from uniform hypercube

Here, we provide a general corollary where the latent factors are drawn from a uniform hypercube. This case covers Cor. 1.

Corollary 2 (Uniform measure on hypercube). *Let $\mathcal{H}_{row} = [0, 1]^d$ for some $d \geq 1$. Let Assums. 1 to 3 hold. Let d_{row} and d_{col} be the Euclidean measure and μ_{row} be the uniform measure. Let $N, M,$ and p be fixed. Let $n_{-j} = n_v$ for $v \neq j$. Conditioned on the number of neighbors being at least $\frac{1}{4}(Np)^{\frac{2}{d+2}}$ (event $\mathcal{E}_1$) and the number of shared columns between row i and row u being at least $\frac{1}{2}Mp^2$ for $u \in [N+1]$ (event $\mathcal{E}_2$), we have*

	$W_2^2(\hat{\mu}_{ij}, \mu_{ij})$	$= \tilde{\mathcal{O}}_p\left(\frac{1}{n_j(Np)^{\frac{2}{d+2}}} + \frac{1}{p\sqrt{M}} + \frac{1}{n_{-j}}\right) \quad \text{as} \quad n_j \to \infty,$	
	$* \mathcal{E}_1$	$\geq 1 - (N+1)\exp\left(-Mp^2/8\right), \text{and}$	
	$* \mathcal{E}_2$	$\geq 1 - 2\exp\left(-(Np)^{\frac{2}{d+2}}/16\right).$	

Proof of Cor. 2. Similar to Corollary 2 in[17], we define $B(x, r) \overset{\triangle}{=} \{x' \in \mathcal{H}_{row} : d_{row}(x, x') \leq r\}$ for $r > 0$. Then, we have if $d_{row}(x, x') \leq \frac{1}{L}\left(\frac{\eta'}{3} - \frac{6c_1}{n}\right)$, then by Lipschitzness of f, we have $\mathrm{E}_{v \sim \mu_{col}}[W_2^2(f(x, v) - f(x', v))] \leq \frac{\eta'}{3} - \frac{6c_1}{n}$. Thus, we have

	$\phi\left(x, \frac{\eta'}{3} - \frac{6c_1}{n}\right)$	$\geq \mu_{row}\left(B\left(x, \frac{1}{L}\left(\frac{\eta'}{3} - \frac{6c_1}{n}\right)\right)\right) = Vol\left(B\left(x, \frac{1}{L}\left(\frac{\eta'}{3} - \frac{6c_1}{n}\right)\right)\right).$	

There are positive universal constants α and β such that for any $d \geq 1, x \in [0, 1]^d, r > 0$

$$Vol(B(x,r)) \geq \min(1, \alpha\beta^d r^d).$$

Plugging this into the inequality above, we have

$$\phi\left(x, \frac{\eta'}{3} - \frac{6c_1}{n}\right) \geq \min\left\{1, \alpha\beta^d\left(\frac{1}{L}\left(\frac{\eta'}{3} - \frac{6c_1}{n}\right)\right)^d\right\}$$

Next, $\forall v \in \mathcal{H}_{col}, x, x' \in \mathcal{H}_{row}$, $W_2(f(x,v), f(x',v)) \leq L\sqrt{d}$. Let

$$\frac{\eta'}{3} = \frac{6c_1}{n} + \alpha^{2/d}\beta^2 L^2 (Mp)^{-2/(d+2)}.$$

So, we have

$$Np \cdot \phi\left(x, \eta' - \frac{c_1}{n} - \frac{c_2}{\sqrt{n}}\right) \geq Np \cdot \frac{\left(\alpha^{-2/d}\beta^{-2}L^2(Np)^{-2/(d+2)}\right)^{d/2}}{\alpha\beta^d L^d} = (Np)^{\frac{2}{d+2}}.$$

Putting this into the bounds in Lems. 4 and 5, and letting η' be equal to its lower bound, we get our result.

□

C.2. Proof of Lem. 4: Lower bound on number of neighbors

Now, we place a high-probability lower bound on the number of neighbors in order to remove the conditioning on $\mathcal{N}_{i,\eta}$. Now, we must remove the conditioning on $|\mathcal{N}_{i,\eta}|$. We do this by finding a high-probability bound on a another set which can be more easily analyzed:

$$\Omega \stackrel{\triangle}{=} \left\{u \in [N+1] \setminus \{i\} : A_{uj} = 1, \mathrm{E}_{\mathcal{H}_{col}}[\mathrm{E}[W_2^2(Y_{iv}, Y_{uv})]] \leq \eta'\right\}.$$

Next, consider the set

$$\tilde{\Omega} \stackrel{\triangle}{=} \left\{ u \in \Omega : \rho_{iu} - \mathrm{E}_{\mathcal{H}_{col}}[\mathrm{E}[W_2^2(Y_{iv}, Y_{uv})]] \leq \eta - \eta' \right\}.$$

Since for $u \in \Omega$, $\mathcal{H}_{col} * * W_2^2(Y_{iv}, Y_{uv}) \leq \eta'$, then $\tilde{\Omega} \subseteq \mathcal{N}_{i,\eta}$. Thus, if we can provide a lower bound on $|\tilde{\Omega}|$, then this provides an upper bound on $\frac{1}{|\mathcal{N}_{i,\eta}|}$.

Next, we claim (proven at the end of this section)

$$\mathrm{P}(u \in \Omega \,|\, x_{row}^{(i)}) \geq p \cdot \phi\left(x_{row}^{(i)}, \frac{\eta'}{3} - \frac{6c_1}{n_{-j}}\right) \text{ and} \tag{34}$$

$$\mathrm{P}(u \in \tilde{\Omega} \,|\, u \in \Omega) \geq 1 - \frac{1}{N^2} - \exp\left(-\frac{Mp^2}{8}\right).$$

Putting these together, we find that

$$\begin{aligned} \mathrm{P}(u \in \tilde{\Omega} \,|\, x_{row}^{(i)}) &= \mathrm{P}(u \in \tilde{\Omega} \,|\, u \in \Omega, x_{row}^{(i)}) \mathrm{P}(u \in \Omega \,|\, x_{row}^{(i)}) \\ &\geq \left(1 - \frac{1}{N^2} - \exp\left(-\frac{Mp^2}{8}\right)\right) \cdot p \cdot \phi\left(x_{row}^{(i)}, \frac{\eta'}{3} - \frac{6c_1}{n_{-j}}\right) \\ &\stackrel{\triangle}{=} \tilde{p}_{i,\eta'}. \end{aligned}$$

So, by the Binomial Chernoff bound, we get

$$\mathrm{P}(|\tilde{\Omega}| \geq \frac{1}{2} N \tilde{p}_{i,\eta'} \,|\, x_{row}^{(i)}) \geq 1 - \exp\left(-\frac{N\tilde{p}_{i,\eta'}}{8}\right)$$

and since $\tilde{\Omega} \subseteq \mathcal{N}_{i,\eta}$, then we have our result.

Proof of claim (34). We have that a row u is in Ω and satisfies the above inequality with probability $p \cdot \psi(x_{row}^{(i)}, \eta')$ where

	$\psi(x_{row}^{(i)}, \eta') = \mathrm{P}(\mathrm{E}[W_2^2(Y_{ij}, Y_{uj})] \leq \eta' \mid x_{row}^{(i)})$.	

By the Binomial Chernoff Bound and conditioning on the i-th latent row vector, we have

	$\mathrm{P}(\lvert\Omega\rvert = 0 \mid x_{row}^{(i)})$	$\leq \mathrm{P}(\lvert\Omega\rvert \leq \frac{1}{2}Np \cdot \psi(x_{row}^{(i)}, \eta'))$	
		$\leq \exp\left(-\frac{Np}{8}\psi(x_{row}^{(i)}, \eta')\right)$.	

Next, there exists a universal constant c_1 such that for two empirical distributions μ_n and ν_n with corresponding true distributions μ and ν, we have

	$\mathrm{E}[W_2^2(\mu_n, \nu_n)]$	$\leq 3W_2^2(\mu, \nu) + \frac{6c_1}{n}$.	

This follows from this line of reasoning: Let $X^{(k)}$ and $Y^{(k)}$ denote the k-th order statistics of the samples from μ and ν respectively. Let F_μ^{-1} and F_ν^{-1} denote the quantile functions of μ and ν, respectively. Then, we have

	$\mathrm{E}[W_2^2(\mu_n, \nu_n)]$
	$= \frac{1}{n}\sum_{k=1}^{n} \mathrm{E}[\left(X^{(k)} - Y^{(k)}\right)^2]$
	$= \sum_{k=1}^{n} \int_{(k-1)/n}^{k/n} \mathrm{E}[\left(X^{(k)} - F_\mu^{-1}(t) + F_\mu^{-1}(t) - F_\nu^{-1} + F_\nu^{-1} - Y^{(k)}\right)^2]dt$
	$\leq 3\sum_{k=1}^{n} \int_{(k-1)/n}^{k/n} \mathrm{E}[\left(X^{(k)} - F_\mu^{-1}(t)\right)^2 + \left(F_\mu^{-1}(t) - F_\nu^{-1}\right)^2 + \left(F_\nu^{-1} - Y^{(k)}\right)^2]dt$
	$= 3W_2^2(\mu, \nu) + 3\mathrm{E}[W_2^2(\mu_n, \mu)] + 3\mathrm{E}[W_2^2(\nu_n, \nu)]$
	$\leq 3W_2^2(\mu, \nu) + \frac{6c_1}{n_{-j}}$.

Then, let $\mathrm{E}[W_2^2(\mu_{ij}, \mu_{uj})] \leq \frac{\eta'}{3} - \frac{6c_1}{n_{-j}}$. So, we get

	$\mathrm{E}[W_2^2(Y_{ij}, Y_{uj})] \leq \eta'.$	

Thus, we have that the bound on $\mathrm{E}[W_2^2(\mu_{ij}, \mu_{uj})]$ implies the bound on $\mathrm{E}[W_2^2(Y_{ij}, Y_{uj})]$. So, we have

	$\mathrm{P}(\mathrm{E}[W_2^2(\mu_{ij}, \mu_{uj})] \leq \frac{\eta'}{3} - \frac{6c_1}{n_{-j}}) \leq \mathrm{P}(\mathrm{E}[W_2^2(Y_{ij}, Y_{uj})] \leq \eta').$	

Rewriting this in the ϕ and ψ notation, we have that

	$\phi\left(x_{row}^{(i)}, \frac{\eta'}{3} - \frac{6c_1}{n_{-j}}\right) \leq \psi(x_{row}^{(i)}, \eta').$	

Thus, we obtain the first part of our claim. For the second part, we have

	$\mathrm{P}(\rho_{iu} - \mathrm{E}_{\mathcal{H}_{col}}[\mathrm{E}[W_2^2(Y_{iv}, Y_{u_0v})]] > \eta - \eta' \left\| x_{row}^{(i)}, \|\Omega\| \geq 1\right.)$	
	$\leq \mathrm{P}(\rho_{iu} - \mathrm{E}_{\mathcal{H}_{col}}[\mathrm{E}[W_2^2(Y_{iv}, Y_{u_0v})]] > \eta - \eta' \left\| x_{row}^{(i)}, \|\mathcal{C}_{iu}\| \geq \frac{1}{2}Mp^2\right.)$	
	$+\mathrm{P}(\|\mathcal{C}_{iu}\| < \frac{1}{2}Mp^2 \| x_{row}^{(i)})$	
	$\leq \exp\left(-c\frac{(\eta-\eta')^2}{K^2}\frac{1}{2}Mp^2\right) + \exp\left(-\frac{Mp^2}{8}\right)$	
	$\leq \exp(-2\log(N)) + \exp\left(-\frac{Mp^2}{8}\right)$	
	$= \frac{1}{N^2} + \exp\left(-\frac{Mp^2}{8}\right)$	

which completes the proof of our claim.

Appendix D. Continuous uniform location-scale case

If we restrict ourselves to just the continuous uniform case, we can analyze the rate with more precision. The uniform distribution is one of the only cases where the expected squared Wasserstein distance

between an empirical distribution and its true distribution can be exactly calculated. Let Θ denote asymptotic upper and lower bounds. From[33], we have for $\mu = Unif(0, 1)$ and μ_n being the empirical distribution of n samples from μ:

$$\mathrm{E}[W_2^2(\mu_n, \mu)] = \Theta\left(\frac{1}{n}\right)$$

and if we take the barycenter of m i.i.d. empirical distributions, then the expected squared Wasserstein distance between the Wasserstein barycenter of the empirical distributions and the true distribution is given in the following lemma:

Lemma 6 (Expected error for the empirical barycenter of uniform distrbutions). *Let $\mu_1, \ldots, \mu_m$ be Uniform distributions on $[a_i, b_i]$. Let μ be the barycenter of $\mu_1, \ldots, \mu_m$. Let $X_{i,1}, \ldots, X_{i,n} \sim \mu_i$. Let $X_i^{(k)}$ denote the k-th order statistic for the i-th distribution. Let $\hat{\mu}$ denote the empirical barycenter of the m empirical distributions. Let $\mathbf{\mu}$ denote the random measure drawn from the empirical distribution of $\{\mu_i\}$. Let $\bar{X}^{(k)}$ denote the k-th order statistic μ. Then, we have*

$$\mathrm{E}[W_2^2(\hat{\mu}, \mu)] = \Theta\left(\frac{1}{m} + \frac{1}{mn} + \frac{1}{n^2}\right)$$

where Big-Θ notation denotes both upper and lower rates.

This case is a scenario where increasing the number of neighbors not only reduces the error, but also improves the error decay rate with respect to the number of samples, n. We see that as the number of neighbors increases, the rate with respect to the number of samples improves from $O\left(\frac{1}{n}\right)$ to $O\left(\frac{1}{n^2}\right)$. This is expected since as the number of neighbors increases, the Wasserstein barycenter's support points approach their expected values, which provide a better quantization of the true distribution. From[[51], Thm. 2.1(b)], we know that for quantization of 1-dimensional probability distributions, $O\left(\frac{1}{n^2}\right)$ is the best rate. While we do not have a proof for the location-scale Gaussian case, we show empirically in Sec. 5 that even in the Gaussian case, the sample error rate improves as the number of neighbors increases. Note

that for our theorem, we require η to be lower bounded by a term that is on the order of $O\left(\frac{1}{n}\right)$. Note that while the uniform case motivates why adding more neighbors improves our error rate, it does not provide a proof that this improvement will always occur.

Proof of Lem. 6. First, we calcualate the distribution of the barycenter μ:

$$F_{\mu}^{-1}(t) = \frac{1}{m}\sum_{i=1}^{m} F_{\mu_i}^{-1}(t) = \frac{1}{m}\sum_{i=1}^{m}\left[a_i + t(b_i - a_i)\right] = \left(\frac{1}{m}\sum_{i=1}^{m} a_i\right) + t\left(\frac{1}{m}\sum_{i=1}^{m} b_i - \frac{1}{m}\sum_{i=1}^{m} a_i\right).$$

Let $\bar{a} = \frac{1}{m}\sum_{i=1}^{m} a_i$ and $\bar{b} = \frac{1}{m}\sum_{i=1}^{m} b_i$. Then, $\mu = Unif(\bar{a}, \bar{b})$. Next, from[37, Thm. 3.1] we have

$$\begin{aligned} \mathrm{E}[W_2^2(\hat{\mu}, \mu)] &= \frac{1}{m}\mathrm{E}[W_2^2(\boldsymbol{\mu}, \mu)] + \frac{1}{mn}\sum_{k=1}^{n} Var\left(\bar{X}^{(k)}\right) \\ &\quad + \sum_{k=1}^{n}\int_{(k-1)/n}^{k/n}\left(\mathrm{E}[\bar{X}^{(k)}] - F_{\mu}^{-1}(t)\right)^2 dt. \end{aligned}$$

We analyze each term independently:

$$\frac{1}{m}\mathrm{E}[W_2^2(\boldsymbol{\mu}, \mu)] = \frac{1}{m^2}\sum_{i=1}^{m} W_2^2(\mu_i, \mu) = \Theta\left(\frac{1}{m}\right)$$

since the Wasserstein distance between two Uniform distributions is $O(1)$ as shown in Lem. 7. Next, we have

$$\frac{1}{mn}\sum_{k=1}^{n} Var\left(\bar{X}^{(k)}\right) = \frac{(\bar{b}-\bar{a})^2}{mn}\sum_{k=1}^{n}\frac{k(n-k+1)}{(n+1)^2(n+2)} \overset{(a)}{=} \frac{(\bar{b}-\bar{a})^2}{6m(n+1)} = \Theta\left(\frac{1}{mn}\right)$$

where (a) follows from the proof of Theorem 4.7 in[33]. Finally, we have

doi.org/10.32388/DX6MDB

	$\sum_{k=1}^{n} \int_{(k-1)/n}^{k/n} \left(* \bar{X}^{(k)} - F_{\mu}^{-1}(t) \right)^2 dt$	$= (\bar{b} - \bar{a})^2 \sum_{k=1}^{n} \int_{(k-1)/n}^{k/n} \left(\frac{k}{n+1} - t \right)^2 dt$	
		$\overset{(a)}{=} (\bar{b} - \bar{a})^2 \left(\frac{1}{6n} - \frac{1}{6(n+1)} \right)$	
		$= \frac{(\bar{b} - \bar{a})^2}{6n(n+1)}$	
		$= \Theta\left(\frac{1}{n^2} \right)$	

where (a) follows again from Theorem 4.7 in[33]. Putting these together, we recover the result in Eq. (3.3) in[37]:

	$* \; W_2^2\left(\hat{\mu}, \mu\right) = \Theta\left(\frac{1}{m} + \frac{1}{mn} + \frac{1}{n^2} \right).$	

Lemma 7. *Let $\mu = Unif(a, b)$ and $\nu = Unif(c, d)$. Then, we have*

	$W_2^2(\mu, \nu) = \frac{1}{3}\left[(a-c)^2 + (b-d)^2 + (a-c)(b-d) \right].$	

Proof. From the definition of the 2-Wasserstein metric, we have

	$W_2^2(\mu, \nu) = \int_0^1 \left(F_{\mu}^{-1}(t) - F_{\nu}^{-1}(t) \right)^2 dt$	$= \int_0^1 (a + (b-a)t - c - (d-c)t)^2 dt$	
		$= \frac{1}{3}\left[(a-c)^2 + (b-d)^2 + (a-c)(b-d) \right].$	

□

Lemma 8. *Let $X_1, \ldots, X_n \sim Unif(a, b)$. Let $X^{(k)}$ denote the k-th order statistic. Denote the law of X_i as μ and the empirical distribution as μ_n. Then, we have*

	$\mathrm{E}[W_2^2(\mu_n, \mu)]$	$= \frac{(b-a)^2}{6n}.$	

Next, let $Y_1, \ldots, Y_n \sim Unif(c, d)$*. Let* $Y^{(k)}$ *denote the* k*-th order statistic. Denote the law of* Y_i *as* ν *and the empirical distribution as* ν_n*. Then, we get*

	$\mathrm{E}[W_2^2(\mu_n, \nu_n)] = W_2^2(\mu, \nu) + \frac{(b-a)(d-c)}{3(n+1)}.$	

Proof. We utilize identities from the proof of Theorem 4.7 in[33]. We know that for $U_1, \ldots, U_n \sim Unif(0, 1)$, we have that $U^{(k)} \sim Beta(k, n-k+1)$. So, we have

	$\mathrm{E}[U^{(k)}] = \frac{k}{n+1}, Var\left(U^{(k)}\right) = \frac{k(n-k+1)}{(n+1)^2(n+2)}, \mathrm{E}[X^{(k)}] = a + (b-a)\frac{k}{n+1}, \text{and}$	

	$Var\left(X^{(k)}\right) = Var\left(a + (b-a)U^{(k)}\right) = (b-a)^2 Var\left(U^{(k)}\right) = (b-a)^2 \frac{k(n-k+1)}{(n+1)^2(n+2)}.$	

Putting these together, we get

	$\mathrm{E}[W_2^2(\mu_n, \mu)]$	$= \frac{1}{n}\sum_{k=1}^{n} Var\left(X^{(k)}\right) + \sum_{k=1}^{n} \int_{(k-1)/n}^{k/n} \left(\mathrm{E}[X^{(k)}] - F_\mu^{-1}(t)\right)^2 dt$	
		$= \frac{1}{n}\sum_{k=1}^{n} \frac{(b-a)^2 k(n-k+1)}{(n+1)^2(n+2)}$	
		$+ \sum_{k=1}^{n} \int_{(k-1)/n}^{k/n} \left(a + (b-a)\frac{k}{n+1} - a - (b-a)t\right)^2 dt$	
		$= (b-a)^2 \frac{1}{n}\sum_{k=1}^{n} \frac{k(n-k+1)}{(n+1)^2(n+2)} + (b-a)^2 \sum_{k=1}^{n} \int_{(k-1)/n}^{k/n} \left(\frac{k}{n+1} - t\right)^2 dt$	
		$= \frac{(b-a)^2}{6n}.$	

For the second part, we have

	$\mathrm{E}[W_2^2(\mu_n, \nu_n)]$	$= \frac{1}{n}\sum_{k=1}^{n} \mathrm{E}[\left(X^{(k)} - Y^{(k)}\right)^2]$	
		$= \frac{1}{n}\sum_{k=1}^{n} \left[\mathrm{E}[\left(X^{(k)}\right)^2] + \mathrm{E}[\left(Y^{(k)}\right)^2] - 2\mathrm{E}[X^{(k)}]\mathrm{E}[Y^{(k)}]\right]$	
		$= \frac{1}{n}\sum_{k=1}^{n} \left[Var\left(X^{(k)}\right) + Var\left(Y^{(k)}\right) + \left(\mathrm{E}[X^{(k)}] - \mathrm{E}[Y^{(k)}]\right)^2\right]$	
		$= \frac{(b-a)^2+(d-c)^2}{6(n+1)} + \frac{1}{n}\sum_{k=1}^{n} \left(\mathrm{E}[X^{(k)}] - \mathrm{E}[Y^{(k)}]\right)^2.$	

Next, we have

	$\frac{1}{n}\sum_{k=1}^{n} \left(\mathrm{E}[X^{(k)}] - \mathrm{E}[Y^{(k)}]\right)^2$
	$= \frac{1}{n}\sum_{k=1}^{n} \left[a + (b-a)\frac{k}{n+1} - \left(c + (d-c)\frac{k}{n+1}\right)\right]^2$
	$= \frac{1}{n}\sum_{k=1}^{n} \left[(a-c) + ((b-a)-(d-c))\frac{k}{n+1}\right]^2$
	$= \frac{1}{n}\sum_{k=1}^{n} \left[(a-c)^2 + 2(a-c)((b-a)-(d-c))\frac{k}{n+1} + ((b-a)-(d-c))^2\frac{k^2}{(n+1)^2}\right]$
	$= (a-c)^2 + \frac{2(a-c)((b-a)-(d-c))}{n(n+1)}\left(\sum_{k=1}^{n} k\right) + \frac{((b-a)-(d-c))^2}{n(n+1)^2}\left(\sum_{k=1}^{n} k^2\right)$
	$= (a-c)^2 + \frac{2(a-c)((b-a)-(d-c))}{n(n+1)} \cdot \frac{n(n+1)}{2}$
	$+ \frac{((b-a)-(d-c))^2}{n(n+1)^2} \cdot \frac{n(n+1)(2n+1)}{6}$
	$= (a-c)^2 + (a-c)((b-a)-(d-c)) + \frac{((b-a)-(d-c))^2(2n+1)}{6(n+1)}$
	$= (a-c)(b-d) + \frac{((b-a)-(d-c))^2(2n+1)}{6(n+1)}$

Putting these together, we get

doi.org/10.32388/DX6MDB

	$$\mathrm{E}[W_2^2(\mu_n, \nu_n)]$$	
	$$= \frac{(b-a)^2+(d-c)^2}{6(n+1)} + \frac{1}{n}\sum_{k=1}^{n}\left(\mathrm{E}[X^{(k)}] - \mathrm{E}[Y^{(k)}]\right)^2$$	
	$$= \frac{(b-a)^2+(d-c)^2}{6(n+1)} + (a-c)(b-d) + \frac{((b-a)-(d-c))^2(2n+1)}{6(n+1)}$$	
	$$= \frac{((b-a)-(d-c))^2}{3} + (a-c)(b-d) + \frac{(b-a)(d-c)}{3(n+1)}$$	
	$$= \frac{1}{3}\left[(a-c)^2 + (b-d)^2 + (a-c)(b-d)\right] + \frac{(b-a)(d-c)}{3(n+1)}$$	
	$$= W_2^2(\mu, \nu) + \frac{(b-a)(d-c)}{3(n+1)}$$	

□

Footnotes

[1] A stochastic process B is a standard Brownian bridge if it is a Gaussian process where for $s, t \in (0, 1)$, $\ast\ \mathrm{B}(t) = 0$, $Cov(\mathrm{B}(s), \mathrm{B}(t)) = \min(s, t) - st$.[48, Prop. 8.1.1]

[2] The Orlicz ψ_2 norm is defined as $\|X\|_{\psi_2} = \inf_{t>0} \ast \exp(|X|^2/t^2) \leq 2$.

References

1. ^Su X, Khoshgoftaar TM. A survey of collaborative filtering techniques. Advances in Artificial Intelligence. 2009;2009.
2. ^Ramlatchan A, Yang M, Liu Q, Li M, Wang J, Li Y. A survey of matrix completion methods for recommendation systems. Big Data Mining and Analytics. 2018;1(4):308–323.
3. ^Kang Z, Peng C, Cheng Q. Top-n recommender system via matrix completion. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2016;30.
4. ^Liu Z, Vandenberghe L. Interior-point method for nuclear norm approximation with application to system identification. SIAM Journal on Matrix Analysis and Applications. 2010;31(3):1235–1256.
5. ^Zhou H, Zhang D, Xie K. Accurate traffic matrix completion based on multi-gaussian models. Computer Communications. 2017;102:165–176.
6. ^Du R, Chen C, Yang B, Lu N, Guan X, Shen X. Effective urban traffic monitoring by vehicular sensor networks. IEEE Transactions on Vehicular Technology. 2015;64(1):273–286.

7. ^Du R, Chen C, Yang B, Guan X. Vanet based traffic estimation: A matrix completion approach. In: 2013 IEEE Global Communications Conference (GLOBECOM). IEEE; 2013. p. 30–35.
8. ^Nguyen LT, Kim J, Kim S, Shim B. Localization of iot networks via low-rank matrix completion. IEEE Trans actions on Communications. 2019;67(8):5833–5847.
9. ^Xie K, Li X, Wang X, Xie G, Wen J, Zhang D. Active sparse mobile crowd sensing based on matrix completio n. In: Proceedings of the 2019 International Conference on Management of Data. 2019. p. 195–210.
10. ^Dwivedi R, Tian K, Tomkins S, Klasnja P, Murphy S, Shah D. Counterfactual inference for sequential experi ments. arXiv preprint arXiv:2202.06891. 2022.
11. ^Yang S, Kalpakis K, Mackenzie CF, Stansbury LG, Stein DM, Scalea TM, et al. Online recovery of missing val ues in vital signs data streams using low-rank matrix completion. In: 2012 11th International Conference on Machine Learning and Applications. IEEE; 2012. p. 281–287.
12. ^Cai TT, Zhou W-X. Matrix completion via max-norm constrained optimization. arXiv preprint arXiv:1303.0 341. 2013.
13. a, b, cAgarwal A, Dahleh M, Shah D, Shen D. Causal matrix completion. In: The Thirty Sixth Annual Confere nce on Learning Theory. PMLR; 2023. p. 3821–3826.
14. a, bChi Y. Low-rank matrix completion [lecture notes]. IEEE Signal Processing Magazine. 2018;35(5):178–181.
15. ^Dou Y, Yang H, Deng X. A survey of collaborative filtering algorithms for social recommender systems. In: 2 016 12th International Conference on Semantics, Knowledge and Grids (SKG). IEEE; 2016. p. 40–46.
16. ^Khan BM, Mansha A, Khan FH, Bashir S. Collaborative filtering based online recommendation systems: A survey. In: 2017 International Conference on Information and Communication Technologies (ICICT). IEEE; 2017. p. 125–130.
17. a, b, c, d, e, fLi Y, Shah D, Song D, Yu CL. Nearest neighbors for matrix estimation interpreted as blind regressio n for latent variable model. IEEE Transactions on Information Theory. 2019;66(3):1760–1784.
18. a, b, c, d, e, fDwivedi R, Tian K, Tomkins S, Klasnja P, Murphy S, Shah D. Doubly robust nearest neighbors in fa ctor models. arXiv preprint arXiv:2211.14297. 2022.
19. a, bCandes EJ, Plan Y. Matrix completion with noise. Proceedings of the IEEE. 2010;98(6):925–936.
20. ^Soleymani M, Liu Q, Mahdavifar H, Balzano L. Matrix completion over finite fields: Bounds and belief pro pagation algorithms. In: 2023 IEEE International Symposium on Information Theory (ISIT). IEEE; 2023. p. 1 166–1171.
21. ^Monge G. Mémoire sur la théorie des déblais et des remblais. De l'Imprimerie Royale; 1781.

22. ^Feydy J, Charlier B, Vialard F-X, Peyré G. Optimal transport for diffeomorphic registration. In: Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20. Springer; 2017. p. 291–299.

23. ^Chen Y, Georgiou TT, Pavon M. Optimal transport in systems and control. Annual Review of Control, Robotics, and Autonomous Systems. 2021;4:89–113.

24. ^Kolouri S, Park SR, Thorpe M, Slepcev D, Rohde GK. Optimal mass transport: Signal processing and machine-learning applications. IEEE Signal Processing Magazine. 2017;34(4):43–59.

25. ^Thorpe M. Introduction to optimal transport. Notes of Course at University of Cambridge; 2018.

26. ^Cuturi M, Doucet A. Fast computation of wasserstein barycenters. In: Xing EP, Jebara T, editors. Proceedings of the 31st International Conference on Machine Learning. PMLR; 2014. p. 685–693.

27. ^Bigot J, Klein T, et al. Consistent estimation of a population barycenter in the wasserstein space. ArXiv e-prints. 2012;49.

28. ^Bigot J, Gouet R, Klein T, López A. Geodesic PCA in the Wasserstein space by convex PCA. Annales de l'Institut Henri Poincaré, Probabilités et Statistiques. 2017;53(1):1–26.

29. a, b, c, d, e, f, g, hBigot J. Statistical data analysis in the wasserstein space*. ESAIM: ProcS. 2020;68:1–19.

30. a, bAltschuler JM, Boix-Adsera E. Wasserstein barycenters are np-hard to compute. SIAM Journal on Mathematics of Data Science. 2022;4(1):179–203.

31. a, bThibaut Le Gouic, Quentin Paris, Philippe Rigollet, and Austin J Stromme. "Fast convergence of empirical barycenters in alexandrov spaces and the wasserstein space." Journal of the European Mathematical Society. 25 (6): 2229–2250, 2022.

32. ^Nicolas Fournier and Arnaud Guillin. "On the rate of convergence in wasserstein distance of the empirical measure." Probability theory and related fields. 162 (3-4): 707–738, 2015.

33. a, b, c, d, e, fSergey Bobkov and Michel Ledoux. "One-dimensional empirical measures, order statistics, and Kantorovich transport distances, volume 261." American Mathematical Society, 2019.

34. a, bFlorian F Gunsilius. "Distributional synthetic controls." Econometrica. 91 (3): 1105–1117, 2023.

35. ^Alberto Abadie. "Using synthetic controls: Feasibility, data requirements, and methodological aspects." Journal of Economic Literature. 59 (2): 391–425, June 2021.

36. ^L Wasserman. "Optimal transport and wasserstein distance," 2017.

37. a, b, cJérémie Bigot, Raúl Gouet, Thierry Klein, and Alfredo López. "Upper and lower risk bounds for estimating the Wasserstein barycenter of random measures on the real line." Electronic Journal of Statistics. 12 (2): 2

253 – 2289, 2018.

38. ^*Sourav Chatterjee. "Matrix estimation by universal singular value thresholding." The Annals of Statistics. 43 (1): 177–214, 2015.*

39. a, b*Guillaume G Martinet, Alexander Strzalkowski, and Barbara Engelhardt. "Variance minimization in the wasserstein space for invariant causal prediction." In International Conference on Artificial Intelligence and Statistics, pages 8803–8851. PMLR, 2022.*

40. a, b*Eric W Weisstein. "Bonferroni correction." https://mathworld.wolfram.com/, 2004.*

41. ^*James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. "Algorithms for hyper-parameter optimiz ation." Advances in neural information processing systems. 24, 2011.*

42. a, b, c, d, e*Godfrey Harold Hardy, John Edensor Littlewood, and George Pólya. "Inequalities." Cambridge univ ersity press, 1952.*

43. ^*Galen R Shorack and Jon A Wellner. "Empirical processes with applications to statistics." SIAM, 2009.*

44. a, b, c, d, e, f, g, h, i, j*Miklos Csorgo and Pál Révész. "Strong approximations of the quantile process." The Anna ls of Statistics, pages 882–894, 1978.*

45. a, b*Robert J Adler and Jonathan E Taylor. "Random fields and geometry." Springer Science & Business Medi a, 2009.*

46. a, b*Roman Vershynin. "High-dimensional probability." University of California, Irvine, 2020.*

47. ^*Papp T, Sherlock C. Bounds on Wasserstein distances between continuous distributions using independent samples. arXiv preprint arXiv:2203.11627. 2022.*

48. a, b*Sheldon M Ross. "Stochastic processes." John Wiley & Sons, 1995.*

49. a, b*Neal L Carothers. "Real analysis." Cambridge University Press, 2000.*

50. ^*Adriano M Garsia, Eugene Rodemich, Howard Rumsey, and M Rosenblatt. "A real variable lemma and the continuity of paths of some gaussian processes." Indiana University Mathematics Journal. 20 (6): 565–578, 1 970.*

51. ^*Gilles Pagès. "Introduction to vector quantization and its applications for numerics." ESAIM: proceedings and surveys. 48: 29–79, 2015.*

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.