Fast approximate Bayesian multidimensional scaling with consistency guarantees

Anonymous authors
Paper under double-blind review

Abstract

Bayesian multidimensional scaling (BMDS) embeds n objects in a low-dimensional space to approximately preserve an observed dissimilarity matrix. Compared to classic MDS, BMDS is more robust to model misspecification and supports posterior uncertainty quantification and joint estimation within hierarchical models. However, standard BMDS inference is computationally prohibitive, requiring $O(n^2)$ operations per MCMC iteration to evaluate the likelihood. We propose Barnes–Hut BMDS (BH-BMDS), which uses a tree-based approximation to the likelihood and a Gibbs sampler that leverages this structure, remaining compatible with hierarchical extensions. BH-BMDS reduces computational complexity to $O(n \log n)$ while preserving the geometric fidelity of the embedding. We further establish consistency for the stationary measure of BH-BMDS, proving that it concentrates around the true latent configuration even as the total error of the surrogate likelihood diverges. Notably, this consistency holds in the infinite-dimensional limit. We evaluate the approximation on datasets with diverse structure, including air traffic networks, arXiv abstracts, MNIST images and neural activity recordings from mouse models of tau pathology. Across all settings, BH-BMDS closely matches BMDS while achieving substantial computational gains, with approximately 10-fold speedups at $n = 1{,}000$ and 70-fold speedups at $n = 10{,}000$. These gains increase with n, demonstrating strong empirical scalability.

1 Introduction

Dissimilarity data arise in many domains, including psychology, physics and biology, where a goal is to quantify how pairs of complex or high-dimensional objects differ from one another. A common approach is to embed objects into a low-dimensional space, enabling visualization and downstream modeling while preserving the observed dissimilarities. Bayesian multidimensional scaling (BMDS) provides a probabilistic framework for this task, mapping n objects to latent coordinates in a low-dimensional Euclidean space while modeling observed dissimilarities through a likelihood-based formulation. Unlike classical approaches such as PCA (Pearson, 1901; Hotelling, 1933) and MDS (Torgerson, 1952), or more recent approaches such as t-SNE (van der Maaten & Hinton, 2008) and UMAP (McInnes et al., 2018), BMDS enables posterior uncertainty quantification over latent embeddings and supports integration into hierarchical models. These properties have made BMDS a valuable tool in applications such as viral phylogeography (Bedford et al., 2015; Holbrook et al., 2021; Li et al., 2023), incorporation of auxiliary information (Lin & Fong, 2019) and clustering (Oh & Raftery, 2007; Sheth et al., 2026).

Despite these advantages, BMDS is computationally constrained by the $O(n^2)$ cost of likelihood evaluation, which requires summing over all pairwise distances (Equation 2). As n grows, this quadratic scaling becomes prohibitive, restricting BMDS inference to relatively small datasets. Existing approximation methods (Sheth et al., 2026) lack rigorous posterior consistency guarantees for two or more dimensional latent spaces, raising concerns about their practical reliability. To address these limitations, we propose an approximation to BMDS based on the Barnes–Hut algorithm (BH-BMDS). This framework scales as $O(n \log n)$, preserves the original Bayesian formulation and guarantees the consistency for the stationary measure under the approximation and its stochastic realizations.

2 Related Work

Several approaches attempt to mitigate the cost of BMDS. Bedford et al. (2015) adopt a misspecified Gaussian likelihood to bypass the computational bottleneck imposed by non-negativity constraints; however, this assumption places probability mass outside the support of the observed data. Holbrook et al. (2021) leverage multi-core central processing units, vectorization and graphic processing units to significantly reduce runtime but depend on specialized hardware for scalability. Both methods require $O(n^2)$ floating point operations. To achieve sub-quadratic scaling, Sheth et al. (2026) induce sparsity in the dissimilarity matrix, achieving $O(mn)$ complexity, where $m \ll n$ represents the number of landmarks. Although this approach yields substantial speedups, m must scale with n to maintain accuracy, and its selection introduces an additional tuning parameter. Moreover, they do not provide posterior consistency guarantees for subsampled dissimilarity matrices in dimensions $d > 1$.

Tree-based methods offer an alternative strategy for approximating pairwise interactions. For instance, van der Maaten (2014) use the Barnes–Hut algorithm to accelerate t-SNE by approximating long-range interactions in an n-body system. While this scales as $O(n \log n)$, the paper lacks formal accuracy bounds for the approximation error. Noting that the BMDS likelihood admits a similar n-body interpretation, we adapt the Barnes–Hut algorithm to the BMDS setting (BH-BMDS), enabling $O(n \log n)$ likelihood approximations within an MCMC framework. We further establish consistency under uniform vanishing pairwise loss error for BH-BMDS, proving the stationary measure concentrates on the true configuration despite errors introduced by the approximation and noisy MCMC updates.

3 Preliminaries

3.1 Bayesian multidimensional scaling

Suppose we have high-dimensional observations, e.g., $\mathbf{y}_1, \dots \mathbf{y}_n \in \mathbb{R}^{\Omega}$, and let $D_{ij} = \mathcal{D}(\mathbf{y}_i, \mathbf{y}_j)$ denote the dissimilarity between observations. Bayesian multidimensional scaling (BMDS) (Oh & Raftery, 2001) assumes that the observed dissimilarities follow

$$D_{ij} \;=\; \left[\, r_{ij} + \sigma Z_{ij} \,\right]_+, \qquad r_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|\,, \qquad Z_{ij} \overset{\text{i.i.d.}}{\sim} N(0,1), \tag{1}$$

independent across pairs and conditioned on the low-dimensional latent locations $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^d$. Here, $[t]_+ = \max\{t, 0\}$ represents truncation at zero, and the BMDS variance σ^2 represents the level of noise distorting the latent distances. This yields the log-likelihood

$$\ell(\mathbf{D} \mid \mathbf{X}, \sigma^2) = -\frac{N}{2}\log(2\pi\sigma^2) - \sum_{i<j}\left[\frac{(D_{ij} - r_{ij})^2}{2\sigma^2} + \log\Phi\left(\frac{r_{ij}}{\sigma}\right)\right], \tag{2}$$

where $\mathbf{D} = \{D_{ij}\}$ is the observed symmetric $n \times n$ dissimilarity matrix, $N = \binom{n}{2}$, and $\Phi(\cdot)$ is the standard normal CDF. Because the likelihood involves a summation over all unordered pairs, its evaluation scales as $O(n^2)$. Consequently, we seek to reduce this computational burden using tree-based approximation methods, such as Barnes–Hut algorithms, that scale as $O(n \log n)$.

3.2 Barnes–Hut algorithm

The gravitational n-body problem involves computing the gravitational forces exerted by all pairs of particles on one another. Direct evaluation of these forces requires $O(n^2)$ computations which becomes prohibitive for large datasets. To reduce this cost, Barnes & Hut (1986) propose a hierarchical approximation algorithm that estimates the influence of distant particles as a single aggregate interaction.

To organize groups of particles efficiently, the Barnes–Hut algorithm uses a tree data structure, e.g., a quadtree $\mathcal{T} \in \mathbb{R}^2$ (Figure 1). A quadtree $\mathcal{T}$ is built by recursively subdividing the embedding space into four quadrants based on particle locations (Algorithm B.1). A cell refers to a fixed spatial region, and a node corresponds to a cell in the tree. Each leaf node contains at most one particle, while internal nodes

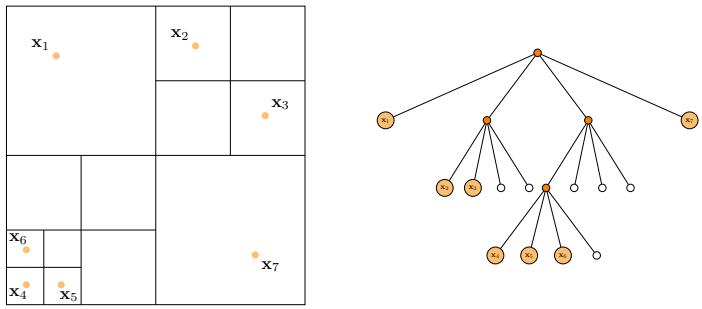

Figure 1: Barnes–Hut structural components. The left figure is the spatial representation of a recursive subdivision of the embedding space into four quadrants while the right figure is the respective quadtree data structure. One builds the quadtree using Algorithm B.1.

store summary statistics of all particles within the cell as a pseudo-particle. As particles are inserted, these summaries are updated recursively. The quadtree $\mathcal{T}$ is constructed in $O(n \log n)$ time and enables efficient spatial queries and interaction approximations.

To compute interactions, the quadtree $\mathcal{T}$ is traversed starting from the root node. At each node $k \in \{0, \ldots, K\}$ where the total number of nodes $K = O(n)$, we consider its corresponding cell and pseudo-particle. Let w_k be the width of this cell, $\bar{r}_{ik}$ the Euclidean distance between the target particle $\mathbf{x}_i$ and the pseudo-particle's center of mass, and $\rho_{ik} = w_k/\bar{r}_{ik}$. If $\rho_{ik} < \theta$, where θ is a user-specified threshold, the pseudo-particle is considered distant and used as an approximation of its constituent particles. Otherwise, we recursively examine the node's children; interactions with particles in leaf nodes are computed directly. The parameter θ controls the trade-off between accuracy and computational efficiency; larger values of θ result in fewer tree traversals and faster computation at the cost of reduced accuracy. $\theta \geq 0.5$ typically yields $O(n \log n)$, while $\theta = 0$ recovers the exact $O(n^2)$ computation.

4 A tree-based framework for Bayesian multidimensional scaling

4.1 Barnes–Hut Bayesian multidimensional scaling

Given latent embeddings $\{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^2$, we construct a quadtree $\mathcal{T}$ (Algorithm B.1) where each node k corresponds to a spatial cell C_k in the latent space with index set $\mathcal{I}_k = \{i \in \{1, \ldots, n\} : \mathbf{x}_i \in \mathcal{C}_k\}$. While leaf nodes contain single observations, internal nodes store pseudo-particles summarizing the number of points ($N_k = |\mathcal{I}_k|$) and the centers of mass of the high-dimensional ($\bar{\mathbf{y}}_k = \frac{1}{N_k} \sum_{l \in \mathcal{I}_k} \mathbf{y}_l$) and low-dimensional ($\bar{\mathbf{x}}_k = \frac{1}{N_k} \sum_{l \in \mathcal{I}_k} \mathbf{x}_l$) points within $\mathcal{C}_k$. The root node ($k = 0$) summarizes the entire dataset.

Unlike t-SNE (van der Maaten, 2014), BMDS (Equation 2) depends on squared differences between high-dimensional and low-dimensional dissimilarities. To preserve the $O(n \log n)$ complexity, we approximate the high-dimensional dissimilarity $\bar{D}_{ik} = \|\mathbf{y}_i - \bar{\mathbf{y}}_k\|$ using the Euclidean distance between the observed point $\mathbf{y}_i$ and the pseudo-particle's high-dimensional center of mass $\bar{\mathbf{y}}_k$. This choice avoids storing or computing all pairwise dissimilarities which would require $O(n^2)$ memory or operations, at the cost of introducing approximation error.

During likelihood computations, we perform a depth-first traversal of quadtree $\mathcal{T}$ for each point $\mathbf{x}_i$ and compute $\bar{r}_{ik} = \|\mathbf{x}_i - \bar{\mathbf{x}}_k\|$, the Euclidean distance between the target point $\mathbf{x}_i$ and the pseudo-particle's low-dimensional center of mass $\bar{\mathbf{x}}_k$. This quantity is used within the Barnes–Hut criterion to determine tree traversal. The Barnes–Hut BMDS (BH-BMDS) framework reduces the computational burden of evaluating the BMDS log-likelihood through the use of summary statistics; dissimilarities between a target point and points within a sufficiently distant cell can be represented by their aggregate summary.

Because each unordered pairwise interaction is visited twice (one per target point), the normalization term is adjusted by a factor of two. The BH-BMDS log-likelihood is

$$\begin{aligned}\tilde{\ell}(\mathbf{D} \mid \mathbf{X}, \sigma^2, \mathcal{T}) &= -N \log(2\pi\sigma^2) - \sum_{i=1}^{n} \sum_{k \in \mathcal{B}_i} N_k \left[\frac{(\bar{D}_{ik} - \bar{r}_{ik})^2}{2\sigma^2} + \log \Phi\left(\frac{\bar{r}_{ik}}{\sigma} \right) \right], \\ &\equiv -N \log(2\pi\sigma^2) - \sum_{i=1}^{n} \tilde{\ell}(\mathbf{x}_i \mid \mathcal{T}), \end{aligned} \tag{3}$$

where $\mathcal{B}_i$ denotes the set of nodes satisfying the Barnes–Hut acceptance criterion (or leaf nodes) for target point $\mathbf{x}_i$, and $\tilde{\ell}(\mathbf{x}_i \mid \mathcal{T})$ is the pointwise contribution to the BH-BMDS log-likelihood from a Barnes–Hut traversal of $\mathbf{x}_i$ starting at the root node (Algorithm B.3). The BH-BMDS log-likelihood approximates the full BMDS log-likelihood, and when $\theta = 0$, it reduces to the exact formulation since $\mathcal{B}_i$ contains only leaf nodes.

4.2 Noisy Barnes–Hut algorithm

Under the Barnes–Hut approximation, the BMDS log-likelihood becomes a piecewise function of the latent positions due to the discrete tree structure. For a fixed set of points, small changes in the embeddings can alter traversal paths or cell assignments, leading to discontinuous changes in the BH-BMDS log-likelihood. Therefore, we introduce a probabilistic acceptance rule to smooth the transition of accepting pseudo-particles near the threshold θ, mitigating abrupt changes in the Barnes–Hut approximation as ρ_{ik} varies for small movements in $\mathbf{x}_i$.

Noisy Barnes–Hut follows the same tree traversal as the deterministic version but relaxes the hard acceptance criterion. Instead of the binary indicator $\mathbb{I}(\rho_{ik} < \theta)$, noisy Barnes–Hut treats this decision probabilistically. A pseudo-particle is accepted if $u < p(\rho_{ik} \mid \theta)$, where $u \sim \text{Unif}(0, 1)$ and the acceptance probability is defined by the logistic function

$$p(\rho \mid \theta) = \left(1 + \exp\left[-\frac{c}{\theta}(\theta - \rho) \right] \right)^{-1}. \tag{4}$$

The parameter θ is a user-specified threshold, while c is a constant controlling the slope of the logistic function. We fix c so that the transition between acceptance and rejection occurs over a narrow band of ρ_{ik} values; specifically, $p(\rho \mid \theta) = 0.8$ when $(\theta - \rho)/\theta = 0.2$. In our implementation, this corresponds to $c = 6.93$. As $c \to \infty$, the probability $p(\rho \mid \theta)$ converges to the indicator function $\mathbb{I}(\rho < \theta)$, recovering the deterministic rule. The method preserves the expected computational complexity of $O(n \log n)$ with occasional additional node evaluations near the acceptance boundary.

By treating $\mathcal{B}_i$ as a realization of a random process with controlled variance, the stochastic approach improves mixing in MCMC settings (Section 4.4). However, it introduces approximation noise that must be rigorously accounted for. We address this by proving that the stationary measure of BH-BMDS is consistent (Section 4.3).

4.3 Consistency for noisy BH-BMDS

We prove consistency for a class of surrogate-likelihood algorithms, even in the regime where the surrogate is not pointwise accurate. See Section A.1 for an expanded discussion of this theoretical framework.

Fix dimension $d \in \mathbb{N}$ and noise level $\sigma > 0$. For each $n \in \mathbb{N}$, we assume the data are generated from "true" latent positions $\mathbf{x}_1^\star, \ldots, \mathbf{x}_n^\star \overset{\text{i.i.d.}}{\sim} \text{Unif}([0, 1]^d)$, and the observed dissimilarities D_{ij} are generated according to Model 1 where $r_{ij} = r_{ij}^\star \equiv \left\| \mathbf{x}_i^\star - \mathbf{x}_j^\star \right\|$. We define the truncation function $T_M(u) = \min\{u, M\}$ for $u \geq 0$, and a maximum distance $M \geq \sqrt{d}$. We then set the capped dissimilarities and their expected values as

$$\tilde{D}_{ij} = T_M(D_{ij}), \qquad \mu_M(r) = \mathbb{E}\big[T_M([r + \sigma Z]_+)\big], \qquad r \geq 0 \tag{5}$$

such that the conditional law of $\tilde{D}_{ij}$ given $r^\star_{ij} = r$ has a piecewise density $q_{M,r}(y)$ (Equation A.2). Since the maximum possible distance within this unit hypercube is bounded by $\sqrt{d}$, the capped distances are consistent with the underlying data-generation process.

For a candidate configuration $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in ([0,1]^d)^n$ and $r_{ij}(\mathbf{X}) = \|\mathbf{x}_i - \mathbf{x}_j\|$, define the loss (negative log-likelihood) as

$$\ell_M(y, r) \;=\; -\log q_{M,r}(y), \qquad y \in [0, M],\; r \in [0, \sqrt{d}], \tag{6}$$

and the associated empirical risk as

$$R_n(\mathbf{X}) \;=\; \frac{1}{N} \sum_{1 \le i < j \le n} \ell_M(\tilde{D}_{ij},\, r_{ij}(\mathbf{X})). \tag{7}$$

We consider a family of surrogate log-likelihoods indexed by $\xi \in \Xi$, where each ξ induces an MCMC algorithm. Let $\xi_0 \in \Xi$ represent the deterministic reference algorithm, e.g., standard Barnes–Hut. By defining a probability measure on Ξ, we draw a stochastic realization ξ, yielding the randomized surrogate log-likelihood that drives our noisy MCMC. We denote $\widehat{\ell}^{(\xi)}_{ij}(\mathbf{X})$ as the resulting approximation of $\ell_M(\tilde{D}_{ij}, r_{ij}(\mathbf{X}))$ and define the surrogate risk as

$$\widehat{R}^{(\xi)}_n(\mathbf{X}) \;=\; \frac{1}{N} \sum_{1 \le i < j \le n} \widehat{\ell}^{(\xi)}_{ij}(\mathbf{X}). \tag{8}$$

Note that $\widehat{\ell}^{(\xi)}_{ij}$ depends on the entirety of $\mathbf{X}$ due to the quadtree construction.

Since distances are invariant under Euclidean isometries, e.g., rigid motion, we measure convergence using the orbit distance $d_{\mathcal{G}}$ defined in Remark A.1. Specifically, we prove that the mass of the stationary measure outside any fixed $d_{\mathcal{G}}$-ball around the true configuration $\mathbf{X}^\star$ goes to 0 (in probability) as $n \to \infty$. To achieve this, we assume a standard data-generating process characterized by a random design, independent paired noise and sufficient prior regularity around the true configuration (Assumption A.2). The key assumption is *uniform vanishing error* for the *pairwise losses*:

Assumption 1 (Small Approximation Error) *There exists a deterministic sequence $\varepsilon_n \downarrow 0$ such that, with probability $\to 1$ under the joint law of $(\mathbf{X}^\star, \mathbf{D})$, the reference approximation $\widehat{\ell}^{(\xi_0)}$ satisfies*

$$\sup_{\mathbf{X} \in ([0,1]^d)^n} \; \max_{1 \le i < j \le n} \left| \widehat{\ell}^{(\xi_0)}_{ij}(\mathbf{X}) - \ell_M(\tilde{D}_{ij}, r_{ij}(\mathbf{X})) \right| \;\le\; \varepsilon_n. \tag{9}$$

While the difference in the pairwise approximation error must go to 0 as n grows, this assumption allows the difference in the total log-likelihood to diverge to infinity, provided the growth remains sub-quadratic.

Theorem 1 (Consistency for noisy MCMC) *Fix $d \ge 1$, $\sigma > 0$, $M \ge \sqrt{d}$ and a parameter space Ξ. Suppose there exists a reference parameter $\xi_0 \in \Xi$ such that the chain P_{ξ_0} is uniformly ergodic and let $\widehat{\ell}^{(\xi_0)}_{ij}$ be an approximation satisfying Assumption 1. Let P_ξ be a Markov chain with stationary measure $\widehat{\Pi}^{(\xi)}_n$, and define the mixture kernel $P = \int_\Xi P_\xi \, \mu(d\xi)$ for a probability measure μ on Ξ. Let $\widetilde{\Pi}_n$ be an invariant probability measure of P and set κ_n and Δ_n as Equations A.10 and A.11, respectively. If $\kappa_n \Delta_n \xrightarrow{\mathbb{P}} 0$, then for every $\delta > 0$,*

$$\widetilde{\Pi}_n\Big(\big\{\mathbf{X} :\; d_{\mathcal{G}}(\mathbf{X}, \mathbf{X}^\star) > \delta\big\}\Big) \xrightarrow{\mathbb{P}} 0. \tag{10}$$

Remark 1 *Theorem 1 applies directly to the deterministic Barnes–Hut algorithm ($\xi = \xi_0$, Section 3.2, Algorithm B.3) by choosing a sequence $\theta = \theta_n \to 0$. More subtly, the result also extend to the noisy quadtree algorithm (Section 4.2, Algorithm B.4) for sufficiently small noise. Because the target distribution is supported on a compact state space and possesses a density strictly bounded away from both zero and infinity, utilizing any standard proposal distribution with a similarly bounded density ensures that the reference Metropolis-Hastings chain is uniformly ergodic. Furthermore, the one-step total variation distance Δ_n is governed by the logistic slope c in the noisy traversal (Equation 4). By letting $c \to \infty$, Δ_n can be made arbitrarily small, fulfilling the final theoretical assumptions.*

Proof Sketch The full proof, detailed in Appendix A, proceeds in two main stages. First, we establish consistency for the deterministic reference (ξ_0). We demonstrate that the baseline empirical risk uniformly concentrates around the expected "population" risk. Because this population risk is strictly minimized at the true configuration $\mathbf{X}^\star$ (up to isometry), the mass of the stationary measure must concentrate around $\mathbf{X}^\star$ in orbit distance. Second, we apply perturbation arguments to show this consistency is preserved under the noisy quadtree algorithm.

4.4 Bayesian computation

To infer the model parameters, we implement a Metropolis-within-Gibbs (MwG) sampler (Algorithm 1) targeting the posterior distribution

$$p(\mathbf{X}, \sigma^2 \mid \mathcal{T}) \propto \tilde{\ell}(\mathbf{D} \mid \mathbf{X}, \sigma^2, \mathcal{T})\pi(\mathbf{X})\,\pi(\sigma^2), \tag{11}$$

where $\pi(\cdot)$ is the prior specification.

The sampler proceeds by sequentially updating each model parameter $\gamma = (\mathbf{x}_1, \ldots, \mathbf{x}_n, \sigma^2)$ using Metropolis–Hastings (MH) steps within a Gibbs framework. At iteration s and for a generic parameter γ_k with proposal $\gamma_k' \sim q(\cdot \mid \gamma_k^{(s)})$, the acceptance probability is

$$\alpha(\gamma_k' \mid \gamma^{(s)}, \mathcal{T}) = \min\left\{1, \frac{p(\gamma_k', \gamma_{-k}^{(s)} \mid \mathcal{T})\; q(\gamma_k^{(s)} \mid \gamma_k')}{p(\gamma_k^{(s)}, \gamma_{-k}^{(s)} \mid \mathcal{T})\; q(\gamma_k' \mid \gamma_k^{(s)})}\right\}, \tag{12}$$

where γ_{-k} denotes all parameters except γ_k. We propose a candidate latent location from $\mathbf{x}_n' \sim N\big(\mathbf{x}_n^{(s)}, \tau_n^{2(s)}\mathbf{I}\big)$ and a candidate precision from a truncated normal proposal, $1/\sigma^{2\prime} \sim N_{(0,\infty)}(1/\sigma^{2(s)}, \tau_{1/\sigma^2}^{2(s)})$.

Algorithm 1: BH-BMDS via Metropolis-within-Gibbs

Input: $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \mathbb{R}^\Omega$
Output: posterior samples for $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^2$
Initialize $\mathbf{X}^0 = \{\mathbf{x}_i^0\}_{i=1}^n \subset \mathbb{R}^2$ via dimension reduction (e.g., MDS or PCA);
Build quadtree $\mathcal{T}$ on $\mathbf{X}^0$;
for *each MCMC iteration* **do**
  Permute indices $1, \ldots, n$;
  for $i = 1, \ldots, n$ **do**
    Compute current $\tilde{\ell}(\mathbf{x}_i \mid \mathcal{T})$ using Algorithm B.4;
    Propose $\mathbf{x}_i' \sim N(\mathbf{x}_i, \tau_i^2\mathbf{I})$;
    Update $\mathcal{T}$ locally using Algorithm B.2;
    Compute proposed $\tilde{\ell}(\mathbf{x}_i' \mid \mathcal{T})$ using Algorithm B.4;
    Accept/reject via MH criterion (Equation 12);
    if *rejected* **then**
      Restore cached node summaries in $\mathcal{T}$;
    else
      Keep updated $\mathcal{T}$;
  Compute $\tilde{\ell}(\mathbf{D} \mid \mathbf{X}, \sigma^2, \mathcal{T})$ using Algorithm B.3;
  Propose $1/\sigma^{2\prime} \sim N_{(0,\infty)}(1/\sigma^{2(s)}, \tau_{1/\sigma^2}^{2(s)})$;
  Compute $\tilde{\ell}(\mathbf{D} \mid \mathbf{X}, \sigma^{2\prime}, \mathcal{T})$ using Algorithm B.3;
  Accept/reject via MH criterion (Equation 12);
  Rebuild $\mathcal{T}$

To efficiently evaluate the posterior, we implement different Barnes–Hut approximations depending on the parameter being updated. Latent locations use Algorithm B.4 which evaluates a pointwise contribution to the BH-BMDS likelihood via a stochastic tree traversal. Noisy Barnes–Hut mitigates discontinuities in the likelihood approximation that arise when small changes in $\mathbf{x}_i$ cause crossings of the deterministic Barnes–Hut acceptance boundary $\mathbb{I}(\rho_{ik} < \theta)$. By replacing this hard threshold with a probabilistic acceptance rule (Equation 4), the resulting likelihood approximation varies more smoothly as a function of $\mathbf{x}_i$ which increases the stability of Metropolis-Hastings updates. In contrast, the variance parameter σ^2 uses deterministic Barnes–Hut to approximate the full BMDS likelihood (Algorithm B.3). This strategy avoids introducing additional noise into likelihood evaluations, thereby improving numerical stability.

Under this MwG sampler, for fixed quadtree $\mathcal{T}$, a single latent location update requires $O(\log n)$ expected time, and a full sweep over all locations costs $O(n \log n)$. We assume proposals are sufficiently local such that $\mathbf{x}_i$ and $\mathbf{x}_i'$ lie in the same quadtree path, so that only center-of-mass updates are required. While running experiments on data (Section 5), we found that this assumption holds true, and within a sweep, points' cell allocations do not change. At the end of each sweep, the quadtree $\mathcal{T}$ is rebuilt using the updated latent locations, which also requires $O(n \log n)$ time. The cost of updating σ^2 is of the same order as a full sweep.

5 Experiments

We conduct experiments on different datasets to evaluate the performance of Barnes–Hut BMDS. All algorithmic results arise from single-core CPU implementations carried out on a single Linux machine with two 26-core Intel Xeon Gold processors (2.1 GHz). Tree related data structures and likelihood computations are

implemented in `C` while MCMC samplers are implemented in `Python`. For visualization, we use the `ggplot2` (Wickham, 2016) package in `R` (R Core Team, 2023).

5.1 Data

We describe each of the four datasets that span structured mobility data, text embeddings, image representations and neural activity recordings, allowing us to evaluate BH-BMDS across diverse modalities. We also explain our preprocessing step for defining high-dimensional inputs which are mapped to two-dimensional latent representations for BH-BMDS. Dimensions are chosen to balance information retention and computational efficiency.

Influenza This dataset consists of air traffic-dervied pairwise "effective distances" between n = 5,392 taxa. The taxa belong to four influenza subtypes: 1,370, 1,389, 1,393 and 1,240 observations of H1N1, H3N1, Victoria and Yamagata, respectively. Effective distances inversely measure the probability of traveling between airports, such that shorter distances correspond to more heavily trafficked routes. Compared to geometric distances, effective distances better capture the structure of the global mobility network and are therefore more appropriate for modeling disease spread (Brockmann & Helbing, 2013). Since multiple taxa are associated with the same country, the dataset exhibits strong clustering, with only 66 unique inter-country distances. As preprocessing, we apply classical MDS to the dissimilarity matrix to obtain Euclidean embeddings $\mathbf{y}_i \in \mathbb{R}^6$.

ArXiv This dataset consists of $n = 10{,}000$ abstracts collected from arXiv, each labeled with one of five subject areas: math (`math`), astrophysics (`astro-ph`), optics (`physics.optics`), quantum physics (`quant-ph`) and computer science / machine learning (`cs/ml`). Each abstract is embedded into 768-dimensional vector $\mathbf{w}_i$ using `SentenceTransformers` (Reimers & Gurevych, 2019) under the `all-mpnet-base-v2` model (Song et al., 2020) in `Python`. This large language model captures semantic similarity at the sentence level. We compute the pairwise cosine dissimilarities, e.g., $1 - \cos(\mathbf{w}_i, \mathbf{w}_j) = 1 - \frac{\mathbf{w}_i \cdot \mathbf{w}_j}{\|\mathbf{w}_i\|\|\mathbf{w}_j\|}$ for $\mathbf{w}_i \in \mathbb{R}^{768}$ between embeddings and then apply classical MDS to obtain Euclidean embeddings $\mathbf{y}_i \in \mathbb{R}^{50}$.

MNIST This dataset is a random subset $n = 10{,}000$ of grayscale handwritten digit images from the MNIST dataset, where each image has dimension $d = 28 \times 28 = 784$ with pixel intensity in $[0, 1]$ and labels from 0 to 9. To extract meaningful features, we employ a convolutional autoencoder that maps each image to a 30-dimensional latent representation $\mathbf{y}_i \in \mathbb{R}^{30}$. The model is trained to minimize pixel-wise reconstruction loss using the Adam optimizer in `PyTorch` (Paszke et al., 2019). The resulting 30-dimensional embeddings serve as the high-dimensional input for BH-BMDS.

Neural activity Data come from experiments reported in Zwang et al. (2025). This dataset comprises local field potential (LFP) recordings from 32-channel probes implanted in the CA1 hippocampus and medial entorhinal cortex of healthy and ThyTau22 mice. Recordings were captured over six months during a goal-oriented virtual reality task that lasted 15 minutes. We employ a pre-trained MaGNet GNN (Zhou et al., 2025)–trained to decode behavioral states (stalled vs. moving)–to generate high-dimensional latent embeddings $\mathbf{y}_i \in \mathbb{R}^{51}$, independently across mouse types and weeks. We use BH-BMDS to project these embeddings into a low-dimensional space to evaluate manifold interpretability and identify progressive signs of dementia.

5.2 Experimental setup

For our simulations, we assume that our posterior distribution (Equation 11) is specified with priors $\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{\text{i.i.d.}}{\sim} N(\mathbf{0}, \mathbf{I}_2)$ and $1/\sigma^2 \sim \text{Gamma}(1, 1)$, where $1/\sigma^2$ is the BMDS precision parameter. The proposal variances $\tau_i^{2(s)}$ and $\tau_{\sigma^{-2}}^{2(s)}$ are adaptively tuned to maintain target acceptance rates, thereby facilitating efficient exploration of the posterior. Let $\hat{\alpha}$ denote the empirical acceptance rate over a fixed adaptation window and α the target acceptance rate. Define $R = \hat{\alpha}/\alpha$, truncated to lie in $[0.5, 2]$. After each adaptation window, the proposal variances are updated via $\tau^{(s+1)} = R\tau^{(s)}$. Adaptation is performed

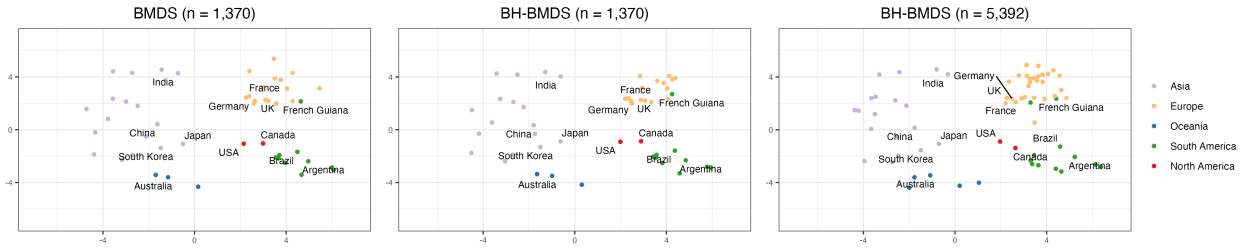

Figure 2: Posterior Procrustes-aligned means across MCMC iterations and taxa for only H1N1 ($n = 1{,}370$) and all influenza subtypes ($n = 5{,}392$) under full Bayesian multidimensional scaling (BMDS) and Barnes–Hut Bayesian multidimensional scaling (BH-BMDS) frameworks.

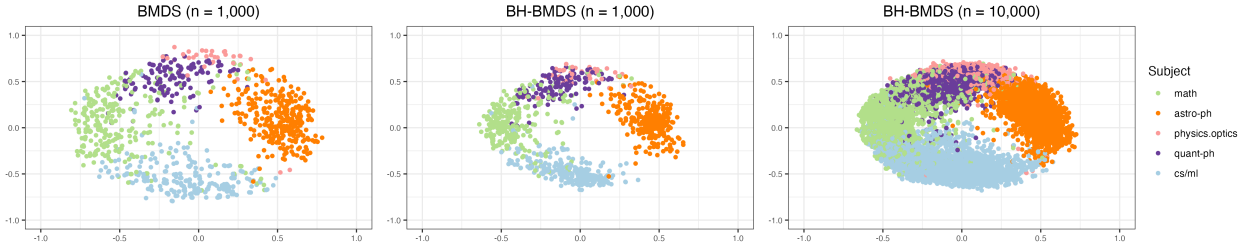

Figure 3: Posterior Procrustes-aligned means across MCMC iterations for 1,000 and 10,000 embedded ArXiv abstracts under full Bayesian multidimensional scaling (BMDS) and Barnes–Hut Bayesian multidimensional scaling (BH-BMDS) frameworks. Figures C.1 and C.2 show analogous results for the MNIST and neural activity datasets.

every 50 MCMC iterations. We set $\theta = 2$ for all experiments as it empirically provides a favorable trade-off between accuracy and computational efficiency.

5.3 Accuracy

Building on the theoretical guarantees for consistent posterior behavior (Theorem 1), we empirically evaluate BH-BMDS across four datasets. We visualize the posterior Procrustes-aligned mean embeddings using every 50th MCMC iteration. Across all benchmarks (Figures 2–C.2), BH-BMDS closely recovers the clustering structure of the exact BMDS, with shared labels consistently co-located. We observe minor coordinate shrinkage toward the origin, a known effect of Gaussian priors when the likelihood approximation introduces additional variance. Notably, exact BMDS results are only provided for smaller subsets where computation is feasible; for larger n, exact inference is prohibitive. These results demonstrate that BH-BMDS provides a faithful, scalable approximation to the full Bayesian formulation.

5.4 Computational Efficiency

We evaluate efficiency using effective sample size (ESS) per hour, computed via the `coda` (Plummer et al., 2006) package in `R`. ESS inversely reflects the degree of autocorrelation in MCMC samples, with higher values indicating more efficient exploration of the posterior. We calculate ESS using thinned MCMC samples (retaining every 50th iteration) on a random subset of 1,000 pairwise latent distances to avoid identifiability issues inherent in raw locations.

Figure 4 compares the computational efficiency across three inference frameworks: Hamiltonian Monte Carlo (HMC) and MwG for exact BMDS and Algorithm 1 for BH-BMDS and five datasets: H1N1, ArXiv, MNIST and the ThyTau22 and Wild-Type mouse at Week 9. HMC efficiently explores the posterior by leveraging gradient information but incurs additional computational cost due to gradient evaluations (Neal, 2012). In our implementation, HMC jointly samples latent locations under the full BMDS likelihood using 20 leapfrog steps. BMDS and BH-BMDS follow Algorithm 1 with $\theta = 0$ and $\theta = 2$, respectively. For BH-BMDS,

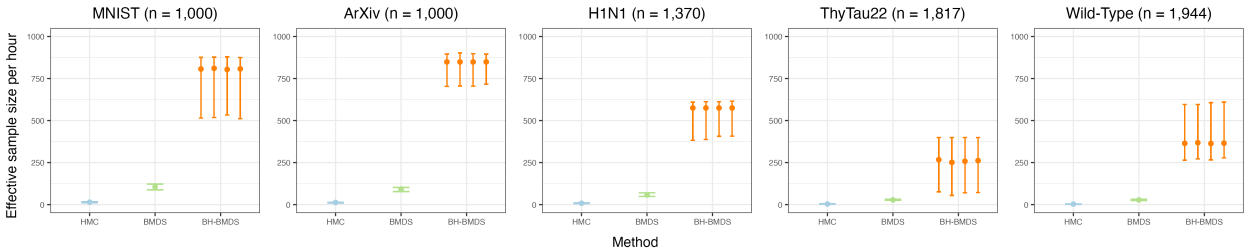

Figure 4: 95% empirical intervals of computational efficiency for H1N1, ArXiv, MNIST and ThyTau22 and Wild-Type cohorts at Week 9 datasets. Computational efficiency is measured as the effective sample size (ESS) per hour, in which ESS is a function of asymptotic auto-correlation. We compare efficiency across three frameworks: 1) BMDS log-likelihood and gradient using HMC (HMC), 2) BMDS log-likelihood using MwG (BMDS) and 3) BH-BMDS log-likelihood using MwG (BH-BMDS, Algorithm 1). ThyTau22 has a lower ESS per hour compared to Wild-Type for similar number of points due to weaker clustering.

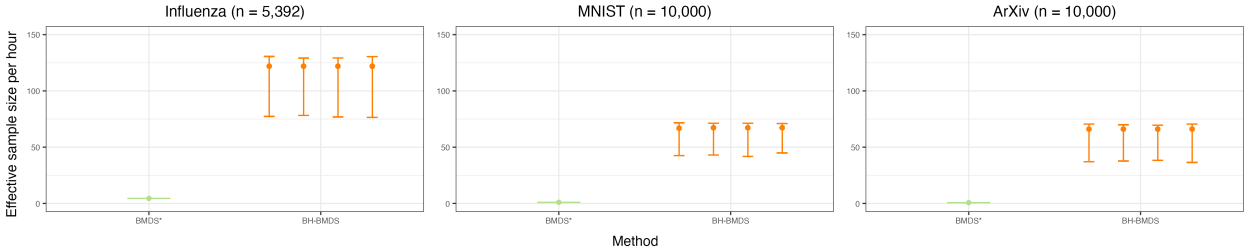

Figure 5: 95% empirical intervals of computational efficiency for influenza, ArXiv and MNIST datasets. Computational efficiency is measured as the effective sample size (ESS) per hour, in which ESS is a function of asymptotic auto-correlation. Since efficiency would be infeasible to compute for such large data under BMDS, we provide a theoretical value. Time at $n = 10{,}000$ is predicted from a quadratic regression using observed times at $n = 100, 500, 1{,}000, 5{,}000$ while ESS is generously assumed to be perfect. BMDS* equals 4.5, 1.0 and 0.7 for influenza, MNIST and ArXiv, respectively.

we repeat each experiment four times to account for variability induced by the quadtree-based likelihood approximation. As shown in Figure 4, BH-BMDS achieves approximately 60-fold speedup relative to HMC and 9-fold speedup relative to BMDS. Gains are most pronounced in datasets with distinct clustering; for instance, the ThyTau22 cohort shows slightly lower relative efficiency than Wild-Type due to more diffuse posterior mass (Figure C.2).

As exact BMDS is infeasible at scale, we extrapolate its performance to the full datasets, generously assuming idealized sampling efficiency by setting $ESS = S$ for S retained (thinned) samples. For the influenza dataset, we use observed BMDS runtimes; for the ArXiv and MNIST datasets, we estimate BMDS runtime at $n = 10{,}000$ via a quadratic regression on benchmarked runtimes ($n \in \{100, 500, 1{,}000 \text{ and } 5{,}000\}$). Even with this optimistic assumption, BH-BMDS achieves substantial gains: approximately 27-fold speedup for influenza, 67-fold for MNIST and 93-fold for ArXiv (Figure 5). These gains increase with n, consistent with the improved scaling behavior of the BH-BMDS approximation.

6 Discussion

We propose a scalable approximation to the BMDS likelihood using Barnes–Hut algorithms (BH-BMDS), reducing computational complexity from $O(n^2)$ to $O(n \log n)$. By embedding this approximation within a Metropolis-within-Gibbs sampler, our approach enables efficient posterior inference while maintaining a probabilistic framework. Crucially, we use a novel approach (Section A.1) to establish consistency of the stationary measure for BH-BMDS and evaluate the proposed method on datasets with diverse clustering patterns.

Several limitations suggest directions for future work. First, our inference scheme relies on coordinate-wise updates of latent locations, which may mix slowly in settings with strong dependence across points. We explored joint updates using HMC, discontinuous HMC and Metropolis–Hastings but encountered difficulties achieving both stable convergence and favorable scaling for larger n. More broadly, these findings highlight the challenges of combining discrete approximation schemes with MCMC methods. Second, the noisy Barnes–Hut approximation (Section 4.2) introduces stochasticity through the traversal criterion. While this improves mixing by mitigating abrupt changes in the approximated likelihood, it also complicates the theoretical scaling behavior and may deviate from ideal $O(n \log n)$ performance, although such deviations were not observed in practice. Developing more stable approximation schemes that preserve both accuracy and computational guarantees is an important direction for future work. Finally, as with standard BMDS, low-dimensional embeddings may not capture all relevant structure in complex datasets, particularly when restricted to two dimensions. Extending BH-BMDS to higher-dimensional embeddings or adaptive representations is a promising avenue for future work.

References

Pierre Alquier, Nial Friel, Richard Everitt, and Aidan Boland. Noisy monte carlo: convergence of markov chains with approximate transition kernels. *Statistics and Computing*, 26:29–47, 2016.

Ery Arias-Castro, Adel Javanmard, and Bruno Pelletier. Perturbation bounds for Procrustes, classical scaling, and trilateration, with applications to manifold learning. *Journal of Machine Learning Research*, 21(15):1–37, 2020.

Josh Barnes and Piet Hut. A hierarchical o(nlogn) force-calculation algorithm. *Nature*, 324:446–449, 1986.

Trevor Bedford, Steven Riley, Ian G. Barr, Shobha Broor, Mandeep Chadha, Nancy J. Cox, Rodney S. Daniels, C. Palani Gunasekara, Aeron C. Hurt, Anne Kelso, Alexander Klimov, Nicola S. Lewis, Xiyan Li, John W. McCauley, Takato Odagiri, Varsha Potdar, Andrew Rambaut, Yuelong Shu, Eugene Skepner, Derek J. Smith, Marc A. Suchard, Masato Tashiro, Dayan Wang, Xiyan Xu, Philippe Lemey, and Colin A. Russell. Global circulation patterns of seasonal influenza virsuses vary with antigentic drift. *Nature*, 523: 217–220, 2015.

Sebastian Bieringer, Gregor Kasieczka, Maximilian F. Steffen, and Mathias Trabs. The surrogate Gibbs-posterior of a corrected stochastic MALA: Towards uncertainty quantification for neural networks. *Journal of Machine Learning Research*, 27(1):1–50, 2026. URL `https://www.jmlr.org/papers/v27/22-0483.html`.

Dirk Brockmann and Dirk Helbing. The hidden geometry of complex, network-driven contagion phenomena. *Science*, 342(6164):1337–1342, 2013.

Andrew J. Holbrook, Philippe Lemey, Guy Baele, Simon Dellicour, Dirk Brockmann, Andrew Rambaut, and Marc A. Suchard. Massive parallelization boosts big bayesian multidimensional scaling. *Journal of Computational and Graphical Statistics*, 30(1):11–24, 2021.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2013.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 1933.

YQ Li, M Ghafari, AJ Holbrook, I Boonen, N Amor, S Catalano, JP Webster, YY Li, HT Li, V Vergote, P Maes, YL Chong, A Laudisoit, P Baelo, S Ngoy, SG Mbalitini, GC Gembu, AP Musaba, J Goüy de Bellocq, H Leirs, E Verheyen, OG Pybus, A Katzourakis, AN Alagaili, S Gryseels, YC Li, MA Suchard, M Bletsa, and P Lemey. The evolutionary history of hepaciviruses. *bioRxiv : the preprint server for biology*, 2023.

Lin Lin and Duncan K.H. Fong. Bayesian multidimensional scaling procedure with variable selection. *Computational Statistics & Data Analysis*, 129:1–13, 2019.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.

F. J. Medina-Aguayo, Anthony Lee, and Gareth O. Roberts. Stability of noisy Metropolis–Hastings. *Statistics and Computing*, 26:1187–1211, 2016.

A. Yu. Mitrophanov. Sensitivity and convergence of uniformly ergodic markov chains. *Journal of Applied Probability*, 42(4):1003–1014, 2005.

Radford M. Neal. *MCMC using Hamiltonian dynamics*, chapter 5. Campman & Hall, 2012.

Man-Suk Oh and Adrian E. Raftery. Bayesian multidimensional scaling and choice of dimension. *Journal of the American Statistical Association*, 96(455):1031–1044, 2001.

Man-Suk Oh and Adrian E. Raftery. Model-based clustering with dissimilarities: A bayesian approach. *Journal of Computational and Graphical Statistics*, 16(3):559–585, 2007.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. URL `https://arxiv.org/abs/1912.01703`.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11, 2006. URL `https://journal.r-project.org/archive/`.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL `https://www.R-project.org/`.

Riccardo Rastelli, Florian Maire, and Nial Friel. Computationally efficient inference for latent position network models. *Electronic Journal of Statistics*, 18(1):2531 – 2570, 2024. doi: 10.1214/24-EJS2256. URL `https://doi.org/10.1214/24-EJS2256`.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL `https://arxiv.org/abs/1908.10084`.

Daniel Rudolf and Nikolaus Schweizer. Perturbation theory for markov chains via Wasserstein distance. *Bernoulli*, 24(4A):2610–2639, 2018.

Ami Sheth, Aaron Smith, and Andrew J. Holbrook. Sparse bayesian multidimensional scaling(s). *Computational Statistics*, 2026.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*, 2020.

Yee Whye Teh, Alexandre H. Thiery, and Sebastian J. Vollmer. Consistency and fluctuations for stochastic gradient langevin dynamics. *Journal of Machine Learning Research*, 17(7):1–33, 2016.

Warren S. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, 1952.

Laurens van der Maaten. Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, 15:1–21, 2014.

Laurens van der Maaten and Geoffery Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Roman Vershynin. High-dimensional probability: An introduction with applications in data science, 2025. Second edition. Prepublication PDF dated Nov. 28, 2025.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `https://ggplot2.tidyverse.org`.

Wenzhuo Zhou, Annie Qu, Keiland W. Cooper, Norbert Fortin, and Babak Shahbaba. A model-agnostic graph neural network for integrating local and global information. *Journal of the American Statistical Association*, 120(550):1225–1238, 2025.

Theodore J Zwang, Jasen Zhang, Rudy Gelb-Bicknell, Nina Wolf, Viktor Chanchykov, Haoyang Zhu, Ergina Agastra, Ava Devine, Andrew J. Holbrook, Rachel E. Bennett, and Bradley T. Hyman. Tau drives cell specific functional isolation of the hippocampal formation. *bioRxiv*, 2025. doi: 10.1101/2025.08.10.669580.

A Consistency of approximate BMDS

We restate and then prove Theorem 1 from the main document.

We begin by recalling some basic notation from Section 4.3. Throughout Appendix A, fix $d \geq 1$, $\sigma > 0$, and $M \geq \sqrt{d}$. We use the shorthand $N = \binom{n}{2}$, $\mathbf{J} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^\top$, and

$$r_{ij}^\star = \left\|\mathbf{x}_i^\star - \mathbf{x}_j^\star\right\|, \qquad r_{ij}(\mathbf{X}) = \|\mathbf{x}_i - \mathbf{x}_j\|.$$

Our overall goal is to infer the latent positions $\mathbf{x}_1^\star, \mathbf{x}_2^\star, \ldots, \mathbf{x}_n^\star \in [0,1]^d$, which we write in matrix form as $\mathbf{X}^\star$ given the observed dissimilarity matrix $\tilde{D}_{ij} \in [0, M]$. For any matrix $\mathbf{X}$ or the actual latent positions $\mathbf{X}^\star$, define the centered configurations $\mathbf{X}_c = \mathbf{J}\mathbf{X}$ and $\mathbf{X}_c^\star = \mathbf{J}\mathbf{X}^\star$.

Relative to the reference measure

$$\nu = \delta_0 + \lambda|_{(0,M)} + \delta_M \tag{A.1}$$

on $[0, M]$, the conditional law of $\tilde{D}_{ij}$ given $r_{ij}^\star = r$ has density

$$q_{M,r}(y) = \begin{cases} \Phi(-r/\sigma), & y = 0, \\ \dfrac{1}{\sigma}\phi\left(\dfrac{y-r}{\sigma}\right), & 0 < y < M, \\ 1 - \Phi\left(\dfrac{M-r}{\sigma}\right), & y = M, \end{cases} \tag{A.2}$$

where λ is the Lebesgue measure and ϕ, Φ denote the standard normal density and CDF, respectively.

We also recall the following notation from Section 4.3: the loss $\ell_M(y, r) = -\log q_{M,r}(y)$ (Equation 6), the empirical risk $R_n(\mathbf{X}) = N^{-1}\sum_{i<j} \ell_M(\tilde{D}_{ij}, r_{ij}(\mathbf{X}))$ (Equation 7), and for an algorithmic realization ξ, the surrogate risk $\widehat{R}_n^{(\xi)}(\mathbf{X}) = N^{-1}\sum_{i<j} \widehat{\ell}_{ij}^{(\xi)}(\mathbf{X})$ (Equation 8). These yield the exact posterior

$$\Pi_n(d\mathbf{X} \mid \mathbf{D}) \propto \Pi_0(d\mathbf{X}) \exp\{-NR_n(\mathbf{X})\} \tag{A.3}$$

and the surrogate posterior

$$\widehat{\Pi}_n^{(\xi)}(d\mathbf{X} \mid \mathbf{D}) \propto \Pi_0(d\mathbf{X}) \exp\{-N\widehat{R}_n^{(\xi)}(\mathbf{X})\}. \tag{A.4}$$

Since distances are invariant under Euclidean isometries, the posterior distribution cannot possibly be consistent under the Euclidean metric. Instead, we prove consistency in the following "orbit distance" obtained by taking the Frobenius distance $\|\mathbf{A}\|_F = (\sum_{i,j} A_{ij}^2)^{1/2}$ and "modding out" by isometries:

Definition A.1 (Orbital distance) *Let $\mathcal{G} = \{(\mathbf{Q}, \mathbf{b}) : \mathbf{Q} \in \mathcal{O}(d), \mathbf{b} \in \mathbb{R}^d\}$ act on configurations via $(\mathbf{Q}, \mathbf{b}) \cdot \mathbf{x}_i = \mathbf{Q}\mathbf{x}_i + \mathbf{b}$. Define the orbit distance*

$$d_{\mathcal{G}}(\mathbf{X}, \mathbf{X}^\star) \;=\; \inf_{(\mathbf{Q},\mathbf{b})\in\mathcal{G}} \frac{1}{\sqrt{n}} \left\|\mathbf{X} - (\mathbf{X}^\star\mathbf{Q} + \mathbf{1}\mathbf{b}^\top)\right\|_F \;=\; \inf_{\mathbf{Q}\in\mathcal{O}(d)} \frac{1}{\sqrt{n}} \left\|\mathbf{X}_c - \mathbf{X}_c^\star\mathbf{Q}\right\|_F . \tag{A.5}$$

Before proving our main result (Theorem 1), we will first establish posterior consistency for the baseline algorithm with parameter $\xi = \xi_0$. For notational simplicity, we drop the algorithmic superscript until Section A.7 and let $\widehat{\ell}_{ij} \equiv \widehat{\ell}_{ij}^{(\xi_0)}$, $\widehat{R}_n \equiv \widehat{R}_n^{(\xi_0)}$, and $\widehat{\Pi}_n \equiv \widehat{\Pi}_n^{(\xi_0)}$ denote the surrogate quantities under the baseline.

The key assumption is that the surrogate likelihoods have errors going to 0:

Assumption A.1 (Small Approximation Error) *There exists a deterministic sequence $\varepsilon_n \downarrow 0$ such that, with probability $\to 1$ under the joint law of $(\mathbf{X}^\star, \mathbf{D})$,*

$$\sup_{\mathbf{X}\in([0,1]^d)^n} \max_{1\le i<j\le n} \left|\widehat{\ell}_{ij}(\mathbf{X}) - \ell_M(\tilde{D}_{ij}, r_{ij}(\mathbf{X}))\right| \;\le\; \varepsilon_n. \tag{A.6}$$

Note that, on the high-probability event (Equation A.6), the error in the risk is also small:

$$\sup_{\mathbf{X}\in([0,1]^d)^n} \left|\widehat{R}_n(\mathbf{X}) - R_n(\mathbf{X})\right| \le \varepsilon_n. \tag{A.7}$$

We also restate the assumptions about the model:

Assumption A.2 (Nice Data-Generating Process) *The data generating process and prior satisfy:*

(T1) ***(Random design)*** $\mathbf{x}_1^\star, \ldots, \mathbf{x}_n^\star \overset{i.i.d.}{\sim} \mathrm{Unif}([0,1]^d)$.

(T2) ***(Independent pair noise)*** *Conditional on $\mathbf{X}^\star$, the noises $\{Z_{ij}\}_{i<j}$ are i.i.d. $N(0,1)$.*

(T3) ***(Prior regularity)*** *The prior Π_0 is supported on $([0,1]^d)^n$ and has density π_0 satisfying $e^{-cn} \le \pi_0(\mathbf{X}) \le e^{cn}$ on $([0,1]^d)^n$, for some constant $c > 0$.*

For the remainder of Appendix A, all probabilistic statements involving the data-generating law are understood under Assumption A.2, unless otherwise stated. Statements about the baseline surrogate posterior are understood in addition under Assumption A.1 or the high-probability event (Equation A.7), as indicated.

Proposition A.1 (Consistency of the Baseline Algorithm) *Fix $d \ge 1$, $\sigma > 0$, and $M \ge \sqrt{d}$. Consider Model (Equation 1) where $r_{ij} = r_{ij}^\star$, the exact posterior (Equation A.3) based on ℓ_M in Equation 6, and the approximate posterior $\widehat{\Pi}_n$ (Equation A.4). Suppose Assumptions A.1 and A.2 hold. Then, for every $\delta > 0$,*

$$\widehat{\Pi}_n\Big(\big\{\mathbf{X} :\; d_{\mathcal{G}}(\mathbf{X}, \mathbf{X}^\star) > \delta\big\} \;\Big|\; \mathbf{D}\Big) \;\xrightarrow{\mathbb{P}}\; 0. \tag{A.8}$$

In particular, the approximate posterior of the baseline algorithm is consistent for the latent configuration up to rigid motion.

We restate Theorem 1 slightly more formally:

Theorem A.1 (Theorem 1, Extended) *Fix $d \ge 1$, $\sigma > 0$, $M \ge \sqrt{d}$ and a parameter space Ξ. Suppose Assumption A.2 holds, and suppose there exists a reference parameter $\xi_0 \in \Xi$ such that the chain P_{ξ_0} on $([0,1]^d)^n$ is uniformly ergodic and $\widehat{\ell}_{ij}^{(\xi_0)}$ satisfies Assumption A.1. For each $\xi \in \Xi$, let P_ξ be a Markov chain*

on $([0,1]^d)^n$ *with stationary measure* $\widehat{\Pi}_n^{(\xi)}$*, and define the mixture kernel* $P = \int_\Xi P_\xi \, \mu(d\xi)$ *for a probability measure* μ *on* Ξ*. Let* $\widetilde{\Pi}_n$ *be an invariant probability measure of* P*.*

Suppose that there exist deterministic constants $C_n < \infty$ *and* $\rho_n \in (0,1)$ *such that*

$$\sup_{\mathbf{X}_0 \in ([0,1]^d)^n} \left\| \delta_{\mathbf{X}_0} P_{\xi_0}^m - \widehat{\Pi}_n^{(\xi_0)} \right\|_{\mathrm{TV}} \leq C_n \rho_n^m, \qquad m \geq 0. \tag{A.9}$$

Set

$$\lambda_n = \max\left\{0, \left\lceil \frac{\log(1/C_n)}{\log(\rho_n)} \right\rceil \right\}, \qquad \kappa_n = \lambda_n + \frac{C_n \rho_n^{\lambda_n}}{1-\rho_n}, \tag{A.10}$$

and assume that the possibly data-dependent quantity

$$\Delta_n = \sup_{\xi \in \Xi} \sup_{\mathbf{X}_0 \in ([0,1]^d)^n} \left\| P_\xi(\mathbf{X}_0, \cdot) - P_{\xi_0}(\mathbf{X}_0, \cdot) \right\|_{\mathrm{TV}} \tag{A.11}$$

satisfies $\kappa_n \Delta_n \xrightarrow{\mathbb{P}} 0$*.*

Then for every $\delta > 0$*,*

$$\widetilde{\Pi}_n\Big(\big\{\mathbf{X} : \; d_{\mathcal{G}}(\mathbf{X}, \mathbf{X}^\star) > \delta\big\}\Big) \xrightarrow{\mathbb{P}} 0. \tag{A.12}$$

A.1 Proof strategy for Proposition A.1

The basic idea is:

1. Check that the empirical risk R_n concentrates near an associated "population" risk R_n^0 (Section A.3),
2. Check that the population risk $R_n^0(\mathbf{X})$ grows quadratically in the *orbit* (not Euclidean) distance to the truth $\mathbf{X}^\star$, with reasonable constant (Section A.4),
3. Check that the posterior associated with the population risk is consistent (Section A.5).

We note that this approach is not the most common approach to analyzing the error of "approximate" or "surrogate" MCMC algorithms. Typically, one aims to have a surrogate that is pointwise close to the target log-likelihood, then use perturbation theory to show that the stationary measure is close to the posterior distribution of interest (e.g., Rastelli et al. (2024) for a typical analysis, applied to a different MCMC algorithm for inferring latent positions). In the present paper, we allow the surrogate to be very far from the log-likelihood pointwise, as long as the error grows only sub-quadratically in the number of points n.

We are aware of several other results that prove consistency of MCMC algorithms even as the pointwise error is allowed to grow, though only in the much simpler situation of i.i.d. samples and a finite-dimensional set of parameters to be learned. The influential analysis of SGLD in Teh et al. (2016) is an early example, and this viewpoint is still being refined through works such as Bieringer et al. (2026).

We emphasize the following differences between our analysis and most or all previous work:

1. Directly analyzing the risk is always more difficult than merely treating an approximate MCMC chain as a perturbation of an ideal chain, where more generic results are available (Mitrophanov, 2005; Alquier et al., 2016; Medina-Aguayo et al., 2016; Rudolf & Schweizer, 2018).
2. The current problem is effectively infinite-dimensional, since the number of latent positions to learn grows with the size of the dataset. For this reason, naive bounds based on, essentially, Taylor expansions and concentration do not work. Compare this to the more classical finite-dimensional analyses of stochastic-gradient MCMC in Teh et al. (2016).
3. The ground truth is only identifiable up to an isometry, so all estimates need to effectively "mod out" by this isometry.

A.2 A standard matrix inequality

Lemma 1 *Define the linear "double-centering" operator $\mathcal{C}(\mathbf{A}) = \mathbf{JAJ}$ on $\mathbb{R}^{n\times n}$, where $\mathbf{J} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ is the orthogonal projector onto $\mathbf{1}^\perp$. We have the operator norm bound*

$$\|\mathcal{C}\|_{F\to F} \le 1.$$

Proof 1 *Note $\mathbf{J}^\top = \mathbf{J}$, $\mathbf{J}^2 = \mathbf{J}$ and $\|\mathbf{J}\|_2 = 1$. Using the standard inequality $\|\mathbf{UAV}\|_F \le \|\mathbf{U}\|_2\|\mathbf{A}\|_F\|\mathbf{V}\|_2$ (e.g., the chapter on matrix norms Horn & Johnson (2013)), we obtain for all $\mathbf{A}$ that*

$$\|\mathbf{JAJ}\|_F \le \|\mathbf{J}\|_2^2\|\mathbf{A}\|_F = \|\mathbf{A}\|_F,$$

so $\|\mathcal{C}\|_{F\to F} \le 1$.

A.3 Uniform concentration of the empirical risk

Define the conditional (given $\mathbf{X}^\star$) population risk

$$R_n^0(\mathbf{X}) \;=\; \mathbb{E}[R_n(\mathbf{X}) \mid \mathbf{X}^\star] \;=\; \frac{1}{N}\sum_{i<j} g\big(r_{ij}(\mathbf{X}); r_{ij}^\star\big), \tag{A.13}$$

where, for $Z \sim N(0,1)$,

$$g(r; r^\star) = \mathbb{E}\Big[\ell_M\big(T_M([r^\star + \sigma Z]_+), r\big)\Big]. \tag{A.14}$$

Lemma 2 (Uniform concentration) *There exist constants $C_1 = C_1(d,\sigma,M) > 0$, $C_2 = C_2(d,\sigma,M) > 0$, and $c = c(d,\sigma,M) > 0$ such that, for all $0 < t < C_1$,*

$$\mathbb{P}\Big(\sup_{\mathbf{X}\in([0,1]^d)^n} \big|R_n(\mathbf{X}) - R_n^0(\mathbf{X})\big| \;>\; t \;\Big|\; \mathbf{X}^\star\Big) \;\le\; 2\exp\Big(C_2\, nd\log\Big(\frac{C_1}{t}\Big) \;-\; c\,Nt^2\Big).$$

Proof 2 *Fix $\mathbf{X}$. Conditional on $\mathbf{X}^\star$, the random variables*

$$Y_{ij}(\mathbf{X}) = \ell_M(\tilde{D}_{ij}, r_{ij}(\mathbf{X}))$$

are independent across pairs (i,j).

On $0 < y < M$ we have $|y - r| \le M$, while at the atoms

$$q_{M,r}(0) = \Phi(-r/\sigma) \ge \Phi(-\sqrt{d}/\sigma), \qquad q_{M,r}(M) = 1 - \Phi((M-r)/\sigma) \ge \Phi(-M/\sigma).$$

Thus, because $r \in [0,\sqrt{d}]$ and $y \in [0,M]$, the density $q_{M,r}(y)$ is uniformly bounded above and below by constants $0 < q_- \le q_+ < \infty$ depending only on (d,σ,M). Hence

$$Y_{ij}(\mathbf{X}) \in [-\log q_+, -\log q_-] \tag{A.15}$$

almost surely. By Equation A.15, writing $W_\ell = \log(q_+/q_-)$, Hoeffding's inequality gives

$$\mathbb{P}\Big(\big|R_n(\mathbf{X}) - R_n^0(\mathbf{X})\big| > t \;\Big|\; \mathbf{X}^\star\Big) \;\le\; 2\exp\Big(-\frac{2Nt^2}{W_\ell^2}\Big). \tag{A.16}$$

To make the bound uniform in $\mathbf{X}$, we use a net argument on $([0,1]^d)^n$. Equip this space with the sup-norm on rows metric $\|\mathbf{X} - \mathbf{X}'\|_\infty = \max_i \|\mathbf{x}_i - \mathbf{x}_i'\|$. It is well-known (Vershynin, 2025, Corollary 4.2.11) that there exists a universal $0 < C_0 < \infty$ so that, for any fixed $\eta > 0$, there exists an η-net $\mathcal{N}_\eta$ satisfying

$$|\mathcal{N}_\eta| \;\le\; \Big(\frac{C_0}{\eta}\Big)^{nd}. \tag{A.17}$$

Next, note that since $\mathbf{X}, \mathbf{X}' \in ([0,1]^d)^n$*, we have that* $r_{ij}(\mathbf{X})$ *is 2-Lipschitz in* $\|\cdot\|_\infty$*. Also, for each fixed* $y \in [0, M]$*, the map* $r \mapsto \ell_M(y, r)$ *is* C^1 *on* $[0, \sqrt{d}]$*, with derivative*

$$\frac{\partial}{\partial r}\ell_M(y,r) = \begin{cases} \dfrac{\phi(r/\sigma)}{\sigma\Phi(-r/\sigma)}, & y = 0, \\ \dfrac{r-y}{\sigma^2}, & 0 < y < M, \\ -\dfrac{\phi((M-r)/\sigma)}{\sigma(1-\Phi((M-r)/\sigma))}, & y = M. \end{cases}$$

Each of these three branches is uniformly bounded on the relevant compact interval, so there exists $L_\ell = L_\ell(d, \sigma, M) < \infty$ *such that*

$$|\ell_M(y,r) - \ell_M(y,r')| \le L_\ell |r - r'|, \qquad y \in [0, M],\ r, r' \in [0, \sqrt{d}].$$

Combining this with the fact that $r_{ij}(\mathbf{X})$ *is 2-Lipschitz in* $\|\cdot\|_\infty$*, we have*

$$|\ell_M(\tilde{D}_{ij}, r_{ij}(\mathbf{X})) - \ell_M(\tilde{D}_{ij}, r_{ij}(\mathbf{X}'))| \le 2L_\ell \|\mathbf{X} - \mathbf{X}'\|_\infty . \tag{A.18}$$

Averaging over pairs gives

$$|R_n(\mathbf{X}) - R_n(\mathbf{X}')| \le 2L_\ell \|\mathbf{X} - \mathbf{X}'\|_\infty . \tag{A.19}$$

Since $R_n^0(\mathbf{X})$ *is a conditional expectation of* $R_n(\mathbf{X})$*, we have the same bound:*

$$|R_n^0(\mathbf{X}) - R_n^0(\mathbf{X}')| \le 2L_\ell \|\mathbf{X} - \mathbf{X}'\|_\infty . \tag{A.20}$$

Pick $\eta = t/(8L_\ell)$*. For any* $\mathbf{X}$*, choose* $\mathbf{X}' \in \mathcal{N}_\eta$ *with* $\|\mathbf{X} - \mathbf{X}'\|_\infty \le \eta$*. By the triangle inequality, Equation A.19 and Equation A.20,*

$$\begin{aligned} |R_n(\mathbf{X}) - R_n^0(\mathbf{X})| &\le |R_n(\mathbf{X}') - R_n^0(\mathbf{X}')| + |R_n(\mathbf{X}) - R_n(\mathbf{X}')| + |R_n^0(\mathbf{X}) - R_n^0(\mathbf{X}')| \\ &\le |R_n(\mathbf{X}') - R_n^0(\mathbf{X}')| + 4L_\ell\eta = |R_n(\mathbf{X}') - R_n^0(\mathbf{X}')| + \frac{t}{2}. \end{aligned} \tag{A.21}$$

Therefore, by Equation A.21, a union bound over $\mathcal{N}_\eta$ *together with Equation A.16 applied at level* $t/2$ *gives*

$$\mathbb{P}\Big(\sup_{\mathbf{X} \in ([0,1]^d)^n} |R_n(\mathbf{X}) - R_n^0(\mathbf{X})| > t \ \Big|\ \mathbf{X}^\star \Big) \le 2|\mathcal{N}_\eta| \exp\Big(-\frac{Nt^2}{2W_\ell^2} \Big).$$

Substituting Equation A.17 and $\eta = t/(8L_\ell)$ *and then enlarging constants gives the conclusion for all sufficiently small* t*.*

A.4 A quadratic lower bound for the population risk

We show $R_n^0(\mathbf{X})$ is minimized (up to isometry) at $\mathbf{X}^\star$, and furthermore that the risk increases quadratically as you move away from $\mathbf{X}^\star$. More precisely, we will show

Proposition A.2 *There exists* $c_{\mathrm{mg}} = c_{\mathrm{mg}}(d, \sigma, M) > 0$ *such that the following holds (uniformly in* $\mathbf{X}$*) with probability going to 1 as* n *goes to infinity:*

$$R_n^0(\mathbf{X}) - R_n^0(\mathbf{X}^\star) \ \ge\ c_{\mathrm{mg}}\ d_{\mathcal{G}}(\mathbf{X}, \mathbf{X}^\star)^2, \qquad \forall \mathbf{X} \in ([0,1]^d)^n. \tag{A.22}$$

This is the main result in this section. The proof is deferred to the end, in sub-section A.4.4.

A.4.1 Initial quadratic lower bound in the distance

Lemma 3 *There exists $c_{\mathrm{sc}} = c_{\mathrm{sc}}(\sigma, M, d) > 0$ such that for all $r, r^\star \in [0, \sqrt{d}]$,*

$$g(r; r^\star) - g(r^\star; r^\star) \;\geq\; c_{\mathrm{sc}}\,(r - r^\star)^2,$$

where g is defined in Equation A.14.

Proof 3 *For fixed $r^\star$, let P_r denote the law of $\tilde{D} = T_M([r + \sigma Z]_+)$, with density $q_{M,r}$ with respect to ν; equivalently, $P_r(A) = \int_A q_{M,r}(y)\,\nu(dy)$ for measurable $A \subseteq [0, M]$. Thus the observation generated at the true distance $r^\star$ has law $P_{r^\star}$, while evaluating the loss at a candidate distance r uses the density $q_{M,r}$. Therefore, by the definition of g in Equation A.14,*

$$g(r; r^\star) - g(r^\star; r^\star) = \int q_{M,r^\star}(y) \log\left(\frac{q_{M,r^\star}(y)}{q_{M,r}(y)}\right) \nu(dy) = \mathrm{KL}(P_{r^\star} \,\|\, P_r). \tag{A.23}$$

By Pinsker's inequality and Equation A.23,

$$g(r; r^\star) - g(r^\star; r^\star) \geq 2\|P_{r^\star} - P_r\|_{\mathrm{TV}}^2. \tag{A.24}$$

Recalling the definition of μ_M from Equation 5, we have

$$\begin{aligned} |\mu_M(r) - \mu_M(r^\star)| &= \left| \int_0^M y\,\big(q_{M,r}(y) - q_{M,r^\star}(y)\big)\,\nu(dy) \right| \\ &\leq M \int_0^M \big|q_{M,r}(y) - q_{M,r^\star}(y)\big|\,\nu(dy) \\ &= 2M\|P_r - P_{r^\star}\|_{\mathrm{TV}}. \end{aligned} \tag{A.25}$$

Combining Equation A.24 and Equation A.25 gives

$$g(r; r^\star) - g(r^\star; r^\star) \geq \frac{\big(\mu_M(r) - \mu_M(r^\star)\big)^2}{2M^2}. \tag{A.26}$$

For each fixed Z, the map $r \mapsto T_M([r + \sigma Z]_+)$ is piecewise linear with derivative $\mathbf{1}\{0 < r + \sigma Z < M\}$ for almost every r, so dominated convergence yields

$$\mu_M'(r) = \mathbb{P}(0 < r + \sigma Z < M) = \Phi\left(\frac{M - r}{\sigma}\right) - \Phi\left(-\frac{r}{\sigma}\right). \tag{A.27}$$

Hence, for $r \in [0, \sqrt{d}]$,

$$\mu_M'(r) = \Phi\left(\frac{M - r}{\sigma}\right) + \Phi\left(\frac{r}{\sigma}\right) - 1 \geq \Phi\left(\frac{M}{2\sigma}\right) - \frac{1}{2} =: m_M. \tag{A.28}$$

We note that $m_M > 0$, since $r \in [0, M]$ and therefore at least one of r and $M - r$ is at least $M/2$. By the mean value theorem and Equation A.28,

$$|\mu_M(r) - \mu_M(r^\star)| \geq m_M |r - r^\star|. \tag{A.29}$$

Substituting Equation A.29 into Equation A.26 proves the claim, with $c_{\mathrm{sc}} = m_M^2/(2M^2)$.

A.4.2 From pairwise distance error to Procrustes error

Define the average squared distance discrepancy

$$\Delta_{\mathrm{dist}}(\mathbf{X}, \mathbf{X}^\star) = \frac{1}{N} \sum_{i<j} \big(r_{ij}(\mathbf{X}) - r_{ij}^\star\big)^2.$$

We have the straightforward estimate:

Lemma 4 (Population risk dominates distance discrepancy) *Under the model 1 and loss (Equation 6),*

$$R_n^0(\mathbf{X}) - R_n^0(\mathbf{X}^\star) \;\geq\; c_{\mathrm{sc}}\; \Delta_{\mathrm{dist}}(\mathbf{X}, \mathbf{X}^\star),$$

where c_{sc} is as in Lemma 3.

Proof 4 *By Lemma 3, for each pair (i, j),*

$$g\big(r_{ij}(\mathbf{X}); r_{ij}^\star\big) - g\big(r_{ij}^\star; r_{ij}^\star\big) \;\geq\; c_{\mathrm{sc}}\big(r_{ij}(\mathbf{X}) - r_{ij}^\star\big)^2.$$

Averaging over pairs yields the result.

Next we relate $\Delta_{\mathrm{dist}}(\mathbf{X}, \mathbf{X}^\star)$ to $d_{\mathcal{G}}(\mathbf{X}, \mathbf{X}^\star)$ using a standard Procrustes perturbation bound. Recall that $\mathbf{X}_c = \mathbf{J}\mathbf{X}$ and $\mathbf{X}_c^\star = \mathbf{J}\mathbf{X}^\star$ denote the centered configurations.

We will use the following consequence of Theorem 1 Arias-Castro et al. (2020) as our basic Procrustes perturbation bound: with $\varepsilon^2 = \left\|\mathbf{Y}\mathbf{Y}^\top - \mathbf{X}\mathbf{X}^\top\right\|_F$, if $\left\|\mathbf{X}^\dagger\right\|_{\mathrm{op}} \varepsilon \leq 1/\sqrt{2}$, then

$$\min_{\mathbf{Q}\in\mathcal{O}(d)} \|\mathbf{Y} - \mathbf{X}\mathbf{Q}\|_F \leq (1+\sqrt{2})\left\|\mathbf{X}^\dagger\right\|_{\mathrm{op}} \left\|\mathbf{Y}\mathbf{Y}^\top - \mathbf{X}\mathbf{X}^\top\right\|_F.$$

Note that the theorem in that paper looks slightly different and is quite a bit more complicated; this follows immediately from the second branch in (Arias-Castro et al., 2020, Theorem 1, inequality (5)) after choosing Schatten norm $p = 2$, simplifying and rephrasing in our notation.

Lemma 5 *For every $\alpha > 0$, there exists $c_2 = c_2(\alpha, d) > 0$ such that, if*

$$s_{\min}(\mathbf{X}_c^\star) \;\geq\; \alpha\sqrt{n}, \tag{A.30}$$

then

$$d_{\mathcal{G}}(\mathbf{X}, \mathbf{X}^\star)^2 \;\leq\; c_2\; \Delta_{\mathrm{dist}}(\mathbf{X}, \mathbf{X}^\star), \qquad \forall \mathbf{X} \in ([0,1]^d)^n.$$

Proof 5 *We follow Arias-Castro et al. (2020) and work with centered Gram matrices. Denote by $D^2(\mathbf{X})$ the squared-distance matrix with entries $D^2(\mathbf{X})_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2$. With $\mathbf{J}$ as above, the centered Gram matrix is*

$$G(\mathbf{X}) = -\frac{1}{2}\,\mathbf{J}D^2(\mathbf{X})\mathbf{J} = \mathbf{X}_c\mathbf{X}_c^\top.$$

We begin by bounding $G(\mathbf{X}) - G(\mathbf{X}^\star)$ in terms of $\Delta_{\mathrm{dist}}(\mathbf{X}, \mathbf{X}^\star)$. By Lemma 1, the Frobenius–Frobenius operator norm of $\mathbf{A} \mapsto \mathbf{J}\mathbf{A}\mathbf{J}$ is 1, so

$$\|G(\mathbf{X}) - G(\mathbf{X}^\star)\|_F = \frac{1}{2}\left\|\mathbf{J}\big(D^2(\mathbf{X}) - D^2(\mathbf{X}^\star)\big)\mathbf{J}\right\|_F \leq \frac{1}{2}\left\|D^2(\mathbf{X}) - D^2(\mathbf{X}^\star)\right\|_F. \tag{A.31}$$

For $r, r^\star \in [0, \sqrt{d}]$ we have

$$|r^2 - (r^\star)^2| = |r - r^\star|\,(r + r^\star) \leq 2\sqrt{d}\,|r - r^\star|. \tag{A.32}$$

Therefore, by Equation A.32,

$$\begin{aligned}
\left\|D^2(\mathbf{X}) - D^2(\mathbf{X}^\star)\right\|_F^2 &= \sum_{i\neq j}\big(r_{ij}(\mathbf{X})^2 - (r_{ij}^\star)^2\big)^2 \\
&\leq 4d\sum_{i\neq j}\big(r_{ij}(\mathbf{X}) - r_{ij}^\star\big)^2 \qquad \text{(A.33)}\\
&= 8d\sum_{i<j}\big(r_{ij}(\mathbf{X}) - r_{ij}^\star\big)^2 \\
&= 8dN\,\Delta_{\mathrm{dist}}(\mathbf{X}, \mathbf{X}^\star). \qquad \text{(A.34)}
\end{aligned}$$

Combining Equation A.31 and Equation A.33 gives

$$\|G(\mathbf{X}) - G(\mathbf{X}^\star)\|_F \leq \sqrt{2dN\,\Delta_{\mathrm{dist}}(\mathbf{X}, \mathbf{X}^\star)}. \tag{A.35}$$

Let

$$\varepsilon = \|G(\mathbf{X}) - G(\mathbf{X}^\star)\|_F^{1/2}. \tag{A.36}$$

We now apply (Arias-Castro et al., 2020, Theorem 1) with $\mathbf{X} = \mathbf{X}_c^\star$ *and* $\mathbf{Y} = \mathbf{X}_c$ *and Schatten norm* $p = 2$ *(which is equal to the Frobenius norm). In the notation of that theorem, we have* $\mathbf{Y}\mathbf{Y}^\top - \mathbf{X}\mathbf{X}^\top = G(\mathbf{X}) - G(\mathbf{X}^\star)$ *and hence* $\varepsilon^2 = \|\mathbf{Y}\mathbf{Y}^\top - \mathbf{X}\mathbf{X}^\top\|_F$*. Using the second branch in the statement of (Arias-Castro et al., 2020, Theorem 1, inequality (5)), if*

$$\|(\mathbf{X}_c^\star)^\dagger\|_{\mathrm{op}}\,\varepsilon \leq \frac{1}{\sqrt{2}}, \tag{A.37}$$

then

$$\min_{\mathbf{Q}\in\mathcal{O}(d)} \|\mathbf{X}_c - \mathbf{X}_c^\star\mathbf{Q}\|_F \leq (1+\sqrt{2})\,\|(\mathbf{X}_c^\star)^\dagger\|_{\mathrm{op}}\;\|G(\mathbf{X}) - G(\mathbf{X}^\star)\|_F\,. \tag{A.38}$$

We now give a sufficient condition for Equation A.37. Under Equation A.30, we have $\|(\mathbf{X}_c^\star)^\dagger\|_{\mathrm{op}} = 1/s_{\min}(\mathbf{X}_c^\star) \leq 1/(\alpha\sqrt{n})$*. Moreover, since* $N = \binom{n}{2} \leq n^2/2$*, Equations A.35–A.36 imply*

$$\varepsilon \leq \big(2dN\,\Delta_{\mathrm{dist}}(\mathbf{X}, \mathbf{X}^\star)\big)^{1/4} \leq d^{1/4}n^{1/2}\,\Delta_{\mathrm{dist}}(\mathbf{X}, \mathbf{X}^\star)^{1/4}. \tag{A.39}$$

Hence, by Equation A.30 and Equation A.39, Equation A.37 holds whenever

$$\Delta_{\mathrm{dist}}(\mathbf{X}, \mathbf{X}^\star) \;\leq\; \frac{\alpha^4}{4d}. \tag{A.40}$$

We now obtain our bound in two cases: when Equation A.40 holds, and when it fails. We begin by assuming it holds. Combining Equation A.38 with Equation A.35 and Equation A.30 gives

$$\begin{aligned}\frac{1}{\sqrt{n}}\min_{\mathbf{Q}\in\mathcal{O}(d)} \|\mathbf{X}_c - \mathbf{X}_c^\star\mathbf{Q}\|_F &\leq \frac{1+\sqrt{2}}{\alpha n}\;\sqrt{2dN\,\Delta_{\mathrm{dist}}(\mathbf{X}, \mathbf{X}^\star)} \\ &\leq \frac{1+\sqrt{2}}{\alpha}\;\sqrt{d\,\Delta_{\mathrm{dist}}(\mathbf{X}, \mathbf{X}^\star)}.\end{aligned} \tag{A.41}$$

By the definition of $d_{\mathcal{G}}$ *in Equation A.5, the left-hand side equals* $d_{\mathcal{G}}(\mathbf{X}, \mathbf{X}^\star)$*, so squaring Equation A.41 yields*

$$d_{\mathcal{G}}(\mathbf{X}, \mathbf{X}^\star)^2 \leq \frac{(1+\sqrt{2})^2}{\alpha^2}\;d\;\Delta_{\mathrm{dist}}(\mathbf{X}, \mathbf{X}^\star) \qquad \textit{whenever Equation A.40 holds.} \tag{A.42}$$

Finally, we consider the case that Equation A.40 fails. In this case, $\Delta_{\mathrm{dist}}(\mathbf{X}, \mathbf{X}^\star) \geq \alpha^4/(4d)$*. Since* $\mathbf{x}_i, \mathbf{x}_i^\star \in [0,1]^d$*, choosing* $\mathbf{Q} = \mathbf{I}$ *and* $\mathbf{b} = \mathbf{0}$ *in the uncentered definition of the orbit distance gives the trivial bound* $d_{\mathcal{G}}(\mathbf{X}, \mathbf{X}^\star)^2 \leq d$*. Therefore,*

$$d_{\mathcal{G}}(\mathbf{X}, \mathbf{X}^\star)^2 \leq d \leq \frac{4d^2}{\alpha^4}\;\Delta_{\mathrm{dist}}(\mathbf{X}, \mathbf{X}^\star) \qquad \textit{whenever Equation A.40 fails.} \tag{A.43}$$

Taking

$$c_2 = \max\Big\{\frac{(1+\sqrt{2})^2}{\alpha^2}d,\;\frac{4d^2}{\alpha^4}\Big\}$$

and combining Equation A.42 and Equation A.43 proves the lemma.

A.4.3 Well-conditioning holds with high probability for uniform points

Lemma 6 *Let $\mathbf{x}_i^\star \overset{iid}{\sim} \mathrm{Unif}([0,1]^d)$, let $\mathbf{X}^\star$ be the associated matrix of points, and let $\mathbf{X}_c^\star$ be its centered version. Then there exists $c_1 = c_1(d) > 0$ such that*

$$\lim_{n\to\infty} \mathbb{P}\Big(s_{\min}(\mathbf{X}_c^\star) \geq c_1\sqrt{n}\Big) = 1.$$

Proof 6 *Let $\mathbf{u}_i = \mathbf{x}_i^\star - \mathbb{E}[\mathbf{x}_i^\star] = \mathbf{x}_i^\star - \frac{1}{2}\mathbf{1}_d$, and let $\mathbf{U}$ be the $n \times d$ matrix with rows $\mathbf{u}_i^\top$. Then the rows of $\mathbf{U}$ are i.i.d., mean-zero, bounded random vectors in $\mathbb{R}^d$, and*

$$\mathbf{X}_c^\star = \mathbf{J}\mathbf{U} = \mathbf{U} - \mathbf{1}\bar{\mathbf{u}}^\top, \qquad \bar{\mathbf{u}} = \frac{1}{n}\sum_{i=1}^n \mathbf{u}_i.$$

We first lower-bound $s_{\min}(\mathbf{U})$. Set $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}^{-1/2} = \sqrt{12}\,\mathbf{U}$, where

$$\mathbf{\Sigma} = \mathbb{E}[\mathbf{u}_i\mathbf{u}_i^\top] = \frac{1}{12}\mathbf{I}_d.$$

Then the rows of $\mathbf{A}$ are isotropic, subgaussian and have subgaussian norm bounded by a constant $K_d < \infty$ depending only on d. By (Vershynin, 2025, Theorem 4.6.1), there exists an absolute constant $0 < C < \infty$ so that for every $t > 0$,

$$\mathbb{P}\big[s_{\min}(\mathbf{A}) \geq \sqrt{n} - CK_d^2(\sqrt{d} + t)\big] \geq 1 - 2e^{-t^2}. \tag{A.44}$$

Since $\mathbf{A} = \sqrt{12}\,\mathbf{U}$, choosing $t = \sqrt{n}/(4CK_d^2)$ in Equation A.44 shows that there exist constants $c_U = c_U(d) > 0$ and $c_A = c_A(d) > 0$ such that, for all sufficiently large n,

$$\mathbb{P}\big(s_{\min}(\mathbf{U}) \geq c_U\sqrt{n}\big) \geq 1 - 2e^{-c_A n}. \tag{A.45}$$

We next control the rank-one correction $\mathbf{1}\bar{\mathbf{u}}^\top$. Because each coordinate of $\mathbf{u}_i$ is bounded and mean-zero, Hoeffding's inequality gives, for every $a > 0$,

$$\mathbb{P}\big(\sqrt{n}\,|\bar{u}_k| > a\big) \leq 2e^{-ca^2}, \qquad k = 1, \ldots, d,$$

for a constant $c = c(d) > 0$. Therefore, by a union bound, for every $b > 0$,

$$\mathbb{P}\big(\sqrt{n}\,\|\bar{\mathbf{u}}\| > b\big) \leq 2d\exp\left(-\frac{cb^2}{d}\right). \tag{A.46}$$

Since $\|\mathbf{1}\bar{\mathbf{u}}^\top\|_{\mathrm{op}} = \sqrt{n}\,\|\bar{\mathbf{u}}\|$, substituting $b = (c_U/2)\sqrt{n}$ in Equation A.46 gives

$$\mathbb{P}\left(\|\mathbf{1}\bar{\mathbf{u}}^\top\|_{\mathrm{op}} > \frac{c_U}{2}\sqrt{n}\right) \leq 2d\exp\left(-\frac{cc_U^2}{4d}n\right). \tag{A.47}$$

Let

$$E_n = \big\{s_{\min}(\mathbf{U}) \geq c_U\sqrt{n}\big\} \cap \left\{\|\mathbf{1}\bar{\mathbf{u}}^\top\|_{\mathrm{op}} \leq \frac{c_U}{2}\sqrt{n}\right\}.$$

By Equation A.45 and Equation A.47, $\mathbb{P}(E_n) \to 1$. On E_n, since $\mathbf{X}_c^\star = \mathbf{U} - \mathbf{1}\bar{\mathbf{u}}^\top$, Weyl's inequality for singular values gives

$$s_{\min}(\mathbf{X}_c^\star) = s_{\min}(\mathbf{U} - \mathbf{1}\bar{\mathbf{u}}^\top) \geq s_{\min}(\mathbf{U}) - \|\mathbf{1}\bar{\mathbf{u}}^\top\|_{\mathrm{op}} \geq (c_U/2)\sqrt{n}.$$

Taking $c_1 = c_U/2$ proves the lemma.

A.4.4 Completing the quadratic lower bounds

Proof 7 (Proof of Proposition A.2) *Let $c_1 = c_1(d)$ be the constant from Lemma 6. We prove Equation A.22 on the high-probability event $s_{\min}(\mathbf{X}_c^\star) \geq c_1\sqrt{n}$, applying Lemma 5 with $\alpha = c_1$.*

Combine Lemma 4 and Lemma 5:

$$R_n^0(\mathbf{X}) - R_n^0(\mathbf{X}^\star) \;\geq\; c_{\mathrm{sc}}\,\Delta_{\mathrm{dist}}(\mathbf{X},\mathbf{X}^\star) \;\geq\; c_{\mathrm{sc}} \cdot \frac{1}{c_2}\; d_{\mathcal{G}}(\mathbf{X},\mathbf{X}^\star)^2.$$

Set $c_{\mathrm{mg}} = c_{\mathrm{sc}}/c_2$.

Proposition A.3 *There exists a constant $C_{\mathrm{up}} = C_{\mathrm{up}}(d,\sigma,M) > 0$ such that, for every $\mathbf{X} \in ([0,1]^d)^n$,*

$$R_n^0(\mathbf{X}) - R_n^0(\mathbf{X}^\star) \;\leq\; C_{\mathrm{up}}\, d_{\mathcal{G}}(\mathbf{X},\mathbf{X}^\star)^2. \tag{A.48}$$

Proof 8 *Fix $\mathbf{X}$ and choose $\mathbf{Q} \in \mathcal{O}(d)$ so that $d_{\mathcal{G}}(\mathbf{X},\mathbf{X}^\star) = n^{-1/2}\,\|\mathbf{X}_c - \mathbf{X}_c^\star\mathbf{Q}\|_F$. Write $\mathbf{H} = \mathbf{X}_c - \mathbf{X}_c^\star\mathbf{Q}$, and let $\mathbf{h}_i$ denote the ith row of $\mathbf{H}$. Then $\|\mathbf{H}\|_F^2 = n\, d_{\mathcal{G}}(\mathbf{X},\mathbf{X}^\star)^2$ and the rows of $\mathbf{H}$ sum to 0.*

For each fixed $y \in [0,M]$, the map $r \mapsto \ell_M(y,r)$ is C^2 on $[0,\sqrt{d}]$, and each of the three branches from Equation A.2 has second derivative uniformly bounded on the relevant compact interval. Hence there exists $H_\ell = H_\ell(d,\sigma,M) < \infty$ such that

$$\sup_{y\in[0,M]}\ \sup_{r\in[0,\sqrt{d}]} \left|\partial_r^2 \ell_M(y,r)\right| \leq H_\ell. \tag{A.49}$$

By Equation A.14,

$$g(r;r^\star) = \int \ell_M(y,r) q_{M,r^\star}(y)\,\nu(dy).$$

Using $\partial_r \ell_M(y,r) = -(\partial_r q_{M,r}(y))/q_{M,r}(y)$, we get

$$\begin{aligned}
\partial_r g(r;r^\star)\big|_{r=r^\star} &= -\int q_{M,r^\star}(y)\frac{\partial_r q_{M,r}(y)|_{r=r^\star}}{q_{M,r^\star}(y)}\,\nu(dy)\\
&= -\int \partial_r q_{M,r}(y)\big|_{r=r^\star}\,\nu(dy)\\
&= -\partial_r\left(\int q_{M,r}(y)\,\nu(dy)\right)\Bigg|_{r=r^\star} = 0,
\end{aligned}$$

because $\int q_{M,r}(y)\,\nu(dy) = 1$ for every r. Moreover, Equation A.49 implies $|\partial_r^2 g(r;r^\star)| \leq H_\ell$ for all $r, r^\star \in [0,\sqrt{d}]$. Taylor's theorem therefore gives

$$g(r;r^\star) - g(r^\star;r^\star) \leq \frac{H_\ell}{2}(r-r^\star)^2, \qquad r, r^\star \in [0,\sqrt{d}]. \tag{A.50}$$

Applying Equation A.50 with $r = r_{ij}(\mathbf{X})$ and $r^\star = r_{ij}^\star$ and then using the fact that centering and rotation do not change pairwise distances, we obtain

$$g\big(r_{ij}(\mathbf{X});r_{ij}^\star\big) - g\big(r_{ij}^\star;r_{ij}^\star\big) \leq \frac{H_\ell}{2}\big(r_{ij}(\mathbf{X}) - r_{ij}^\star\big)^2 \leq \frac{H_\ell}{2}\,\|\mathbf{h}_i - \mathbf{h}_j\|^2\,. \tag{A.51}$$

Averaging Equation A.51 over pairs and using the identity $\sum_{i<j}\|\mathbf{h}_i - \mathbf{h}_j\|^2 = n\,\|\mathbf{H}\|_F^2$ yields

$$R_n^0(\mathbf{X}) - R_n^0(\mathbf{X}^\star) \leq \frac{H_\ell}{2N}\sum_{i<j}\|\mathbf{h}_i - \mathbf{h}_j\|^2 = \frac{H_\ell n\,\|\mathbf{H}\|_F^2}{2N} \leq 2H_\ell\, d_{\mathcal{G}}(\mathbf{X},\mathbf{X}^\star)^2,$$

since $N = \binom{n}{2} \geq n^2/4$ for $n \geq 2$. This proves Equation A.48 with $C_{\mathrm{up}} = 2H_\ell$.

A.5 Consistency of the exact posterior

Proposition A.4 (Consistency of the exact posterior) *Suppose Assumption A.2 holds. Then, for every $\delta > 0$,*

$$\Pi_n\Big(\{\mathbf{X}:\ d_{\mathcal{G}}(\mathbf{X},\mathbf{X}^\star) > \delta\}\ \Big|\ \mathbf{D}\Big)\ \xrightarrow{\mathbb{P}}\ 0.$$

Proof 9 *Fix $\delta > 0$, and define the constant*

$$a_0 = \min\Big\{\frac{1}{2},\ \sqrt{\frac{c_{\mathrm{mg}}}{4C_{\mathrm{up}}}}\Big\} \in \Big(0, \frac{1}{2}\Big], \tag{A.52}$$

where c_{mg} is from Proposition A.2 and C_{up} is from Proposition A.3. Let

$$A_\delta = \{\mathbf{X}:\ d_{\mathcal{G}}(\mathbf{X},\mathbf{X}^\star) > \delta\}, \qquad B_{a_0\delta} = \{\mathbf{X}:\ d_{\mathcal{G}}(\mathbf{X},\mathbf{X}^\star) \le a_0\delta\}, \qquad \widetilde{B}_{a_0\delta} = B_{a_0\delta} \cap ([0,1]^d)^n.$$

Because the integrand is nonnegative, restricting the domain of integration gives

$$\int e^{-NR_n(\mathbf{X})}\,\Pi_0(d\mathbf{X})\ \ge\ \int_{\widetilde{B}_{a_0\delta}} e^{-NR_n(\mathbf{X})}\,\Pi_0(d\mathbf{X}). \tag{A.53}$$

Let $\eta_n = \sup_{\mathbf{X}\in([0,1]^d)^n} |R_n(\mathbf{X}) - R_n^0(\mathbf{X})|$. By Lemma 2, $\eta_n \to 0$ in probability. Let

$$E_n^{\mathrm{cond}} = \{s_{\min}(\mathbf{X}_c^\star) \ge c_1\sqrt{n}\},$$

where $c_1 = c_1(d)$ is the constant from Lemma 6; then $\mathbb{P}(E_n^{\mathrm{cond}}) \to 1$. Fix $\gamma > 0$. On the event $\{\eta_n \le \gamma\}$ we have, for all $\mathbf{X}$,

$$R_n(\mathbf{X}) \ge R_n^0(\mathbf{X}) - \gamma, \qquad R_n(\mathbf{X}) \le R_n^0(\mathbf{X}) + \gamma. \tag{A.54}$$

Therefore, using Equation A.53 and then Equation A.54,

$$\begin{aligned}\Pi_n(A_\delta \mid \mathbf{D}) &= \frac{\int_{A_\delta} e^{-NR_n(\mathbf{X})}\Pi_0(d\mathbf{X})}{\int e^{-NR_n(\mathbf{X})}\Pi_0(d\mathbf{X})} \le \frac{\int_{A_\delta} e^{-NR_n(\mathbf{X})}\Pi_0(d\mathbf{X})}{\int_{\widetilde{B}_{a_0\delta}} e^{-NR_n(\mathbf{X})}\Pi_0(d\mathbf{X})} \\ &\le \frac{\int_{A_\delta} e^{-N(R_n^0(\mathbf{X})-\gamma)}\Pi_0(d\mathbf{X})}{\int_{\widetilde{B}_{a_0\delta}} e^{-N(R_n^0(\mathbf{X})+\gamma)}\Pi_0(d\mathbf{X})} = e^{2N\gamma}\cdot\frac{\int_{A_\delta} e^{-NR_n^0(\mathbf{X})}\Pi_0(d\mathbf{X})}{\int_{\widetilde{B}_{a_0\delta}} e^{-NR_n^0(\mathbf{X})}\Pi_0(d\mathbf{X})}.\end{aligned}$$

Next, on E_n^{cond}, Proposition A.2 gives

$$R_n^0(\mathbf{X}) \ge R_n^0(\mathbf{X}^\star) + c_{\mathrm{mg}}\delta^2 \qquad \textit{for all } \mathbf{X} \in A_\delta. \tag{A.55}$$

Also, Proposition A.3 and the choice of a_0 in Equation A.52 imply

$$R_n^0(\mathbf{X}) \le R_n^0(\mathbf{X}^\star) + C_{\mathrm{up}}a_0^2\delta^2 \le R_n^0(\mathbf{X}^\star) + \frac{1}{4}c_{\mathrm{mg}}\delta^2 \qquad \textit{for all } \mathbf{X} \in \widetilde{B}_{a_0\delta}. \tag{A.56}$$

Using Equation A.55 and Equation A.56,

$$\int_{A_\delta} e^{-NR_n^0(\mathbf{X})}\Pi_0(d\mathbf{X}) \le e^{-N(R_n^0(\mathbf{X}^\star)+c_{\mathrm{mg}}\delta^2)}\Pi_0([0,1]^{dn}), \tag{A.57}$$

and

$$\int_{\widetilde{B}_{a_0\delta}} e^{-NR_n^0(\mathbf{X})}\Pi_0(d\mathbf{X}) \ge e^{-N(R_n^0(\mathbf{X}^\star)+c_{\mathrm{mg}}\delta^2/4)}\Pi_0(\widetilde{B}_{a_0\delta}). \tag{A.58}$$

Combining Equations A.57–A.58 on $E_n^{\mathrm{cond}} \cap \{\eta_n \le \gamma\}$ yields

$$\Pi_n(A_\delta \mid \mathbf{D}) \le \frac{\Pi_0([0,1]^{dn})}{\Pi_0(\widetilde{B}_{a_0\delta})}\exp\Big(-N\cdot\tfrac{3}{4}c_{\mathrm{mg}}\delta^2 + 2N\gamma\Big). \tag{A.59}$$

It remains to lower-bound $\Pi_0(\widetilde{B}_{a_0\delta})$*. By Assumption A.2(T3), if* π_0 *denotes the prior density then*

$$\pi_0(\mathbf{X}) \geq e^{-cn}, \qquad \mathbf{X} \in ([0,1]^d)^n, \tag{A.60}$$

for some constant $c > 0$*. Moreover, by the definition*

$$d_{\mathcal{G}}(\mathbf{X}, \mathbf{X}^\star) = \frac{1}{\sqrt{n}} \min_{\mathbf{Q}\in\mathcal{O}(d)} \|\mathbf{X}_c - \mathbf{X}_c^\star \mathbf{Q}\|_F ,$$

we have the inclusion

$$\Big\{\mathbf{X} \in ([0,1]^d)^n : \ \|\mathbf{X} - \mathbf{X}^\star\|_F \leq \sqrt{n}\, a_0\delta\Big\} \ \subseteq \ \widetilde{B}_{a_0\delta}. \tag{A.61}$$

Set

$$b_\delta = \min\Big\{1, \frac{a_0\delta}{\sqrt{d}}\Big\},$$

and write $\|\cdot\|_{\max}$ *for the maximum absolute coordinate over the* $n \times d$ *matrix. Since* $\|\mathbf{X} - \mathbf{X}^\star\|_{\max} \leq a_0\delta/\sqrt{d}$ *implies* $\|\mathbf{X} - \mathbf{X}^\star\|_F \leq \sqrt{n}\, a_0\delta$*, the box*

$$C_\delta = \Big\{\mathbf{X} \in ([0,1]^d)^n : \ \|\mathbf{X} - \mathbf{X}^\star\|_{\max} \leq a_0\delta/\sqrt{d}\Big\}$$

is contained in the ball $\widetilde{B}_{a_0\delta}$*.*

Note that each side of this box has length at least b_δ*, so* $\mathrm{Vol}(C_\delta) \geq b_\delta^{nd}$*. Therefore, by Equation A.60,*

$$\Pi_0(\widetilde{B}_{a_0\delta}) \ \geq \ e^{-cn} \cdot \mathrm{Vol}(C_\delta) \ \geq \ e^{-cn}\, b_\delta^{nd}. \tag{A.62}$$

In particular,

$$-\log \Pi_0(\widetilde{B}_{a_0\delta}) \ \leq \ cn + nd \log\Big(\frac{1}{b_\delta}\Big), \tag{A.63}$$

which is $O(n)$ *for each fixed* $\delta > 0$*.*

Fix $\tau = \frac{3}{16} c_{\mathrm{mg}} \delta^2$*. Since* $\eta_n \to 0$ *in probability and* $\mathbb{P}(E_n^{\mathrm{cond}}) \to 1$*, the event* $\{\eta_n \leq \tau\} \cap E_n^{\mathrm{cond}}$ *has probability tending to* 1*. On this event, Equation A.59 with* $\gamma = \tau$ *and Equation A.63 give*

$$\Pi_n(A_\delta \mid \mathbf{D}) \leq \frac{\Pi_0([0,1]^{dn})}{\Pi_0(\widetilde{B}_{a_0\delta})} \exp\Big(- N \cdot \tfrac{3}{8} c_{\mathrm{mg}} \delta^2\Big),$$

and the right-hand side tends to 0 *because* $N = \binom{n}{2} \asymp n^2$ *while Equation A.63 is* $O(n)$ *for fixed* δ*. This proves the claim.*

A.6 Proof of Proposition A.1

Let ε_n be as in Equation A.6, and define

$$E_n^{\mathrm{app}} = \left\{ \sup_{\mathbf{X}\in([0,1]^d)^n} |\widehat{R}_n(\mathbf{X}) - R_n(\mathbf{X})| \leq \varepsilon_n \right\}.$$

By Equation A.7, $\mathbb{P}(E_n^{\mathrm{app}}) \to 1$. On E_n^{app}, Equation A.7 gives

$$\sup_{\mathbf{X}\in([0,1]^d)^n} |\widehat{R}_n(\mathbf{X}) - R_n(\mathbf{X})| \leq \varepsilon_n.$$

Repeat the proof of Proposition A.4 with $\widehat{R}_n$ in place of R_n. Everywhere that the previously defined quantity η_n appeared, we now have, on E_n^{app},

$$\widehat{\eta}_n = \sup_{\mathbf{X}\in([0,1]^d)^n} |\widehat{R}_n(\mathbf{X}) - R_n^0(\mathbf{X})| \leq \sup_{\mathbf{X}\in([0,1]^d)^n} |\widehat{R}_n(\mathbf{X}) - R_n(\mathbf{X})| + \sup_{\mathbf{X}\in([0,1]^d)^n} |R_n(\mathbf{X}) - R_n^0(\mathbf{X})| \leq \varepsilon_n + \eta_n. \tag{A.64}$$

Since $\mathbb{P}(E_n^{\mathrm{app}}) \to 1$, $\varepsilon_n \to 0$ by assumption, and $\eta_n \to 0$ in probability by Lemma 2, Equation A.64 gives $\widehat{\eta}_n \to 0$ in probability. The remainder of the argument (risk gap $\Rightarrow$ posterior concentration) is unchanged.

A.7 Proof of Theorem A.1

Let

$$\Delta_n^{\text{mix}} = \sup_{\mathbf{X}_0 \in ([0,1]^d)^n} \left\| P(\mathbf{X}_0, \cdot) - P_{\xi_0}(\mathbf{X}_0, \cdot) \right\|_{\text{TV}}.$$

By convexity of total variation and the definition of P,

$$\Delta_n^{\text{mix}} \leq \int_{\Xi} \sup_{\mathbf{X}_0 \in ([0,1]^d)^n} \left\| P_\xi(\mathbf{X}_0, \cdot) - P_{\xi_0}(\mathbf{X}_0, \cdot) \right\|_{\text{TV}} \mu(d\xi) \leq \Delta_n. \tag{A.65}$$

By the uniform-ergodic perturbation bound in Alquier et al. (2016) and Equation A.65,

$$\sup_{\mathbf{X}_0 \in ([0,1]^d)^n} \left\| \delta_{\mathbf{X}_0} P^m - \delta_{\mathbf{X}_0} P_{\xi_0}^m \right\|_{\text{TV}} \leq \kappa_n \Delta_n^{\text{mix}} \leq \kappa_n \Delta_n, \qquad m \geq 0. \tag{A.66}$$

Integrating Equation A.66 against $\mathbf{X}_0 \sim \widetilde{\Pi}_n$ and using the invariance of $\widetilde{\Pi}_n$ under P gives

$$\left\| \widetilde{\Pi}_n - \widetilde{\Pi}_n P_{\xi_0}^m \right\|_{\text{TV}} \leq \kappa_n \Delta_n. \tag{A.67}$$

By the triangle inequality and Equation A.67,

$$\left\| \widetilde{\Pi}_n - \widehat{\Pi}_n^{(\xi_0)} \right\|_{\text{TV}} \leq \left\| \widetilde{\Pi}_n - \widetilde{\Pi}_n P_{\xi_0}^m \right\|_{\text{TV}} + \left\| \widetilde{\Pi}_n P_{\xi_0}^m - \widehat{\Pi}_n^{(\xi_0)} \right\|_{\text{TV}}.$$

Since P_{ξ_0} has invariant distribution $\widehat{\Pi}_n^{(\xi_0)}$ and is uniformly ergodic, the second term tends to 0 as $m \to \infty$. Letting $m \to \infty$ yields

$$\left\| \widetilde{\Pi}_n - \widehat{\Pi}_n^{(\xi_0)} \right\|_{\text{TV}} \leq \kappa_n \Delta_n. \tag{A.68}$$

Now set $A_\delta = \{\mathbf{X} : \ d_{\mathcal{G}}(\mathbf{X}, \mathbf{X}^\star) > \delta\}$. Then Equation A.68 gives

$$\widetilde{\Pi}_n(A_\delta) \leq \widehat{\Pi}_n^{(\xi_0)}(A_\delta) + \kappa_n \Delta_n.$$

The first term goes to 0 in probability by Proposition A.1, and the second does as well by assumption.

B Supplementary algorithms

Algorithm B.1: Quadtree construction

Input: $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^2$, $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \mathbb{R}^\Omega$
Output: quadtree $\mathcal{T}$
Function `Insert`(i, k)
  if $\mathbf{x}_i \notin \mathcal{C}_k$ **then**
    return false
  Update $N_k, \bar{\mathbf{x}}_k, \bar{\mathbf{y}}_k$;
  if *node k is an empty leaf* **then**
    Store $(\mathbf{x}_i, \mathbf{y}_i)$ in *node k*;
    return true
  if *node k is a leaf* **then**
    Subdivide $\mathcal{C}_k$ into children $c \in \mathit{children}(k)$;
    for *each stored point j in node k* **do**
      Determine $c(j)$ and call `Insert`$(j, c(j))$;
    Mark *node k* as internal;
  foreach $c \in \mathit{children}(k)$ **do**
    if `Insert`(i, c) **then**
      return true
  return false
Initialize root node $k = 0$ with cell $\mathcal{C}_0$;
for $i = 1, \ldots, n$ **do**
  `Insert`$(i, 0)$;
return $\mathcal{T}$

Algorithm B.2: Local quadtree update (forward)

Input: quadtree $\mathcal{T}$, current point $\mathbf{x}_i$, proposed point $\mathbf{x}'_i$
Output: locally updated quadtree $\mathcal{T}$
Let $\mathcal{I}_i = \{\text{node } k \in \mathcal{T} : \mathbf{x}_i \in \mathcal{C}_k\}$ be the set of nodes along the path from root to leaf containing $\mathbf{x}_i$;
for *node $k \in \mathcal{I}_i$* **do**
  Cache $\bar{\mathbf{x}}_k$;
  Update $\bar{\mathbf{x}}_k \leftarrow \bar{\mathbf{x}}_k + \frac{(\mathbf{x}'_i - \mathbf{x}_i)}{N_k}$
return $\mathcal{T}$

Algorithm B.3: Deterministic evaluation of the BH-BMDS log-likelihood

Input: $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^2$; $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \mathbb{R}^\Omega$; quadtree $\mathcal{T}$
Output: BH-BMDS log-likelihood $\tilde{\ell}(\cdot \mid \mathcal{T})$
Function `BH`(i, k)
  Compute $\rho_{ik} = w_k / \bar{r}_{ik}$;
  if *node k is leaf* ***or*** $\rho_{ik} < \theta$ **then**
    return $N_k \left[\frac{(\bar{D}_{ik} - \bar{r}_{ik})^2}{2\sigma^2} + \log \Phi\left(\frac{\bar{r}_{ik}}{\sigma}\right) \right]$
  else
    return $\sum_{c \in \mathit{children}(k)}$ `BH`(i, c)
Initialize $\tilde{\ell}(\cdot \mid \mathcal{T}) = -N \log(2\pi\sigma^2)$;
for $i = 1, \ldots, n$ **do**
  $\tilde{\ell}(\mathbf{x}_i \mid \mathcal{T}) \leftarrow$ `BH`$(i, 0)$;
  $\tilde{\ell}(\cdot \mid \mathcal{T}) \leftarrow \tilde{\ell}(\cdot \mid \mathcal{T}) + \tilde{\ell}(\mathbf{x}_i \mid \mathcal{T})$;
return $\tilde{\ell}(\cdot \mid \mathcal{T})$

Algorithm B.4: Stochastic evaluation of the pointwise contribution to the BH-BMDS log-likelihood

Input: target point $\mathbf{x}_i$, quadtree $\mathcal{T}$
Output: $\tilde{\ell}(\mathbf{x}_i \mid \mathcal{T})$
Function `SBH`(i, k)
  Draw $u \sim \mathrm{Unif}(0, 1)$;
  Compute $p(\rho_{ik} \mid \theta)$ (Equation 4);
  if *node k is leaf* ***or*** $u < p(\rho_{ik} \mid \theta)$ **then**
    return $N_k \left[\frac{(\bar{D}_{ik} - \bar{r}_{ik})^2}{2\sigma^2} + \log \Phi\left(\frac{\bar{r}_{ik}}{\sigma}\right) \right]$
  else
    return $\sum_{c \in \mathit{children}(k)}$ `SBH`(i, c)
$\tilde{\ell}(\mathbf{x}_i \mid \mathcal{T}) \leftarrow$ `SBH(i, 0)`;
return $\tilde{\ell}(\mathbf{x}_i \mid \mathcal{T})$

C Additional plots

Figure C.1: Posterior Procrustes-aligned means across MCMC iterations for 1,000 and 10,000 embedded MNIST handwritten digital images under full Bayesian multidimensional scaling (BMDS) and Barnes–Hut Bayesian multidimensional scaling (BH-BMDS) frameworks.

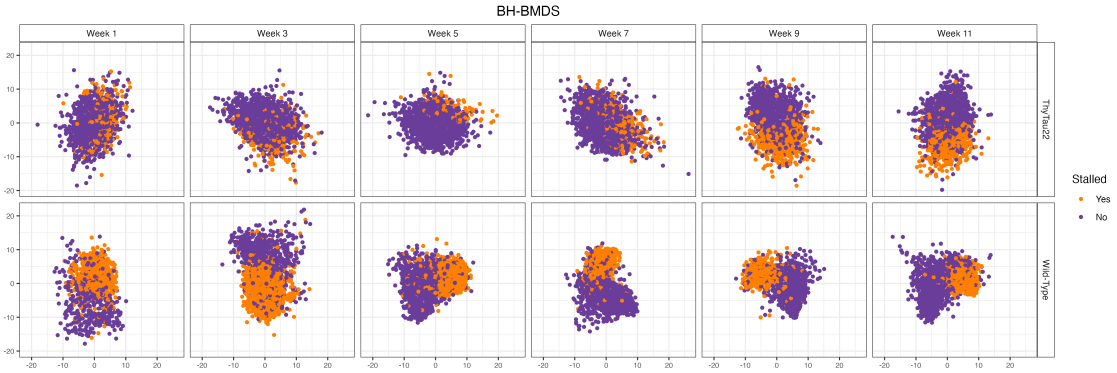

Figure C.2: Posterior Procrustes-aligned means across MCMC iterations for ThyTau22 and Wild-Type mouse under the Barnes–Hut Bayesian multidimensional scaling (BH-BMDS) framework. High-dimensional embeddings are obtain from a trained MaGNet GNN (Zhou et al., 2025) to decode behavior states (stalled vs. moving) on local field potential recordings. We observe more clear clustering for a Wild-Type than ThyTau22 mouse.