



## Sparse Bayesian multidimensional scaling(s)

Ami Sheth<sup>1</sup>  · Aaron Smith<sup>2</sup> · Andrew J. Holbrook<sup>1</sup>

Received: 9 June 2025 / Accepted: 26 September 2025  
© The Author(s) 2025

### Abstract

Bayesian multidimensional scaling (BMDS) is a probabilistic dimension reduction tool that allows one to model and visualize data consisting of dissimilarities between pairs of objects. Although BMDS has proven useful within, e.g., Bayesian phylogenetic inference, its likelihood and gradient calculations require burdensome  $\mathcal{O}(N^2)$  floating-point operations, where  $N$  is the number of data points. Thus, BMDS becomes impractical as  $N$  grows large. We propose and compare two sparse versions of BMDS (sBMDS) that apply log-likelihood and gradient computations to subsets of the observed dissimilarity matrix data. Landmark sBMDS (L-sBMDS) extracts columns, while banded sBMDS (B-sBMDS) extracts diagonals of the data. These sparse variants let one specify a time complexity between  $N^2$  and  $N$ . Under simplified settings, we prove posterior consistency for subsampled distance matrices. Through simulations, we examine the accuracy and computational efficiency across all models using both the Metropolis-Hastings and Hamiltonian Monte Carlo algorithms. We observe approximately 3-fold, 10-fold and 40-fold speedups with negligible loss of accuracy, when applying the sBMDS likelihoods and gradients to 500, 1000 and 5,000 data points with 50 bands (landmarks); these speedups only increase with the size of data considered. Finally, we apply the sBMDS variants to: (1) the phylogeographic modeling of multiple influenza subtypes to better understand how these strains spread through global air transportation networks and (2) the clustering of ArXiv manuscripts based on low-dimensional representations of article abstracts. In the first application, sBMDS contributes to holistic uncertainty quantification within a larger Bayesian hierarchical model. In the second, sBMDS approximates uncertainty quantification for a downstream modeling task.

**Keywords** Dimension reduction · Sparsity · Bayesian hierarchical modeling · Clustering · Phylogeography · Hamiltonian Monte Carlo

---

Extended author information available on the last page of the article

## 1 Introduction

Multidimensional scaling (MDS) is a dimension reduction technique that maps pairwise dissimilarity measurements corresponding to a set of  $N$  objects to a configuration of  $N$  points within a low-dimensional Euclidean space (Torgerson 1952). Classical MDS uses the spectral decomposition of a doubly centered matrix derived from the observed dissimilarity matrix to calculate the objects' coordinates. While classical MDS serves as a valuable data visualization tool, probabilistic extensions further enable uncertainty quantification in the context of Bayesian hierarchical models. Oh and Raftery (2001) propose a Bayesian framework for MDS (BMDS) under the assumption that the observed dissimilarities follow independent truncated normal probability density functions (PDFs). BMDS facilitates Bayesian inference of object configurations in a manner that is robust to violations of the Euclidean model assumption and dimension misspecifications (Oh and Raftery 2001). The key benefits of a Bayesian approach to MDS are that it provides uncertainty quantification for the projection itself and conditional distributions that can be easily integrated with other probability models, enabling fully model-based approaches to analyzing dissimilarity data. For example, one may incorporate BMDS into hierarchical modeling frameworks for Bayesian phylogeography (Bedford et al. 2014; Holbrook et al. 2021; Li et al. 2023), clustering (Oh and Raftery 2007) and variable selection (Lin and Fong 2019).

Bayesian phylogeography uses molecular data from species such as viruses, bacteria or pathogens to probabilistically model their evolution over both time and space (Lemey et al. 2009). For instance, one can reconstruct viral dispersion patterns to better understand the way viruses spread within and between human populations. The incorporation of BMDS within Bayesian phylogeography allows one to place dissimilarity data between species into a low-dimensional spatial representation while also considering their evolutionary dynamics from genetic data. Bedford et al. (2014) simultaneously characterize antigenic and genetic patterns of influenza by combining BMDS with an evolutionary diffusion process on the latent strain locations. They apply BMDS on hemmagglutination inhibition assay data to place the subtypes on a low-dimensional antigenic map. Holbrook et al. (2021) implement a similar Bayesian phylogenetic MDS model but perform phylogeographic inference on pairwise distances arising from air traffic data. Additionally, Li et al. (2023) use phylogenetic BMDS on pairwise distances stemming from hepaciviruses to infer the viral locations in a lower dimensional geographic and host space.

Unfortunately, BMDS is difficult to scale to big data settings; computing the BMDS log-likelihood and gradient each have  $\mathcal{O}(N^2)$  complexity. Bedford et al. (2014) partially circumvent this problem by assuming that the observed data follow non-truncated Gaussian distributions, thereby avoiding the costly floating-point operations necessary to evaluate the Gaussian cumulative density function (CDF) in the truncated normal PDFs (2). However, there are benefits to using the truncated normal distribution: it appropriately accounts for non-negative dissimilarities, and its variance term is always less than that of its corresponding non-truncated normal distribution, resulting in more precise posterior inference. Holbrook et al. (2021) mitigate BMDS's computational burden through massive parallelization using multi-core

central processing units, vectorization and graphic processing units. They obtain substantial performance gains, but parallelization requires expensive hardware. In either case, these models still scale quadratically in the number of objects. We therefore develop a framework that reduces the time complexity to  $\mathcal{O}(N)$  by inducing sparsity on the observed dissimilarity matrix. We perform experiments with simulated data and show that our sparse versions of BMDS (sBMDS) obtain significant speedups while preserving inferential accuracy. We then illustrate how one may use sBMDS within a larger hierarchical model by extending the types of phylogeographical models mentioned above under sparse assumptions, implementing sBMDS phylogenetic frameworks on air traffic data to analyze the geographic spread of four influenza subtypes. Additionally, we apply sBMDS to a collection of ArXiv paper abstracts and perform post-hoc clustering on posterior samples of the low-dimensional embeddings. Using a “bagged estimator” approach, we obtain posterior probabilities over cluster assignments.

In the following, we present two versions of sparse BMDS and prove that under simplistic conditions, the posterior latent locations are consistent for subsampled dissimilarity matrices (Sect. 2). In Sect. 3, we evaluate the empirical accuracy, sensitivity to model misspecification and computational performance of both methods. We apply sBMDS to the phylogeographic modeling of influenza variants and verify that we obtain similar migration rate estimates for both full and sparse BMDS models (Sect. 3.2). In Sect. 3.3, we apply sBMDS to a dataset representing ArXiv paper abstracts and recover the posterior probability that two manuscripts belong to the same subject-matter category. We conclude by summarizing our findings and discussing future research directions (Sect. 4).

## 2 Methods

### 2.1 Bayesian multidimensional scaling

Bayesian multidimensional scaling (BMDS) (Oh and Raftery 2001) models a set of  $N$  objects’ locations as latent variables in low-dimensional space under the assumption that the observed dissimilarity measures follow a prescribed joint probability distribution. To set notation: for  $A \subset \mathbb{R}$ , let  $N_A(\mu, \sigma^2)$  denote the Gaussian distribution truncated to  $A$ ; for  $k \in \mathbb{N}$ , let  $[k] = \{1, 2, \dots, k\}$ . Within BMDS, each observed dissimilarity measure  $\delta_{nn'}$  is the posited latent measure  $\delta_{nn'}^*$  plus a truncated Gaussian error:

$$\delta_{nn'} \sim N_{(0, \infty)}(\delta_{nn'}^*, \sigma^2), \quad n \neq n', \quad n, n' \in [N], \quad (1)$$

where  $\delta_{nn'}^* = \sqrt{\sum_{d=1}^D (x_{nd} - x_{n'd})^2}$  is the Euclidean distance between latent locations  $x_n, x_{n'} \in \mathbb{R}^D$ , and  $N(\cdot, \cdot)$  represents the normal distribution. These assumptions yield the log-likelihood function

$$\ell(\Delta, \sigma^2) = -\frac{m}{2} \log(2\pi\sigma^2) - \sum_{n < n'} \left[ \frac{(\delta_{nn'} - \delta_{nn'}^*)^2}{2\sigma^2} + \log \Phi\left(\frac{\delta_{nn'}^*}{\sigma}\right) \right], \quad (2)$$

where  $\Delta = \{\delta_{nn'}\}$  is the symmetric  $N \times N$  matrix of observed dissimilarities,  $m = N(N - 1)/2$  is the number of dissimilarities, and  $\Phi(\cdot)$  is the standard normal CDF.

Many MCMC algorithms, e.g., Hamiltonian Monte Carlo (HMC) (Sect. 2.4) and Metropolis-adjusted Langevin algorithm (MALA), use evaluations of gradients for efficient state space exploration. For this model, we take the first derivative of the log-likelihood function (2) with respect to a single row  $\mathbf{x}_n$  of  $\mathbf{X}$ , the  $N \times D$  matrix of unknown object coordinates to obtain the log-likelihood gradient function

$$\nabla_{\mathbf{x}_n} \ell(\Delta, \sigma^2) = - \sum_{\substack{n' \in [N], \\ n' \neq n}} \left[ \left( \frac{(\delta_{nn'}^* - \delta_{nn'})}{\sigma^2} + \frac{\phi(\delta_{nn'}^*/\sigma)}{\sigma \Phi(\delta_{nn'}^*/\sigma)} \right) \frac{(\mathbf{x}_n - \mathbf{x}_{n'})}{\delta_{nn'}^*} \right] \equiv - \sum_{\substack{n' \in [N], \\ n' \neq n}} \mathbf{r}_{nn'}. \quad (3)$$

Here  $\phi(\cdot)$  is the PDF of a standard normal variate, and  $\mathbf{r}_{nn'}$  is the contribution of the  $n'$ th location to the gradient with respect to the  $n$ th location.

The BMDS log-likelihood (2) and gradient (3) both involve summing  $\binom{N}{2}$  terms and require  $\mathcal{O}(N^2)$  floating point operations. Given the large number of calculations needed, they become computationally cumbersome as the number of objects grows large. Therefore, we propose using a small subset of the data for likelihood and gradient evaluations, namely the sparse BMDS methods (sBMDS).

## 2.2 Sparse likelihoods and their gradients

For each item  $n$ , let  $J_{n,N} \subset [N] \setminus \{n\}$  be an index set. We consider sparse coupling approaches resulting in log-likelihoods and log-likelihood gradients of the form

$$\ell_{sp}(\Delta, \sigma^2) = - \sum_{n=1}^N \sum_{\substack{n' \in J_{n,N}, \\ n' > n}} \left[ \frac{1}{2} \log(2\pi\sigma^2) + \frac{(\delta_{nn'} - \delta_{nn'}^*)^2}{2\sigma^2} + \log \Phi\left(\frac{\delta_{nn'}^*}{\sigma}\right) \right], \quad (4)$$

and

$$\nabla_{\mathbf{x}_n} \ell_{sp}(\Delta, \sigma^2) = - \sum_{n' \in J_{n,N}} \mathbf{r}_{nn'}. \quad (5)$$

We reduce the computational complexity of BMDS by including a small subset of couplings  $J_{n,N}$  per object  $n$ , where  $|J_{n,N}| \ll N$ . Here, we discuss two possible strategies for choosing  $J_{n,N}$ . By a slight abuse of notation, we use  $[a, b]$  to refer to a closed interval of either reals or integers, where the appropriate set should be obvious from context. The first option is to extract  $B \in [N - 1]$  off-diagonal bands of the observed

dissimilarity matrix such that  $J_{n,N} = [\max(1, n - B), \min(N, n + B)] \setminus \{n\}$  for all  $n$ . The second approach is to choose  $L \in [N]$  objects called “landmarks” and select each landmark’s dissimilarities from the remaining  $N - 1$  objects, e.g.,  $J_{n,N} = [N] \setminus \{n\}$  for  $n \in [L]$  and  $J_{n,N} = [L]$  for  $n \notin [L]$ . Essentially, this strategy retains a rectangular subset of the observed dissimilarity matrix by extracting  $L$  columns (rows) of the data. We note there is no loss of generality in taking the first  $n$  indices as landmarks rather than an arbitrary set because one can relabel the object indices without affecting the learned geometry. We refer to the first method as banded sBMDS (B-sBMDS) and the second method as landmark sBMDS (L-sBMDS). Alternative strategies for selecting index sets are possible, provided they satisfy Assumption 1 (Sect. 2.3). For instance, our B-sBMDS model assumes distances are measured on the real line and bands are defined as a contiguous interval. However, one could explore other forms of banded matrices, e.g., by selecting any set of entries in the distance matrix.

To highlight the difference, we consider a simplified scenario in which the number of objects is five, the latent dimension is two, the BMDS error variance  $\sigma^2$  is 0.25, and the observed dissimilarities are equal to the latent dissimilarity measures ( $\delta_{nn'} = \delta_{nn'}^*$ ). Given the distance and location matrices

$$\Delta = \begin{bmatrix} 0.00 & 1.35 & 2.53 & 0.99 & 1.85 \\ 1.35 & 0.00 & 1.54 & 0.76 & 0.50 \\ 2.53 & 1.54 & 0.00 & 1.54 & 1.26 \\ 0.99 & 0.76 & 1.54 & 0.00 & 1.12 \\ 1.85 & 0.50 & 1.26 & 1.12 & 0.00 \end{bmatrix}, \quad X = \begin{bmatrix} 0.59 & 0.71 \\ -0.11 & -0.45 \\ 0.61 & -1.82 \\ 0.63 & -0.28 \\ -0.28 & -0.92 \end{bmatrix},$$

we compare the sBMDS log-likelihood (Table 1) and gradient (Table 2) calculated from couplings defined by B-sBMDS versus L-sBMDS.

For B-sBMDS, the number of couplings is the number of elements in  $B$  bands. The relationship between the number of bands and number of couplings  $C$  is  $C = \sum_{b=1}^B (N - b)$ . We add one less coupling for each additional band. When the

**Table 1** We extract the  $(n, n')$  pair from the off-diagonals of the observed and latent dissimilarity matrices for banded sBMDS (B-sBMDS) versus the columns for landmark sBMDS (L-sBMDS)

B/L	B-sBMDS	L-sBMDS
Pairs $(n, n')$		
1	(1, 2); (2, 3); (3, 4); (4, 5)	(1, 2); (1, 3); (1, 4); (1, 5)
2	+ (1, 3); (2, 4); (3, 5)	+ (2, 3); (2, 4); (2, 5)
3	+ (1, 4); (2, 5)	+ (3, 4); (3, 5)
4	+ (1, 5)	+ (4, 5)
Log-likelihood values		
1	-0.885	-0.875
2	-1.490	-1.311
3	-1.743	-1.756
4	-1.969	-1.969

*B/L* refers to the number of bands ( $B$ ) or landmarks ( $L$ ), and the + symbol indicates all couplings above are also included. The bottom table shows the calculated log-likelihoods as the number of bands/landmarks increases. Importantly, the bottom log-likelihoods are equal for both sBMDS variants and correspond to the full BMDS log-likelihood

**Table 2** We extract the  $(n, n')$  pair from the observed and latent dissimilarity matrices to calculate the sBMDS gradients

Banded sBMDS				Landmark sBMDS			
Pairs $(n, n')$				Gradient			
1 band	2 bands	3 bands	4 bands	1 band	2 bands	3 bands	4 bands
$x_1(1, 2)$	$+(1, 3)$	$+(1, 4)$	$+(1, 5)$	$[-.010, .017]$	$[-.010, .018]$	$[-.005, .134]$	$[-.006, .135]$
$x_2(2, 3); (2, 1)$	$+(2, 4)$	$+(2, 5)$		$ [.014, .011]$	$ [.275, .074]$	$ [.071, -.468]$	$ [.071, -.468]$
$x_3(3, 4); (3, 2)$	$+(3, 5);$ $(3, 1)$			$[-.003, .013]$	$[-.026, .036]$	$[-.026, .036]$	$[-.026, .036]$
$x_4(4, 5); (4, 3)$	$+(4, 2)$	$+(4, 1)$		$[-.054, -.045]$	$[-.315, -.108]$	$[-.321, .009]$	$[-.321, .009]$
$x_5(5, 4)$	$+(5, 3)$	$+(5, 2)$	$+(5, 1)$	$ [.054, .038]$	$ [.077, .015]$	$ [.281, .557]$	$ [.281, .558]$

Pairs $(n, n')$				Gradient			
1 landmark	2 landmarks	3 landmarks	4 landmarks	1 landmark	2 landmarks	3 landmarks	4 landmarks
$x_1(1, 2 - 5)$				$[-.006, .135]$	$[-.006, .135]$	$[-.006, .135]$	$[-.006, .135]$
$x_2(2, 1)$	$+(2, 3$ $- 5)$			$ [.010, .017]$	$ [.071, -.468]$	$ [.071, -.468]$	$ [.071, -.468]$
$x_3(3, 1)$	$+(3, 2)$	$+(3, 4$ $- 5)$		$ [.000, .000]$	$[-.003, .006]$	$[-.026, .036]$	$[-.026, .036]$
$x_4(4, 1)$	$+(4, 2)$	$+(4, 3)$	$+(4, 5)$	$[-.005, .117]$	$[-.266, .054]$	$[-.266, .047]$	$[-.321, .009]$
$x_5(5, 1)$	$+(5, 2)$	$+(5, 3)$	$+(5, 4)$	$ [.000, .000]$	$ [.204, .543]$	$ [.227, .519]$	$ [.281, .558]$

On the left, the  $-$  symbol, as in  $(n, a - c)$ , indicates pairs  $(n, a)$ ;  $(n, b)$ ;  $(n, c)$ . For example, pair  $(3; 4 - 5)$  means we include both pair  $(3, 4)$  and  $(3, 5)$ . The  $+$  symbol indicates all couplings to the left are also included, and  $[\cdot, \cdot]$  represents a vector. On the right is the gradient computed for each  $x_n$  of  $X$  as function of the number of bands/landmarks. Extracting the entire column of a landmark point gives the full BMDS gradient in  $\mathbb{R}^D$  whereas banded sBMDS incrementally adds information to the row-wise gradients. Importantly, the rightmost gradients are equal for both sBMDS variants and correspond to the full BMDS gradient

number of bands equals  $N - 1$ , we return to the full BMDS case. Using a subset of the observed dissimilarity matrix reduces the burden of computing the BMDS likelihood and gradient to  $\mathcal{O}(NB)$ . Similar arguments hold for L-sBMDS, the likelihoods and gradients of which exhibit  $\mathcal{O}(NL)$  time complexity.

For classical MDS, an analogous strategy to L-sBMDS already exists. In MDS, the rate limiting step is the calculation of the top  $D$  eigenvalues and eigenvectors from a  $N \times N$  matrix. Silva and Tenenbaum (2004) propose applying classical MDS to  $L$  landmark points, e.g., an  $L \times N$  submatrix of the observed dissimilarity matrix, and then following a distance-based triangulation procedure to determine the remaining object coordinates. L-sBMDS uses the concept of randomly selecting  $L$  landmarks as well, but integrates them into the BMDS framework, allowing inference on the entire model. Raftery et al. (2012) approximate the likelihood of their network data by taking a random subset of objects deemed to have no link, reducing the time complexity from  $\mathcal{O}(N^2)$  to  $\mathcal{O}(N)$ . In the context of a very different network model, they incorporate an array of covariates to model the probability of a link between objects  $n$  and  $n'$  while our model is simpler, using no outside information to aid in determining locations in a latent space.

### 2.3 Posterior consistency

For the following theoretical development, we consider the model

$$\delta_{nn'} \sim N_{(0,M)}(\delta_{nn'}^*, \sigma^2), \quad n \neq n', \quad n, n' \in [N], \quad (6)$$

a generalization of (1) insofar as  $M$  can be any number within the interval  $(0, \infty)$ . Let the latent locations  $X$  be sampled from a range of values in the interval  $I$ . The posterior density function of the unknown parameters  $(X, I, \sigma^2, M)$  is proportional to  $\mathcal{L}(\Delta|X, \sigma^2, M)$ , the BMDs likelihood function of model (6), and the priors put on each auxiliary parameter, e.g.,

$$p(X, I, \sigma^2, M|\Delta) \propto \mathcal{L}(\Delta, \sigma^2, M|X) \times p(X|I) \times p(I) \times p(\sigma^2) \times p(M). \quad (7)$$

The marginal posterior density function of  $X$  is

$$p(X|\Delta) = \int p(X, I, \sigma^2, M|\Delta) dI d\sigma^2 dM. \quad (8)$$

We examine the posterior consistency of subsampled dissimilarity matrices under simple conditions. Fixing some interval  $I$ , we sample points  $x_1, \dots, x_N \stackrel{i.i.d.}{\sim} N_I(0, 1)$ . Let  $\Delta^*$  be the Euclidean distance matrix with entries  $\delta_{nn'}^* = |x_n - x_{n'}|$  and  $\delta_{nn'}$  be the noisy observations sampled from the truncated normal model (6). We set a prior on  $I, M$  and  $\sigma^2$  that has compact support and is bounded away from 0 and infinity on its support, e.g., Oh and Raftery (2001), Oh and Raftery (2007). In addition, we fix in advance a collection of indices  $J_{n,N} \subset [N] \setminus \{n\}$  of observations to keep for each object  $n$ , treating this choice as non-random in the following. Next, we make some assumptions about which observations are kept.

**Assumption 1** Fix  $K \in \mathbb{N}$ . Assume there exists a sequence  $\{\ell_N\}_{N \in \mathbb{N}}$  and a collection of partitions  $\{G_{n,N}^{(k(n))}\}_{k=1}^K$  of  $J_{n,N}$  with the following properties:

1. For all  $n \in [N]$  and  $k \in [K]$ ,  $|G_{n,N}^{(k(n))}| \geq \ell_N$ .
2. Say  $n, n' \in [N]$  are linked by an edge if there exists  $k(n), k(n') \in [K]$  so that

$$|J_{n,N} \cap J_{n',N} \setminus (G_{n,N}^{(k(n))} \cup G_{n',N}^{(k(n'))})| \geq \ell_N. \quad (9)$$

Assume that the graph with these edges and vertex set  $[N]$  is a connected graph.

3. The sequence  $\ell_N$  satisfies

$$\lim_{N \rightarrow \infty} \frac{\ell_N}{\log(N)^2} = \infty. \quad (10)$$

**Remark 1** We verify that Assumption 1 holds for  $L$ -sBMDS given  $L \geq (\ell_N + 1)K$  landmarks and  $L \ll N$ . We can think of  $\ell_N$  as the number of retained entries in the sparsest row (up to a universal constant). Let  $\ell_N = \lceil \frac{\sqrt{N}}{2} \rceil$ , so that it satisfies part 3 of Assumption 1. For all  $n \in [N]$ ,  $|J_{n,N}| \geq L$  and for objects  $n, n' \in [N]$ ,  $|J_{n,N} \cap J_{n',N}| \geq L - 1$ . Consider the partition  $G_{n,N}^{(k(n))} = [\lfloor \frac{(k-1)L}{K} \rfloor + 1, \lfloor \frac{kL}{K} \rfloor] \setminus \{n\}$ . Then for any object  $n, n' \in [N]$ ,  $G_{n,N}^{(k(n))} \cup G_{n',N}^{(k(n'))} = \{\lfloor \frac{(k-1)L}{K} \rfloor + 1, \dots, \lfloor \frac{kL}{K} \rfloor\} \equiv G_N^{(k)}$  is independent of  $n, n'$  and of size  $\frac{L}{K}$ , which satisfies part 1 of Assumption 1. Thus in the most conservative case, e.g.,  $n \in [L]$  and  $n' \notin [L]$ ,

$$\begin{aligned} |J_{n,N} \cap J_{n',N} \setminus (G_{n,N}^{(k(n))} \cup G_{n',N}^{(k(n'))})| &= |\{1, \dots, n-1, n+1, \dots, L\} \setminus G_N^{(k)}| \\ &= (L-1) - \frac{L}{K} = \frac{L(K-1)}{K} - 1. \end{aligned}$$

As a result,  $|J_{n,N} \cap J_{n',N} \setminus (G_{n,N}^{(k(n))} \cup G_{n',N}^{(k(n'))})| \geq \ell_N$  when  $L \geq (\ell_N + 1)K$  for

$K > 1$ . Notably, the graph in part 2 of Assumption 1 is connected.

**Remark 2** Similarly, we verify that Assumption 1 holds for  $B$ -sBMDS given  $B \geq 2\ell_N + 1$  bands and  $B \ll N$ . Again, let  $\ell_N = \lceil \frac{\sqrt{N}}{2} \rceil$ , so that it satisfies part 3 of Assumption 1. Under  $B$ -sBMDS, the ends of a distance matrix have the fewest indices. To ensure that the size of each partition is at least  $\ell_N$  at these boundaries, let the number of partitions be  $K = \lfloor \frac{B}{\ell_N} \rfloor$ , satisfying part 1 of Assumption 1. For all  $n \in [N]$ ,  $|J_{n,N}| \geq B$  and for two consecutive objects  $(n < n') \in [N]$ ,  $|J_{n,N} \cap J_{n',N}| \geq B - 1$ . To remove the minimal number of common indices between two consecutive objects, let  $G_{n,N}^{(k(n))} = \{n+1, \dots, n+\ell_N\}$  so that  $|G_{n,N}^{(k(n))}| = \ell_N$ . Then,  $G_{n,N}^{(k(n))} \cup G_{n',N}^{(k(n'))} = \{n', n'+1, \dots, n'+\ell_N\}$  and the cardinality of the intersection is  $\ell_N + 1$ . Finally,

$$\begin{aligned} |J_{n,N} \cap J_{n',N} \setminus (G_{n,N}^{(k(n))} \cup G_{n',N}^{(k(n'))})| &\geq |\{n'+1, \dots, n+B\} \setminus \{n', n'+1, \dots, n'+\ell_N\}| \\ &\geq |\{n'+\ell_N+1, \dots, n+B\}| \geq (B-1) - \ell_N \end{aligned}$$

because we remove  $\ell_N$  elements from the union containing at least  $B - 1$  elements. Thus when  $B \geq 2\ell_N + 1$ ,  $|J_{n,N} \cap J_{n',N} \setminus (G_{n,N}^{(k(n))} \cup G_{n',N}^{(k(n'))})| \geq \ell_N$ , and we obtain a connected graph, fulfilling part 2 of Assumption 1.

Under Assumption 1, we have the following posterior consistency result.

**Theorem 1** Fix  $0 < \alpha < 0.1$  and  $K \in \mathbb{N}$ . Let the sequences  $\{J_{n,N}\}$ ,  $\{G_{n,N}^{(k(n))}\}$  and  $\{\ell_N\}$  satisfy Assumption 1. Let  $\epsilon_N = \ell_N^{-0.5+\alpha}$ . Let  $(x_1^{(N)}, \dots, x_N^{(N)}) \stackrel{i.i.d.}{\sim} N_I(0, 1)$  and let  $\{\delta_{nn'}^{(N)}\}_{1 \leq n < n' \leq N}$  be sampled from model (6). Finally, let

$(\tilde{x}_1^{(N)}, \dots, \tilde{x}_N^{(N)}) \sim p(\cdot | \{\delta_{nn'}^{(N)}\}_{n' \in J_{n,N}})$  be sampled from the associated marginal posterior distribution of the model. Then there exists  $C > 0$  so that the event

$$\{\forall n \in [N], |x_n^{(N)} - \tilde{x}_n^{(N)}| < C\epsilon_N\} \quad (11)$$

occurs asymptotically almost surely.  $\square$

**Proof** See Appendix A.

Theorem 1 establishes posterior consistency for latent locations estimated from a subsampled dissimilarity matrix as we are able to recover the estimated latent locations  $\tilde{x}_n^{(N)}$  up to an additive error of  $\mathcal{O}(\epsilon_N)$  relative to the true latent locations  $x_n^{(N)}$ . Intuitively, each observed distance constrains the feasible location of a point, and with  $J_{n,N}$  such distances, the uncertainty in its position concentrates to a ball of radius  $\mathcal{O}(J_{n,N}^{-1/2})$ . Under Part 1 of Assumption 1, every point satisfies  $J_{n,N} \geq \ell_N$  with  $\ell_N \rightarrow \infty$ , so the worst-case uncertainty scales as  $\mathcal{O}(\ell_N^{-1/2})$ . Theorem 1 formalizes this heuristic by showing posterior concentration at rate  $\epsilon_N = \ell_N^{-0.5+\alpha}$ , matching the  $\ell_N^{-1/2}$ . Thus, as long as the number of observations per point grows with the sample size, the feasible regions shrink asymptotically, yielding consistent estimators of latent positions. Our result parallels the theoretical guarantees of landmark MDS (Silva and Tenenbaum 2004) but in a Bayesian setting. We acknowledge that the main limitation of this proof is the assumption that we have one-dimensional latent objects. See Sect. A.4 for a short discussion of how similar results may be obtained in fixed dimensions greater than 1.

## 2.4 Bayesian computation

Bayesian hierarchical models under the BMDS framework have previously been fit using MCMC algorithms such as Metropolis-Hastings (MH) (Metropolis et al. 1953; Hastings 1970; Oh and Raftery 2001; Bedford et al. 2014) and HMC (Neal 2012; Holbrook et al. 2021). In the following, we experiment with MH and HMC to perform posterior inference with the sBMDS models.

Let  $\theta$  be the random variable of interest and  $\pi(\theta)$  the target distribution. Under MH, a new candidate  $\theta^*$  is sampled from a proposal distribution centered at the value of the current iteration  $s$ ,  $q(\theta^* | \theta^{(s)})$ . One then accepts the candidate with probability

$$\alpha(\theta^* | \theta^{(s)}) = \min \left[ 1, \frac{\pi(\theta^*) q(\theta^{(s)} | \theta^*)}{\pi(\theta^{(s)}) q(\theta^* | \theta^{(s)})} \right]. \quad (12)$$

In the BMDS model (1), the parameters of interest are the latent locations  $X$  and the error variance  $\sigma^2$ , and—within a larger Metropolis-within-Gibbs scheme—the target distributions of interest are their respective conditional posterior distributions.

For our MH-based experiments, we jointly draw each candidate object's latent location  $x_n^*$  from the normal proposal distribution,  $N(x_n^{(s)}, \tau^2)$ , in which the proposal standard deviation  $\tau$  is a tuning parameter. In practice, we find it beneficial to adjust  $\tau$  in a manner that satisfies the diminishing adaptations criterion of Roberts and Rosenthal (2001). Specifically, the acceptance ratio is the number of acceptances in a given sample bound. If the acceptance ratio exceeds the target acceptance ratio, we multiplicatively increase  $\tau$  by  $(1 + \min(0.01, 1/\sqrt{s-1}))$ ; otherwise we multiplicatively decrease  $\tau$  by  $(1 - \min(0.01, 1/\sqrt{s-1}))$ .

For BMDS and its sparse variants, the dimension of the state space grows with the number of objects. Because MH typically breaks down in high-dimensions, we also consider HMC to infer the latent locations. HMC allows one to generate a Markov chain with distant proposals that nonetheless have a high probability of acceptance. It combines a fictitious momentum variable,  $P$ , along with a position variable to create a Hamiltonian system from which we compute the trajectories necessary for state space exploration. The position variable represents the parameters of the target distribution, so in the context of our model, we let the position variable be the latent locations  $X$ . The Hamiltonian function is

$$H(X, P) = U(X) + K(P) \quad (13)$$

where  $U(X)$  is the potential energy defined as the negative log target density, and  $K(P)$  is the kinetic energy defined as  $K(P) = \text{tr}(P^T P)/2$ . The partial derivatives of the Hamiltonian dictate how  $P$  and  $X$  change over time  $t$ :

$$\frac{dX}{dt} = \frac{\partial H(X, P)}{\partial P} = P, \quad \frac{dP}{dt} = \frac{-\partial H(X, P)}{\partial X} = -\nabla_X \ell(\Delta, \sigma^2). \quad (14)$$

For computer implementation, these equations are discretized over time using some small stepsize  $\epsilon$ . We follow Neal (2012) and implement the leapfrog method to numerically integrate Hamilton's equations (14). We tune the stepsize in the same way we change the proposal standard deviation in the adaptive MH algorithm. To propose a new state, we sample an initial momentum variable  $P_0$  and numerically integrate Hamilton's equations with initial state,  $(X^{(s)}, P_0)$ . We then accept the proposed state,  $(X^*, P^*)$ , according to the Metropolis-Hastings-Green (Green 1995; Geyer 2011) probability of

$$\min \left[ 1, \exp(-H(X^*, P^*) + H(X^{(s)}, P_0)) \right] = \min \left[ 1, \exp(-U(X^*) + U(X^{(s)}) - K(P^*) + K(P_0)) \right]. \quad (15)$$

Measured on an iteration by iteration basis, HMC allows for faster exploration of state spaces, especially in higher dimensions, compared to MH (Neal 2012; Beskos et al. 2013). However, HMC is computationally more expensive because it requires the gradient of the target function within every iteration of the leapfrog method. Recall that these gradient evaluations scale  $\mathcal{O}(N^2)$  for BMDS. If we want to learn the BMDS error variance  $\sigma^2$  as well, we again follow the adaptive MH algorithm, drawing a candidate  $\sigma^{2*}$  from a truncated normal proposal distribution with the current iteration's  $\sigma^{2(s)}$  as the mean and a standard deviation with the same adaption scheme

as described above. We account for the asymmetric proposal distribution within the MH acceptance probability (12).

### 3 Results

We explore the accuracy of full and sparse BMDS as well as the computational efficiency of all models in the context of the MH and HMC algorithms. The code for this project is available on Github ([https://github.com/andrewjholbrook/sparseBMD\\_S](https://github.com/andrewjholbrook/sparseBMD_S)). For visualization, we use the `ggplot2` (Wickham 2016) package in R (R Core Team 2023).

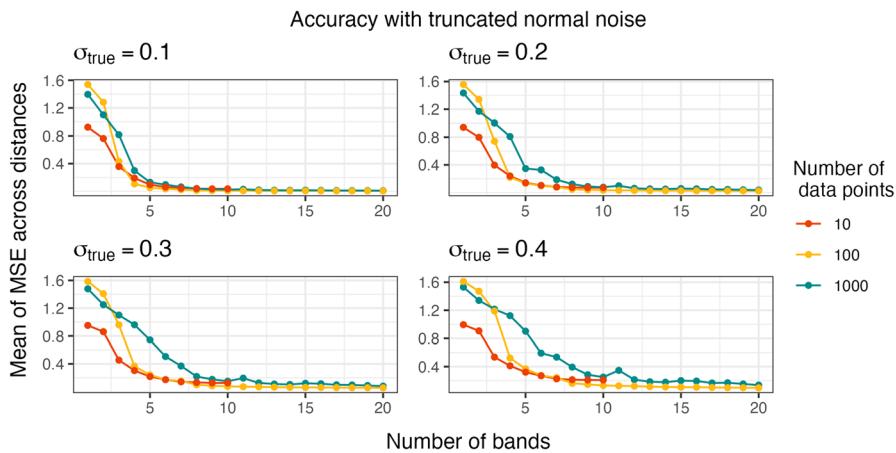
#### 3.1 Simulation studies

For a full Bayesian analysis, we put a D-dimensional multivariate normal distribution with mean 0 and diagonal covariance matrix  $\Lambda$  as the prior for  $x_n$ , independently for  $n = 1, \dots, N$ . The prior for the BMDS error variance  $\sigma^2$  is an inverse gamma with rate  $a$  and shape  $b$ . One can define hyperpriors for  $\Lambda, a, b$ , but we assume those parameters are fixed and known in this section. For our simulations, we set  $\Lambda$  equal to the identity  $I_2$ ,  $a = 1$  and  $b = 1$  so that  $x_n \sim N(0, I_2)$  and  $\sigma^2 \sim IG(1, 1)$ . To create the observed dissimilarity matrix  $\Delta = \{\delta_{nn'}\}$ , we add i.i.d. noise using a truncated normal distribution with mean 0 and variance  $\sigma_{true}^2$  to a “true” distance matrix. For the “true” distance matrix,  $\Delta^{(true)} = \{\delta_{nn'}^{(true)}\}$ , we generate a  $N \times 2$  “true” location matrix  $X$  from standard normal distributions and use  $X$  to calculate the Euclidean distance between pairs  $(n, n')$ .

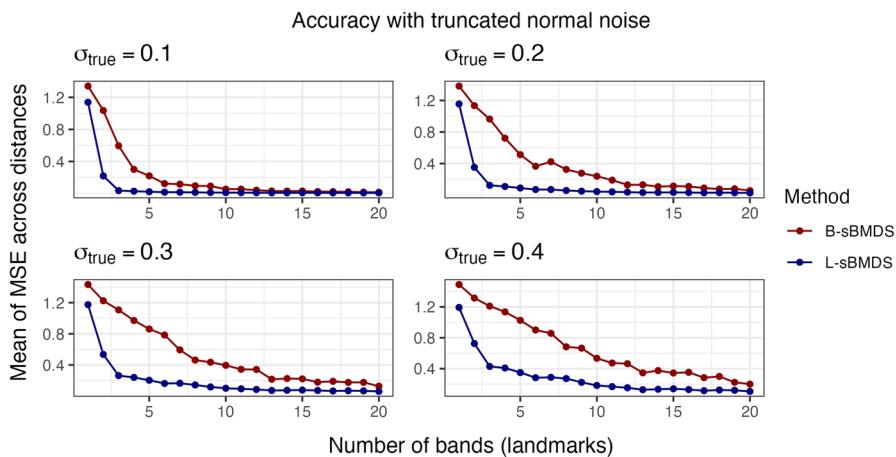
##### 3.1.1 Accuracy

We test the accuracy of the sBMDS models by comparing the simulated “true” dissimilarities to those obtained from HMC using the sBMDS posteriors and gradients. Given  $S$  iterations, we calculate the mean of the mean squared error ( $\overline{MSE}$ ) as  $\overline{MSE} = \frac{1}{Sm} \sum_{s=1}^S \sum_{n \neq n'} (\delta_{nn'}^{*(s)} - \delta_{nn'}^{(true)})^2$  where  $\delta_{nn'}^{*(s)}$  is the Euclidean distance calculated from the inferred locations of object  $n$  and object  $n'$  at iteration  $s$ ,  $\delta_{nn'}^{(true)}$  is the “true” Euclidean distance, and  $m = N(N - 1)$  is the number of dissimilarities. We compare distances instead of locations because the locations are not identifiable under distance preserving transformations. For computational convenience, when the number of objects is greater than 1,000, we randomly sample 1,000 distances to calculate  $\overline{MSE}$ . We set  $\sigma_{true}$  to either 0.1, 0.2, 0.3 or 0.4 to change noise levels and run 110,000 iterations, discarding the first 10,000 as burn-in and retaining every 100th iteration. We establish the initial conditions of the latent locations within HMC from classical MDS output.

Figure 1 plots  $\overline{MSE}$  as function of the number of bands for data with 10, 100 and 1,000 data points at varying levels of noise (see Appendix B, Fig. 12 for landmark results). Likewise, Fig. 2 plots  $\overline{MSE}$  as function of the number of bands/



**Fig. 1** The mean of the mean squared error (MSE) across all distances using 1 to 10 bands for 10 data points and 1 to 20 bands for 100 and 1000 data points. We estimate Euclidean distances from the inferred locations obtained using an adaptive Hamiltonian Monte Carlo algorithm under banded sparse Bayesian multidimensional scaling (B-sBMDS).  $\sigma_{true}^2$  is the variance component of the truncated normal noise centered at 0 added to the “true” distance matrix such that  $\sigma_{true}$  corresponds to the BMDS error standard deviation  $\sigma$



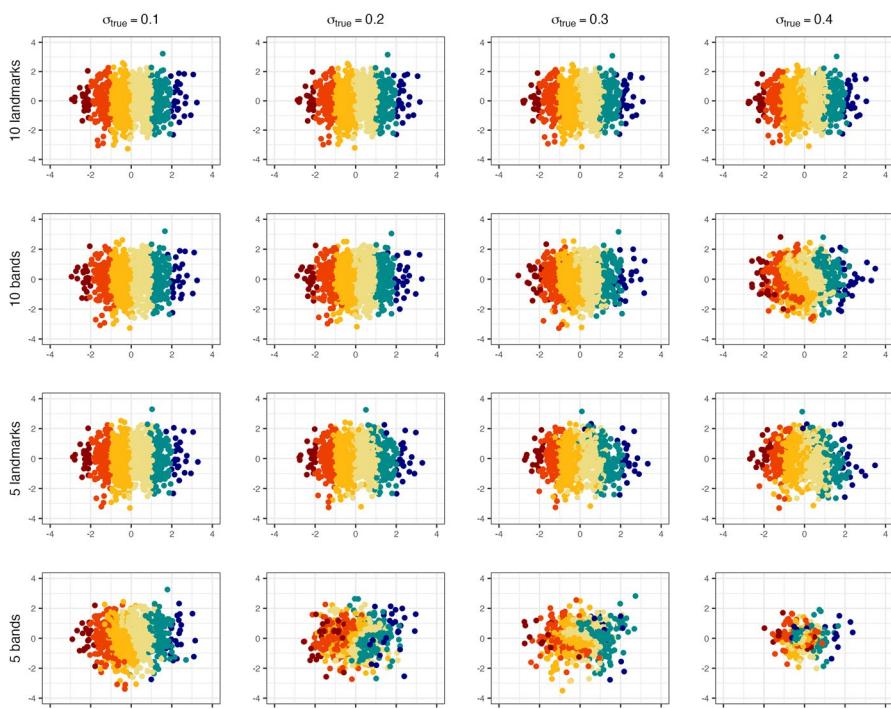
**Fig. 2** The mean of the mean squared error (MSE) across 1000 distances, randomly sampled from distance matrices with 10,000 data points. We estimate Euclidean distances from the inferred locations obtained using an adaptive Hamiltonian Monte Carlo algorithm under both sparse Bayesian multidimensional scaling (sBMDS) variants, banded sBMDS (B-sBMDS) and landmark sBMDS (L-sBMDS) with 1 to 20 bands/landmarks.  $\sigma_{true}^2$  is the variance component of the truncated normal noise centered at 0 added to the “true” distance matrix such that  $\sigma_{true}$  corresponds to the BMDS error standard deviation  $\sigma$

landmarks for data with 10,000 data points under B-sBMDS and L-sBMDS at different noise levels. In both figures, all the plots have identifiable elbows, demonstrating that a small number of bands/landmarks is sufficient to achieve low error. While we need more bands for noisier data, the amount is still modest compared to the number

of objects. Interestingly, we detect an elbow earlier for L-sBMDS than B-sBMDS; L-sBMDS recovers accurate pairwise relationships more efficiently than B-sBMDS. We visually see this difference in Fig. 3. In this simulation, we generate 1000 data points using the same sampling scheme and color-code the x-axis of the “true” locations. After running 110,000 HMC samples, we plot the Procrustes-aligned mean of the inferred latent locations from B-sBMDS and L-sBMDS using 5 and 10 bands/landmarks. From Fig. 3, we observe that while L-sBMDS maintains the integrity of the latent locations, B-sBMDS rapidly loses its accuracy as noise increases for 10 bands and is no longer accurate for 5 bands.

### 3.1.2 Sensitivity to model misspecification

To observe how the sparse variants behave under model misspecifications, we explore two possible situations: 1) a mismatch between the true dimensionality and that specified by the scientist and 2) heavy-tailed, rather than truncated normal, noise. For case 1, we vary the dimension of the “true” location matrix from 2 to 10 while fixing the embedding dimension to 2 and the truncated Gaussian noise variance  $\sigma_{true}^2$  to 0.2.



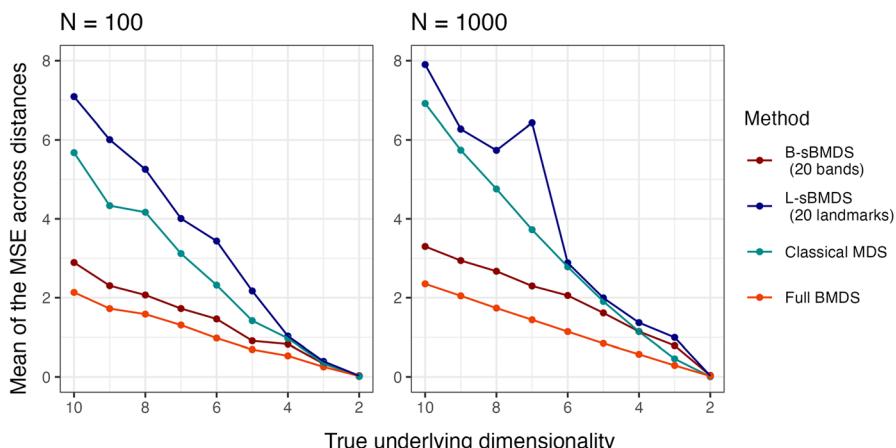
**Fig. 3** Procrustes-aligned means of the inferred locations across 100,000 iterations under the B-sBMDS and L-sBMDS frameworks when the number of bands/landmarks is ten and five. The number of data points is 1000. We simulate the latent locations from a two-dimensional standard normal distribution and assign a color according to their x-coordinate.  $\sigma_{true}^2$  is the variance component of the truncated normal noise centered at 0 added to the “true” distance matrix such that  $\sigma_{true}$  corresponds to the BMDS error standard deviation  $\sigma$

As expected,  $\overline{\text{MSE}}$  decreases as the true underlying dimensionality approaches the embedding dimension. Full BMDS and B-sBMDS with 20 bands are more robust to dimension misspecification than classical MDS, and the accuracy for B-sBMDS with 20 bands closely matches that of full BMDS (Fig. 4).

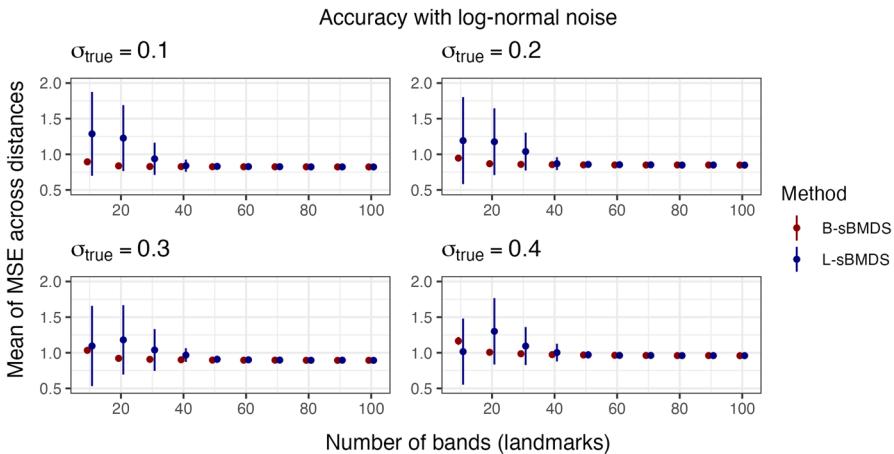
For case 2, we assume a correctly-specified-dimensional Euclidean space, but add i.i.d log-normal noise to the “true” distance matrix. We bootstrap the  $\overline{\text{MSE}}$  across all distances from 100 data points, a 100 times and plot the mean of  $\overline{\text{MSE}}$  along with error bars representing  $\pm$  the standard deviation of  $\overline{\text{MSE}}$ . Figure 5 demonstrates that, even with heavy-tailed data, both sparse variants achieve comparable  $\overline{\text{MSE}}$  s to full BMDS’s at a low number of bands/landmarks. In both cases, we observe that B-sBMDS seems to be less sensitive to model misspecification than L-sBMDS. We further explore the sensitivity of our sparse methods to modeling assumptions in Appendix B, specifically the prior choice (Fig. 15), the number of bands/landmarks relative to dimensionality (Fig. 16) and the assumed likelihood for the dissimilarities (Fig. 17).

### 3.1.3 Computational performance

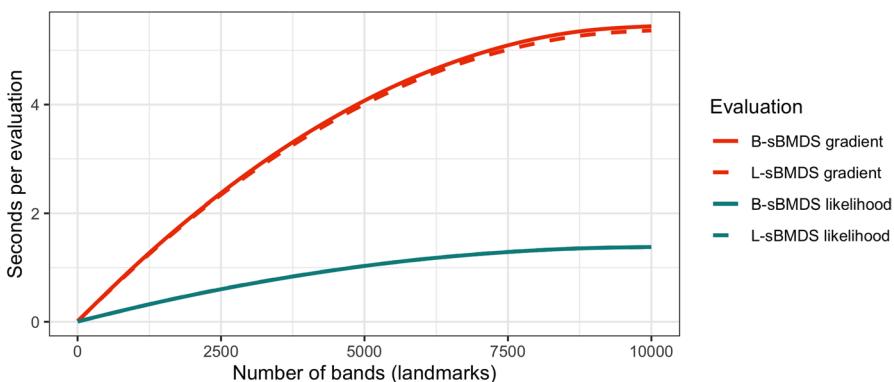
To better understand the computational benefits of the sBMDS variants, we first calculate the log-likelihood and log-likelihood gradient using B-sBMDS and L-sBMDS for a 10,000 by 10,000 Euclidean distance matrix. Recall that the number of couplings decreases per additional band/landmark. As a result, we see a parabolic-like relationship between evaluation time (in seconds) and the number of bands/landmarks (Fig. 6). If we were to plot the number of couplings vs seconds per evaluation, we would observe linear associations instead. When the number of bands/landmarks



**Fig. 4** The mean of the mean squared error (MSE) across 20 bands/landmarks for 100 and 1,000 data points. We vary the dimension space of the “true” latent locations while fixing the latent dimensionality to two. We estimate Euclidean distances from the inferred locations obtained using an adaptive Hamiltonian Monte Carlo algorithm under the B-sBMDS, L-sBMDS and full BMDS frameworks. Additionally, we compare the mean MSE across all distances from the inferred locations using classical MDS

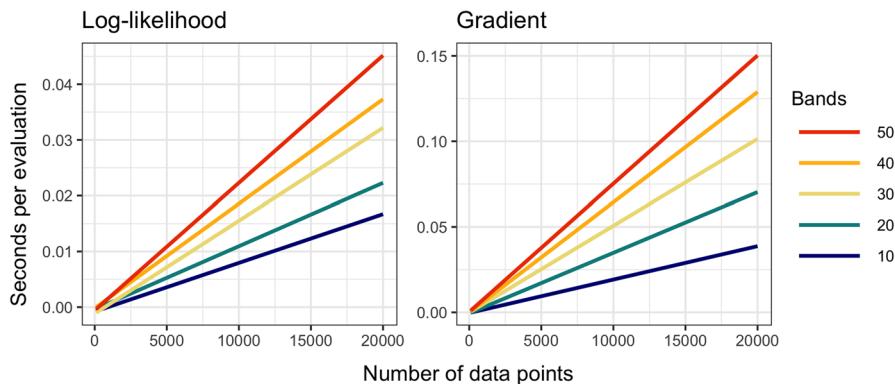


**Fig. 5** The average mean of the mean squared error (MSE) across all distances from 100 data points evaluated at intervals of 10 bands/landmarks, from 10 to 100, repeated 100 times. The dot is the average mean of MSE, and the error bars are  $\pm$  one standard deviation away from this mean. We estimate Euclidean distances from the inferred locations obtained using an adaptive Hamiltonian Monte Carlo algorithm under both sparse Bayesian multidimensional scaling (sBMDS) variants, banded sBMDS (B-sBMDS) and landmark sBMDS (L-sBMDS). When the number of bands (landmarks) equals 99, we return to full BMDS.  $\sigma_{true}^2$  is the variance component of the log-normal noise centered at 0 added to the “true” distance matrix such that the distribution of the observed distance matrix has heavy-tails



**Fig. 6** Time elapsed to calculate the sparse BMDS (B-sBMDS and L-sBMDS) likelihoods (cyan) and gradients (red) as a function of the number of bands/landmarks when the number of data points is 10,000. The seconds per evaluation at 9999 bands/landmarks correspond to the time it takes to calculate the full BMDS likelihoods and gradients. The parabolic curve is due to the number of couplings decreasing per additional band/landmark, causing the differences in computational time to reduce as well. If we plot the number of couplings vs seconds per evaluation, we would observe strictly linear associations

is 9,999, we return to the full case. We observe likelihood (gradient) speedups of 457-fold (773-fold), 91-fold (71-fold), 7-fold (10-fold) and 1.3-fold (1.3-fold) for 5, 50, 500 and 5000 bands (landmarks); there appears to be negligible time differences between B-sBMDS and L-sBMDS. Figure 7 emphasizes this correspondence



**Fig. 7** Time elapsed to calculate the banded sparse Bayesian multidimensional scaling (B-sBMDS) likelihood and gradient using  $B$  bands as a function of the number of data points

between speedups and number of bands, demonstrating the performance gains using a small number of bands relative to the number of objects. We observe approximately 3-fold, 10-fold and 40-fold speedups when applying the sBMDS likelihoods and gradients to 500, 1,000 and 5,000 data points with 50 bands/landmarks. We only scale up to 50 bands because these are reasonable band counts to achieve high accuracy (Figs. 1 and 2). We see similar patterns for landmarks in Fig. 13 (Appendix B).

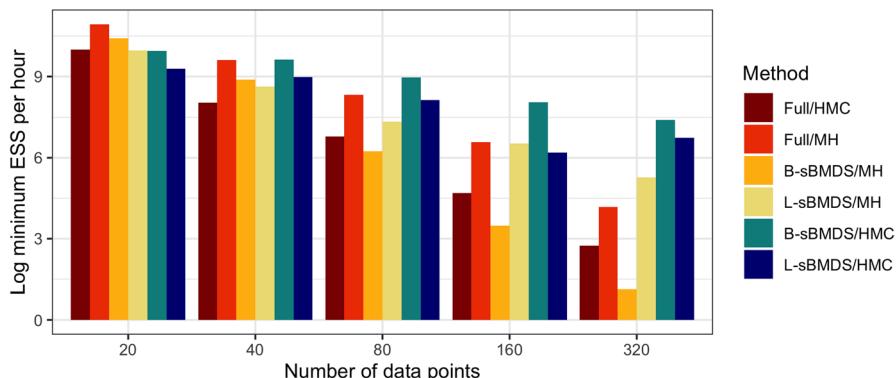
To compare computational performances, we set  $\sigma_{true} = 0.2$ , a value that allows us to establish accurate results while obtaining high acceptance probabilities. We fix the number of bands/landmarks to 10 based on the findings from both Fig. 1 and 12,

which confirm that this number ensures high model accuracy when  $\sigma_{true} = 0.2$  and  $N < 1,000$ . We then conduct MH and HMC under the full BMDS, B-sBMDS, and L-sBMDS models. For a fair comparison, we run all chains until the minimum effective sample size (ESS) is at least 100. ESS is a function of asymptotic auto-correlation,  $ESS = \frac{S}{1+2 \sum_{t=1}^{\infty} \rho_t}$ , where  $\rho_t$  is the autocorrelation between samples separated

by a lag of  $t$  timesteps and  $S$  is the length of a time series input. We calculate ESS using the `coda` package (Plummer et al. 2006) in R. We define efficiency as the minimum ESS per hour and take the natural log of it to allow comparison across scales. Figure 8 compares efficiency across the three models and two MCMC algorithms. The sBMDS variants under HMC outperform the others even in moderately high dimensions. MH begins to break down as the number of data points increases because, while it is computationally faster than HMC, the large dimension of the state space prevents efficient exploration, leading to high auto-correlation and low ESS values.

### 3.2 Analysis of global influenza

When incorporated into a larger Bayesian hierarchical model, sBMDS provides dimension reduction that propagates and accounts for total model uncertainty. To demonstrate the robustness and applicability of our sparse frameworks, we integrate them within a Bayesian hierarchical model for analyzing the global spread of



**Fig. 8** Computational performance measured as the logarithm of the minimum effective sample size (ESS) per hour across different frameworks and number of data point. In the legend, the half before the slash corresponds to the model type. “Full” is the BMDS model; “B-sBMDS” (banded sBMDS) and “L-sBMDS” (landmark sBMDS) are the sparse models using 10 bands/landmarks. The latter half explains the MCMC algorithm used for posterior inference. HMC abbreviates for Hamiltonian Monte Carlo and MH for Metropolis-Hastings

influenza. Every year seasonal influenza affects millions of adults, resulting in about 140,000 to 710,000 influenza-related hospitalizations in the United States alone (Rolfes et al. 2018). The virus’s ability to constantly evolve makes understanding its viral patterns so important for managing prevalence. The use of easily accessible mobility data can improve the readiness in which we learn about viral epidemics. Holbrook et al. (2021) apply the BMDS framework to a phylogeographic analysis of the spread of influenza subtypes through transportation networks. They analyze 1,370, 1,389, 1,393 and 1,240 samples of type H1N1, H3N1, Victoria (VIC) and Yamagata (YAM), spanning 12.9, 14.2, 15.4 and 17.75 years, respectively. To scale BMDS to data of this size, they implement core model likelihood and log-likelihood gradient calculations on large graphics processing units and multi-core central processing units. Unfortunately, such an approach requires time-intensive coding and access to expensive computational hardware. We employ a similar Bayesian hierarchical model, applying the same highly structured stochastic process priors but use sBMDS to transform to a latent network space. We are interested in whether under sBMDS we can accurately and efficiently infer the subtype-specific rates of dispersal across the latent airspace for the four influenza strains.

Our data consists of pairwise “effective distances” (Brockmann and Helbing 2013) between countries, which inversely measures the probability of traveling between airports. More trafficked airports have a shorter “distance” and thus a higher chance of disease transmission. Effective distances are better at predicting disease arrival times and spread compared to geographical distances because they incorporate the underlying mobility network (Brockmann and Helbing 2013). For each influenza subtype, we apply sBMDS to their air traffic data with the following priors for the unknown parameters,  $X_v$ ,  $\sigma_v^2$ , and hyperparameter  $\Sigma_v$ . For strain  $v$ , the prior on the viral latent locations  $X_v$  follows a multivariate Brownian diffusion process along the tree

$$X_v \sim MN(\mu_v, V_{\mathcal{G}_v}, \Sigma_v), \quad (16)$$

in which  $\mu_v$  is the  $N \times D$  mean matrix,  $V_{\mathcal{G}_v}$  is the  $N \times N$  row covariance matrix calculated from a fixed tree  $\mathcal{G}_v$ , and  $\Sigma_v$  is the  $D \times D$  column covariance matrix, independently for  $v = 1, \dots, 4$ . For viral diffusion,  $\Sigma_v$  describes how the virus's location in geographic space covary over lineages. In addition, we assume a priori

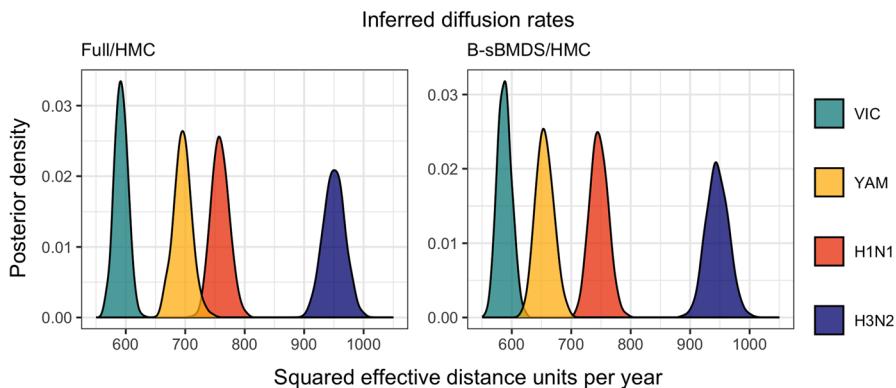
$$\Sigma_v^{-1} \sim Wishart(d_0, T_0) \quad (17)$$

$$\sigma_v^{-2} \sim Gamma(1, 1). \quad (18)$$

$d_0$  is the degree of freedom set as the dimension of the latent space and  $T_0$  is the rate matrix fixed as  $I_D$  in our model. The trace of  $\Sigma_v$  provides the instantaneous rate of diffusion and is of chief scientific interest. One can think of spatial variance as how much the virus diffuses in a geographic dimension, so by summing up the variance in each dimension, we can understand the total spread of a virus across space in a given moment. We want to accurately infer the trace of  $\Sigma_v$  with our phylogenetic sBMDS model trained on a latent airspace. We implement the adaptive HMC algorithm to recover the viral latent locations along with adaptive MH updates on the BMDS precision parameter,  $1/\sigma_v^2$ , and Gibbs updates on  $\Sigma_v^{-1}$ . We let the latent dimension be six as Holbrook et al. (2021) recommended from 5-fold cross-validation. We find 20 leapfrog steps to be adequate as we vary the number of bands/landmarks to 50, 100 and 200.

### 3.2.1 Accuracy

For each subtype and model, we run 120,000 iterations, burning the first 20,000 and saving every 100th iteration. Figure 9 plots the posterior distributions of the strain-

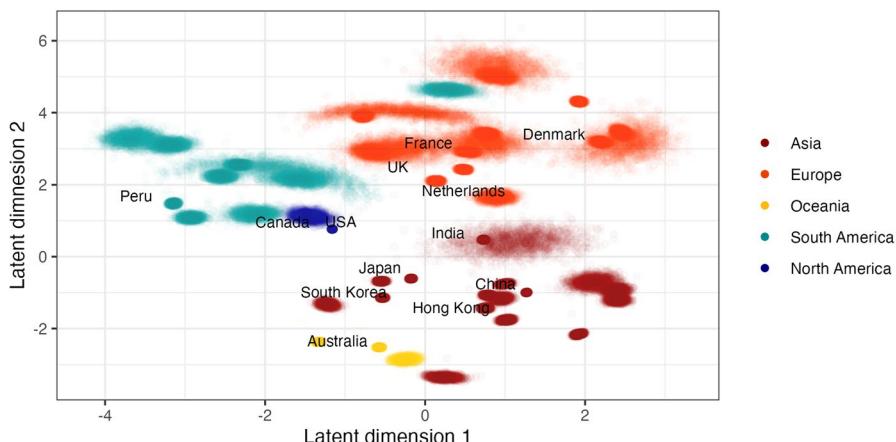


**Fig. 9** Posterior distribution of strain-specific diffusion rates inferred from 6-dimensional Bayesian phylogenetic multidimensional scaling with effective world-wide air traffic space distances for data. Full/HMC refers to the use of the full likelihood and gradient whereas B-sBMDS/HMC uses 50 bands to compute the sparse banded likelihood and gradient for inference within the Hamiltonian Monte Carlo algorithm

specific diffusion rates inferred from the full (left) and banded sparse (right) model. We successfully capture the relative distributions for the B-sBMDS using 50 bands, but note that the posterior modes are slightly off. When we increase the number of bands to 200 (Fig. 14), the distributions appear identical. For Fig. 10, we apply sBMDS with 50 bands to the H1N1 air traffic data and use Procrustes alignment to account for rotation and reflection invariance across MCMC iterations. Since our data has multiple taxon IDs per country, we calculate the mean of the Procrustes-aligned latent locations for each country, yielding a country-level centroid per iteration. We then plot the first two latent dimensions. Figure 10 demonstrates that we obtain a reasonable map; countries in the same continent group together, and within continents, countries with more air traffic are more centrally located. Moreover, the posterior spread of the centroids across iterations provides a measure of uncertainty, revealing both the strength of clustering and the variability in each country's inferred position. Using the `textmineR` package (Jones 2021) in R, we compute the Hellinger distance between the strain-specific posterior distributions of the squared effective distance per year from the full and sparse methods (Table 3). As expected, the Hellinger distance decreases with more bands.

### 3.2.2 Computational performance

We measure efficiency speedups across the four influenza subtypes as the ratio of ESS per hour between the full and sparse versions. From Table 3, we generally observe that B-sBMDS is more efficient than L-sBMDS, which matches our previous findings (Fig. 8). The efficiency speedup decreases with more bands, but is still three times faster for a more than sufficient band count of 200.



**Fig. 10** The first two dimensions of the inferred latent locations for each country using a 6-dimensional sparse Bayesian phylogenetic multidimensional scaling with 50 bands on the H1N1 air traffic data. The plotted latent locations are the country-level Procrustes-aligned posterior means across taxon from 2,000 MCMC samples obtained after thinning

**Table 3** We compare the strain-specific posterior distributions of the inferred diffusion rates from the full and sparse BMDS methods

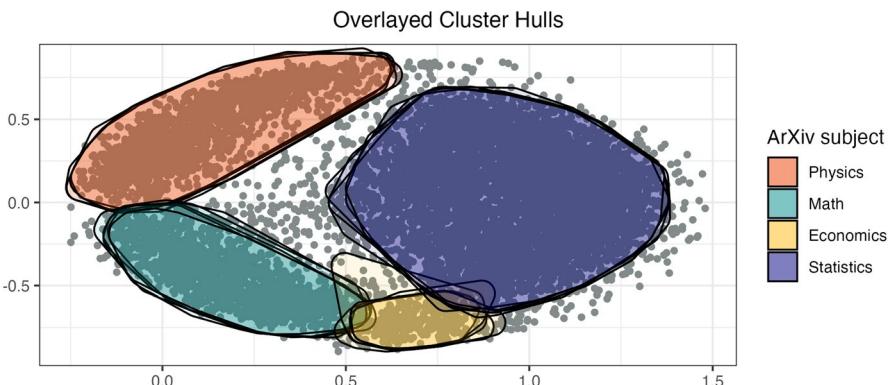
B-sBMDS		
B	Hellinger distance	Average efficiency speedup (min, max)
50	0.024	5.99 (5.58, 6.52)
100	0.021	4.06 (3.99, 4.14)
200	0.019	2.81 (2.76, 2.86)
L-sBMDS		
L	Hellinger distance	Average efficiency speedup (min, max)
50	0.024	5.22 (4.35, 5.63)
100	0.023	3.83 (3.69, 3.90)
200	0.022	2.97 (2.55, 3.52)

We calculate Hellinger distance between the posterior densities obtained using sparse Bayesian multidimensional scaling (sBMDS) and BMDS. Efficiency speedup is the ratio of effective sample size per hour between the full and sparse BMDS versions. We take the average efficiency speedup across the four influenza subtypes

### 3.3 Cluster analysis of ArXiv articles

We explore the utility of sBMDS for approximating posterior uncertainty in downstream tasks by applying cluster analysis to a collection of ArXiv paper abstracts. Using the `arxivscraper` package (Sadjadi 2017) in Python, we scrape articles posted on ArXiv from December 2017 to March 2024 across four subject areas: mathematical logic (math.LO), applied physics (physics.app-ph), statistics/machine learning (stat.ML) and economics (q-fin.ec). Our final dataset includes 9,308 articles with 1,411 related to math, 1,838 to physics, 5,361 to statistics and 698 to economics. We extract each paper’s abstract and embed it into a 768-dimensional numerical vector using `SentenceTransformers` (Reimers and Gurevych 2019) under the `all-mpnet-base-v2` model (Song et al. 2020) in Python. This large language model produces sentence-level embeddings that capture semantic similarity; therefore, abstracts with similar content yield similar embeddings. To form an observed dissimilarity matrix, we compute the pairwise cosine dissimilarities between embeddings, e.g,  $1 - \cos(\mathbf{x}_n, \mathbf{x}_{n'}) = 1 - \frac{\mathbf{x}_n \cdot \mathbf{x}_{n'}}{\|\mathbf{x}_n\| \|\mathbf{x}_{n'}\|}$  for  $\mathbf{x}_n, \mathbf{x}_{n'} \in \mathbb{R}^D$ . Cosine dissimilarity is appropriate for textual data, as it emphasizes directional similarity over magnitude.

We apply B-sBMDS to this 9,308 by 9,308 observed distance matrix using 50, 100, 500, 1,000 and 2,000 bands and obtain the posterior samples of the latent locations in a 2-dimensional space. We use an adaptive HMC algorithm with 20 leapfrog steps to recover the latent locations. We keep every 100th iteration and run enough iterations such that the ESS approximates the number of thinned samples, indicating “near-independence” among samples. To incorporate the uncertainty encoded in the sBMDS posterior samples, we implement a “bagged estimator”-style algorithm. We randomly draw  $S$  iterations from the joint posterior of the latent locations and apply hierarchical density-based spatial clustering of applications with noise (HDBSCAN) to each iteration using the `dbSCAN` package (Hahsler et al. 2019) in R. DBSCAN (Ester et al. 1996) is a non-parametric clustering algorithm that groups points into dense regions based on a user-defined radius parameter  $\epsilon$  and a minimum number of



**Fig. 11** Convex cluster hulls for five posterior samples of the inferred latent locations using hierarchical density-based spatial clustering (HDBSCAN). The inferred locations are estimated from a 2-dimensional B-sBMDS framework using 500 bands and an adaptive Hamiltonian Monte Carlo (HMC) algorithm. The convex hulls are plotted on the posterior mean of the inferred locations across 1,000 iterations, and points outside a hull are considered noise. The ArXiv subject of a cluster is assumed by the most frequent ArXiv category among its members

**Table 4** Accuracy as the number of bands increases using hierarchical density-based spatial clustering (HDBSCAN) on 200 randomly sampled inferred latent locations. The inferred locations are estimated from a 2-dimensional B-sBMDS framework and an adaptive Hamiltonian Monte Carlo (HMC) algorithm. Note that the number of clusters should be 4. HDBSCAN is a non-parametric clustering algorithm and therefore the number of clusters is predicted. Classification error is the proportion of articles whose ArXiv subject label does not match the majority subject label of their assigned cluster, and CI is the credible interval

Band	Mode number of clusters	Mean number of noise points	Mean classification error (95% CI)
50	2	3975	0.226 (0.111, 0.315)
100	3	2954	0.094 (0.038, 0.171)
500	4	1380	0.051 (0.041, 0.061)
1000	4	1248	0.038 (0.031, 0.045)
2000	4	1009	0.034 (0.031, 0.037)

neighbors. Points within a dense region are assigned to the same cluster, while points in low-density regions are labeled as noise. HDBSCAN (Campello et al. 2015) performs DBSCAN for various  $\epsilon$  values and integrates the results to give the most stable output. By applying HDBSCAN, a bounded deterministic function, to the posterior draws of the latent locations from sBMDS, we induce a posterior distribution over the cluster assignments (Christensen et al. 2011).

Figure 11 visualizes the results of HDBSCAN applied to five posterior samples of the latent space plotted on the posterior mean of the latent locations. We set the minimum number of neighbors to 40. Table 4 reports clustering accuracy across band settings. Each cluster is assigned a subject label based on the most frequent ArXiv category among its members. Classification error is the proportion of articles whose ArXiv subject label does not match the subject label of their assigned cluster. We calculate the mean classification error across clusters and note that 500 bands provide a good balance between accuracy and computational cost.

We obtain clustering labels for each article as indirectly induced by uncertainty in the sBMDS embeddings. Across the  $S$  iterations, we compute a co-clustering matrix that estimates the posterior pairwise probability that two articles belong in the same cluster. This matrix can be used for subsequent analyses such as consensus clustering. The combination of sBMDS with HDBSCAN provides an example of using a deterministic function on the state space from a sparse Bayesian embedding model to approximate uncertainty in post-hoc inference.

## 4 Discussion

We present two methods for subsetting the observed dissimilarity data: banded sparse BMDS (B-sBMDS) and landmark sparse BMDS (L-sBMDS). We show that both sparse methods obtain accurate results at low band/landmark counts even with noisy data. Moreover, combining HMC with sBMDS proves effective in inferring thousands of latent locations. We successfully integrate the sBMDS variants within a Bayesian hierarchical model and demonstrate how posterior samples from sBMDS can be combined with deterministic algorithms to approximate uncertainty in downstream tasks.

Based on our experiments, we recommend B-sBMDS as the default choice, since it is more robust to model misspecifications. For example, in our influenza application, in which the Euclidean assumption is violated, B-sBMDS yields a slightly smaller Hellinger distance than L-sBMDS (Table 3). However, when minimizing the number of dissimilarities is critical, L-sBMDS can be highly effective in favorable settings such as (i) the latent dimension is well specified, (ii) the noise distribution is close to Gaussian, and (iii) the dissimilarities can be well approximated by a Euclidean embedding. In such cases, L-sBMDS achieves higher accuracy with fewer landmarks (Figs. 2, 3), leading to greater speedups compared to B-sBMDS. Thus, the computational efficiency of L-sBMDS comes at the price of reduced robustness, making B-sBMDS the safer default across a wide range of applications.

Our simulations show that sBMDS performs well under misspecified noise, e.g., log-normal noise, which is a heavily skewed and heavy-tailed distribution (Fig. 5). Possible extensions to our work include the use of different noise distributions on the observed dissimilarities (Fig. 17). For example, Bakker and Poole (2013) employ Bayesian metric MDS, assuming the observed dissimilarities come from log-normal distributions. While beyond the scope of this paper, sBMDS could also be applied within a generalized BMDS framework (Zhang et al. 2025), which accommodates non-Gaussian errors and diverse dissimilarity metrics. As these distributions still have  $\mathcal{O}(N^2)$  time complexity, sBMDS could improve the computational performance across a wider range of dissimilarity data. Another promising direction is the application of our sparse methods within non-Euclidean latent spaces, as explored in Liu et al. (2024). For example, negatively curved spaces such as the Poincaré disk are effective for embedding sparse or scale-free networks, since their exponential geometry aligns with power-law degree distributions. In such setting, L-sBMDS can improve efficiency by selecting highly connected nodes as landmarks.

Additionally, many potential theoretical developments remain. For example, it appears that one needs approximately  $D\sqrt{N}$  bands where  $N$  is the number of objects and  $D$  is the embedding space. However, we have no formal proof, only experimental results (Sects. 3.2 and 3.3, Fig. 16). Determining the number of desired bands/landmarks is difficult due to its data-dependence. We explain in Sect. A.4 how one could extend Theorem 1's proof of posterior consistency to higher dimensions. The biggest limitations are extending Lemma 3 and obtaining estimates with good dependence on dimension  $D$ . One could also explore treating the coupling matrix  $J_{n,N}$  as a random variable that depends on the observed data (and perhaps changes over the run-time of an algorithm). An appealing feature of Raftery et al. (2012) is that they claim reasonable uncertainty quantification along a truly linear run-time. It seems difficult to formalize such a result with posterior consistency for our sBMDS models as the number of bands (landmarks) grows with the number of objects. We are left with many tantalizing questions: “by including a data-informed approach to model sparsity, can we achieve a linear run-time and still demonstrate posterior consistency?”, “how should we be measuring consistency?”, and “do the datasets Raftery et al. (2012) study have any special features that change the rate of convergence for a sBMDS-like model?”

Lastly, we are interested in further extensions within phylogeography. Holbrook et al. (2021) and Li et al. (2023) select the dimension of the latent diffusion process using cross-validation, which is computationally demanding. Therefore, we want to incorporate a shrinkage prior within the Bayesian phylogenetic MDS framework that penalizes the eigenvalues of the diffusion rate matrix. As long as implementing such a prior does not slow down mixing, this approach may help one learn the latent locations in a faster, more unified manner.

## Proof of Theorem 1

Throughout this section, we fix notation as in the statement of Theorem 1.

### Consistent Estimates of Absolute Values

We note that  $|x_n|$  (but not  $x_n$  itself) is effectively identifiable given the data  $\{\delta_{ny}\}_{y \in G_n^{(k(n))}}$ , and we have the posterior concentration bound:

**Lemma 1** *Fix some  $0 < \alpha < 0.1$  and a sequence  $\epsilon_N = \ell_N^{-0.5+\alpha}$ . Then there exist constants  $c_1, c_2, c_3 > 0$  so that for all  $N$  sufficiently large and all  $n \in [N]$ ,  $k \in [K]$ , we have:*

$$\mathbb{P}[p(\{u : \min(|u - x_n|, |u + x_n|) \leq c_1 \epsilon_N\} | \{\delta_{ny}\}_{y \in G_n^{(k(n))}}) \geq 1 - e^{-c_2 \ell_N \epsilon_N^2}] \geq 1 - e^{-c_3 \ell_N \epsilon_N^2}. \quad (19)$$

**Proof** Given  $x_n$ , the data  $\{\delta_{ny}\}_{y \in G_n^{(k(n))}}$  are i.i.d. with distributions being a finite mixture of truncated Gaussians. Denote the density of this distribution by  $q_{x_n}$ , and let  $\mathcal{F} = \{q_u\}_{u \in I}$  be the associated family of possible distributions.

With  $\epsilon_N$  as above and this choice of  $\mathcal{F}$ , for any fixed  $0 < c < c_{\text{crit}}$  small enough and all  $N > N_0$  large enough, the sequence  $\{\epsilon_N\}$  satisfies Inequality (3.1) of Wong and Shen (1995) for the collection of likelihoods  $\mathcal{F}$ . Applying Theorem 1 of Wong and Shen (1995) (together with the well-known formula for Hellinger distances between Gaussians), there exist constants  $c_1, c_2, c_3$  so that for all  $N$  sufficiently large,

$$\mathbb{P}\left[\sup_{u: \min(|u-x_n|, |u+x_n|) > c_1 \epsilon_N} \prod_{y \in G_{n,N}^{(k(n))}} \frac{q_u(\delta_{n,y}^*)}{q_{x_n}(\delta_{n,y}^*)} \geq e^{-c_2 \ell_N \epsilon_N^2}\right] \leq 4e^{-c_3 \ell_N \epsilon_N^2}, \quad (20)$$

where the outer probability is taken with respect to the distribution of the data  $\{\delta_{ny}\}_{y \in G_{n,N}^{(k(n))}}$  given  $x_n$ . On the other hand, for all  $u$  satisfying  $|u - x_n| < \frac{1}{\ell_N^3}$  and all  $N$  sufficiently large, we have

$$\prod_{y \in G_{n,N}^{(k(n))}} \frac{q_{x_n}(\delta_{n,y}^*)}{q_u(\delta_{n,y}^*)} \leq 2. \quad (21)$$

Combining Inequalities (20) and (21) completes the proof (with possibly different values of  $c_1, c_2, c_3$ ).  $\square$

### Consistent estimates of signs

Fix  $n, n' \in [N]$  and associated indices  $k(n), k(n') \in [K]$ . Fix  $J = J_{n,N} \cap J_{n',N} \setminus (G_{n,N}^{(k(n))} \cup G_{n',N}^{(k(n'))})$  satisfying  $|J| \geq \ell_N$ .

Let  $\hat{x}_n$  be the posterior median of the distribution of  $|x_n|$  given  $\{\delta_{ny}\}_{y \in G_{n,N}^{(k(n))}}$ , and similarly for  $\hat{x}_{n'}$ . For  $j \in J$ , define the Bernoulli random variables  $Z_j = 1_{A_j}$ , where  $A_j$  is the event:

$$A_j = \{\max(\delta_{nj}, \delta_{n'j}) > |\hat{x}_n - \hat{x}_{n'}|\}. \quad (22)$$

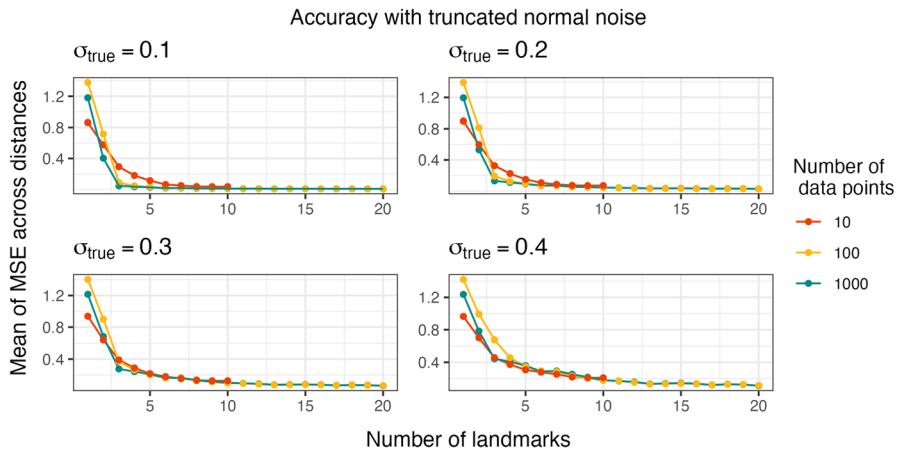
Note that  $\hat{x}_n, \hat{x}_{n'}$  are  $\{\delta_{ny}\}_{y \in G_{n,N}^{(k(n))}} \cup \{\delta_{n'y}\}_{y \in G_{n',N}^{(k(n'))}}$ -measurable, and  $(\delta_{nj}, \delta_{n'j})$  are independent of  $\{\delta_{ny}\}_{y \in G_{n,N}^{(k(n))}} \cup \{\delta_{n'y}\}_{y \in G_{n',N}^{(k(n'))}}$  for each

$j \in J$ , and finally the collection  $\{(\delta_{nj}, \delta_{n'j})\}_{j \in J}$  are independent. Thus, conditional on  $\{\delta_{ny}\}_{y \in G_{n,N}^{(k(n))}} \cup \{\delta_{n'y}\}_{y \in G_{n',N}^{(k(n'))}}$ , the random variables  $\{Z_j\}_{j \in J}$  are i.i.d.

Denote by  $r_{n,n'}$  their common parameter. By the same argument as in Lemma 1, we have the posterior concentration bound:

**Lemma 2** Fix notation  $0 < \alpha < 0.1$ ,  $\epsilon_N = \ell_N^{-0.5+\alpha}$  and notation as above. Then there exist constants  $c_1, c_2, c_3 > 0$  so that, for all  $N$  sufficiently large,

$$\mathbb{P}[p(\{r : |r - r_{n,n'}| \leq c_1 \epsilon_N\} | \{Z_j\}_{j \in J}) \geq 1 - e^{-c_2 \ell_N \epsilon_N^2}] \geq 1 - e^{-c_3 \ell_N \epsilon_N^2}. \quad (23)$$



**Fig. 12** The mean of the mean squared error (MSE) across all distances using 1 to 10 landmarks for 10 data points and 1 to 20 landmarks for 100 and 1000 data points. We estimate Euclidean distances from the inferred locations obtained using an adaptive Hamiltonian Monte Carlo algorithm under landmark sparse Bayesian multidimensional scaling (L-sBMDS).  $\sigma_{true}^2$  is the variance component of the truncated normal noise centered at 0 added to the “true” distance matrix such that  $\sigma_{true}$  corresponds to the BMDS error standard deviation  $\sigma$

We observe that this will allow us to learn whether  $\hat{x}_n, \hat{x}_{n'}$  have the same signs (as long as both are far from 0). More precisely, for  $j \in J$ , define  $Y_j = 1_{B_j}$ , where

$$B_j = \{\max(\delta_{nj}, \delta_{n'j}) > |\hat{x}_n + \hat{x}_{n'}|\}. \quad (24)$$

By the same argument as the one immediately following Equation (22), the  $Y_j$  are i.i.d. Bernoulli. Denote by  $q_{n,n'}$  their common parameter. The following is a direct calculation with Gaussians<sup>1</sup>:

**Lemma 3** *There exists  $C, D > 0$  depending on  $\sigma$  so that, for all  $N$  sufficiently large, the following implication holds:*

$$\{\min(|\hat{x}_n|, |\hat{x}_{n'}|) > C\epsilon_N\} \Rightarrow \{|r_{n,n'} - q_{n,n'}| > D\epsilon_N\}. \quad (25)$$

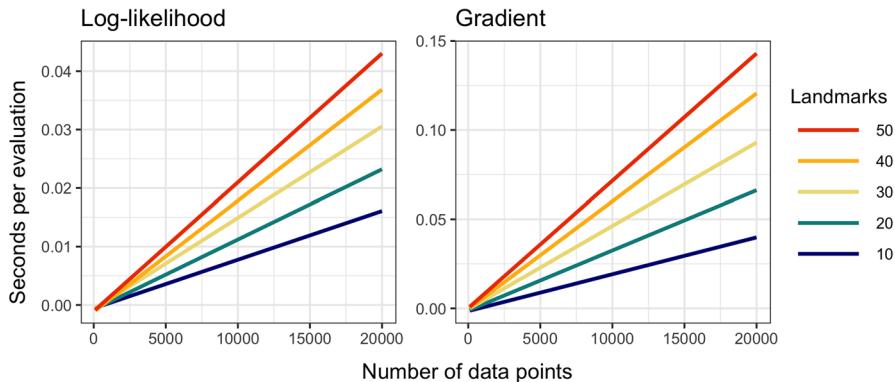
## Completing the Proof

We complete the proof of Theorem 1.

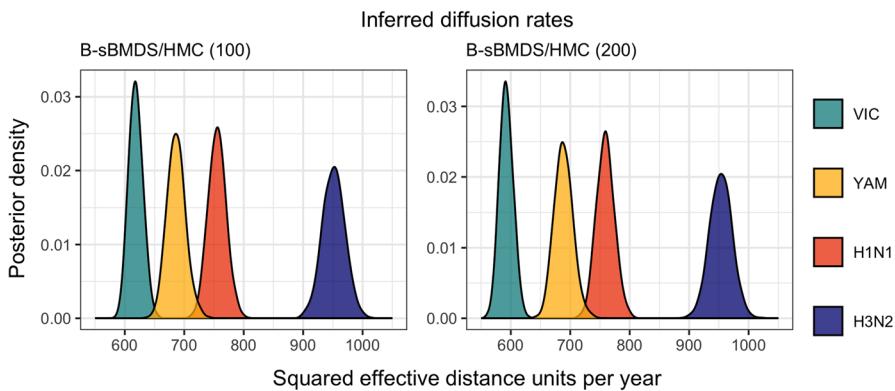
**Proof** For constants  $c_1, c_2, c_3$  to be determined later, we define events

$$\mathcal{A}_N = \left\{ \forall n \in [N], k \in [K], p(\{u : \min(|u - x_n|, |u + x_n|) \leq c_1\epsilon_N\} \mid \{\delta_{ny}\}_{y \in G_{n,N}^{(k(n))}}) \geq 1 - e^{-c_2\ell_N\epsilon_N^2} \right\} \quad (26)$$

<sup>1</sup>If  $\sigma = 0$ , we'd just look at the probability that the latent position is in the interval  $(-\min(|\hat{x}_n|, |\hat{x}_{n'}|), \min(|\hat{x}_n|, |\hat{x}_{n'}|))$ , for which this is obvious. Since  $\sigma > 0$ , a complete calculation needs to add in a few additional cases. These doesn't substantially change the results from the trivial case.



**Fig. 13** Time elapsed to calculate the landmark sparse Bayesian multidimensional scaling (L-sBMDS) likelihood and gradient using  $L$  landmarks as a function of the number of data points



**Fig. 14** Posterior distribution of strain-specific diffusion rates inferred from 6-dimensional Bayesian phylogenetic multidimensional scaling with effective world-wide air traffic space distances for data. B-sBMDS/HMC uses 100 (left) and 200 (right) bands to compute the sparse banded likelihood and gradient for inference within the Hamiltonian Monte Carlo algorithm

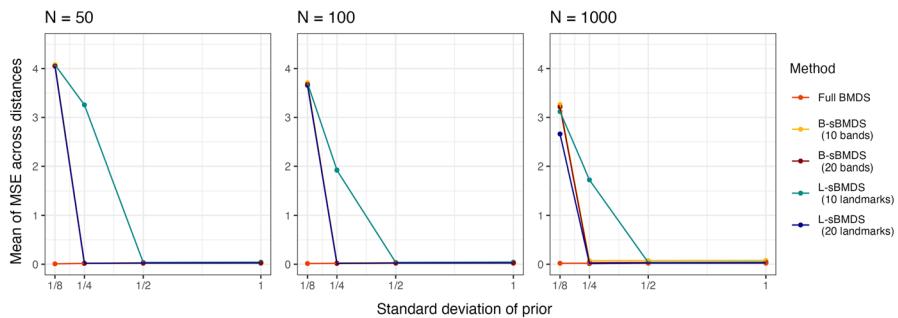
and

$$\mathcal{B}_N = \left\{ \forall n, n' \in [N], p(\{r : |r - r_{n,n'}| \leq c_1 \epsilon_N\} | \{Z_y\}_{y \in J}) \geq 1 - e^{-c_2 \ell_N \epsilon_N^2} \right\}. \quad (27)$$

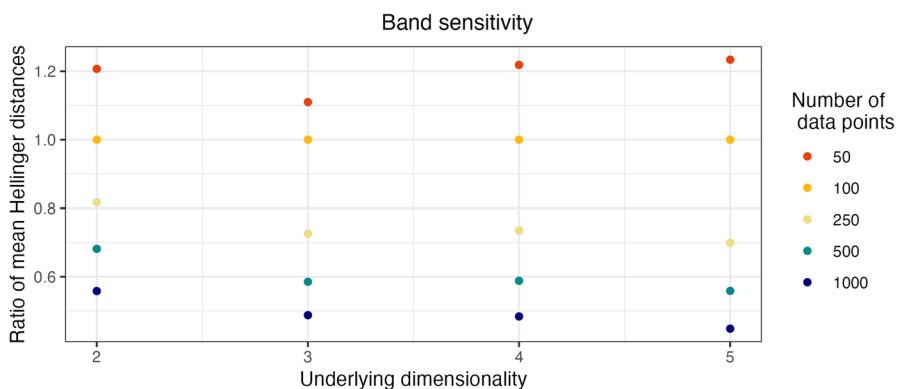
Since we have chosen  $\epsilon_N = \ell_N^{-0.5+\alpha}$  for some  $0 < \alpha < 0.1$ , we have that  $\ell_N \epsilon_N^2 \geq \frac{1}{2} N^\alpha$  for all  $N$  sufficiently large. Thus, by Lemmas 1 and 2, we know that  $\mathcal{A}_N$  and  $\mathcal{B}_N$  occur asymptotically almost surely.

On the event  $\mathcal{A}_N$ , we correctly recover  $|x_n^{(N)}|$  up to additive error  $O(\epsilon_N)$ . We now fix a large constant  $C$  and consider two cases:

1. When  $|x_n^{(N)}| \leq C \epsilon_N$ , recovering  $|x_n^{(N)}|$  up to additive error  $O(\epsilon_N)$  also means recovering  $x_n^{(N)}$  up to additive error  $O(\epsilon_N)$ .

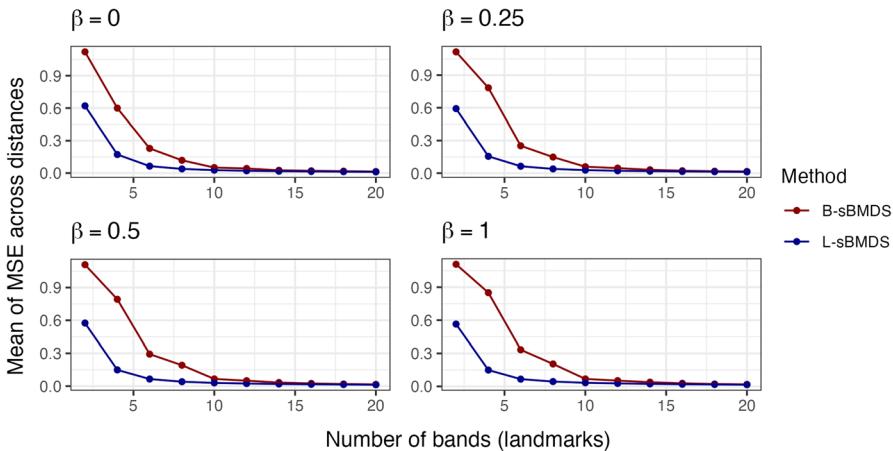


**Fig. 15** The mean of the mean squared error (MSE) across all distances using 10, 20 and all bands/landmarks for 50, 100 and 1,000 data points. We estimate Euclidean distances from the inferred locations obtained using an adaptive Hamiltonian Monte Carlo algorithm under the B-sBMDS, L-sBMDS and full BMDS frameworks. The prior of the latent locations is a multivariate normal distribution centered around 0 with standard deviation equal to  $\frac{1}{8}, \frac{1}{4}, \frac{1}{2}$  or 1. The “true” latent locations are generated from a standard normal multivariate distribution. We sometimes perform worse than simply sampling from the prior because the model “overcompensates” by inflating  $\sigma$ . Since the prior contribution grows with  $N$ , there is a threshold in the ratio between likelihood and prior contributions that must be exceeded for good recovery. With sufficiently large  $N$ , the likelihood eventually dominates. In practice, we recommend that users 1) ensure the BMDS noise scale is small relative to the prior scale, which is scientifically reasonable since observed dissimilarities should be more informative than the latent spread, or 2) learn the prior scale/covariance structure as a hyperparameter in the model as we do in Sect. 3.2 or as in Oh and Raftery (2001)



**Fig. 16** Ratio of the mean Hellinger distance for various dimensionalities ( $D$ ) and number of data points ( $N$ ). Hellinger distances are computed from the posterior distributions of the estimated Euclidean distances between full BMDS and banded sBMDS with  $\lceil D\sqrt{N} \rceil$  bands. Data is generated under Gaussian assumptions as described in Sect. 3. For fixed  $D$ , we calculate the ratio as the mean Hellinger distance with  $N$  data points and with  $N = 100$ . The near horizontal lines across the underlying dimensionality indicate that the selected number of bands obtains results that are very similar to the full model despite the increase in dimensions. The ratio of the Hellinger distance is smaller for larger  $N$  as error goes down at some statistical rate, e.g., of  $\frac{1}{\sqrt{N}}$

- When  $|x_n^{(N)}| \geq C\epsilon_N$  for fixed  $C$  sufficiently large, Lemma 3 implies that on  $\mathcal{B}_N$  we also recover the sign of  $x_n^{(N)}$ . Thus, in either case, we recover  $x_n^{(N)}$  up to additive error  $O(\epsilon_N)$ .  $\square$



**Fig. 17** The mean of the mean squared error (MSE) across all distances from 1 to 20 bands/landmarks for 500 data points. We estimate Euclidean distances from the inferred locations obtained using an adaptive Hamiltonian Monte Carlo algorithm under a one-parameter family of likelihoods interpolating between a Gaussian and Laplace. When  $\beta = 1$ , the likelihood reduces to a Gaussian and when  $\beta = 0$ , it reduces to a Laplace distribution. We vary  $\beta \in [0, 1]$  to explore sensitivity to heavier tails. The “true” latent dissimilarities are Euclidean distances between standard multivariate normal locations, while the observed dissimilarities are perturbed by Laplace noise calibrated to have a standard deviation of 0.2

### Extending Theorem 1 to higher dimensions

It is natural to ask if Theorem 1 holds in higher dimensions. The answer appears to be “yes,” but the only proofs that we are aware of have at least one of the following two substantial flaws: they are noticeably longer or give constants  $C$  that scale very poorly with dimension. We give here a quick sketch of a proof that closely mimics our one-dimensional argument. It requires no new ideas, but gives bounds that scale very poorly with respect to dimension.

In our proof of Theorem 1, we invoke Theorem 1 of Wong and Shen (1995) twice: once in Lemma 1 on the “single row”  $G_{n,N}^{(k(n))}$  to show that we have learned  $|x_n|$  with high accuracy, and again in Lemma 2 on the “pair of rows with large intersection”  $J$  to show that we have learned the sign of  $x_n$  (as long as  $|x_n|$  is sufficiently large). To extend this to a higher dimension  $D$ , we would invoke Theorem 1 of Wong and Shen (1995)  $(D + 1)$  times. On the first invocation, we would show that the posterior distribution of  $x_n$  concentrates near a  $(D - 1)$ -dimensional set that contains the true point. For  $1 \leq d \leq D$ , in the  $d$ 'th invocation, we would show that we have learned that  $x_n$  is on a certain subset of dimension  $(D - d)$  with high accuracy by looking at  $d$  rows of the matrix. Thus, after  $D$  invocations, we would have shown that  $x_n$  is recoverable up to a set of dimension 0. These calculations are nearly identical to the calculations in the current proof.

The last invocation would be used to deal with ambiguity on a finite set, as in the one-dimensional case. Most of the required changes would be routine (e.g., in part 2 of Assumption 1, we would need intersections of  $(D+1)$  parts of the partition to support our  $(D+1)$  invocations of Theorem 1 of Wong and Shen (1995)). The biggest change comes in proving the natural analogue to Lemma 3. This calculation is what describes a quantitative sort of identifiability for the model. To extend our arguments to higher dimensions, we need a result along the lines of: “the set of latent points that (i) lie in a set of dimension  $(D - d + 1)$  and (ii) have a given expected distance  $r$  will lie in a reasonably nice set of dimension  $(D - d)$ .” Ignoring truncations, one can easily check see that the following is true in dimension  $d = 1$ : for  $r > 0$ , the set of points  $x$  such that  $E[\delta_{nn'}|x_n = x] = r$  is a sphere. When we allow for truncations in dimension  $d = 1$ , we merely need to slice off part of the sphere, and so Lemma 3 is straightforward. For fixed  $d > 1$ , the various truncations and conditionings involved in repeatedly using this calculation will result in repeated application of unions, intersections and truncation operations to these spheres. In dimension  $d = 2$ , proving the resulting analogue of Lemma 3 in this way is a straightforward but very messy calculus exercise. Unfortunately, we see no easy way to do quick calculations on the resulting set in arbitrary dimension, and no way at all to obtain estimates with reasonable dependence on  $D$ .

## Additional plots

Figures 12 and 13 are analogous to Figs. 1 and 7 from Sect. 3, but under the sparse model using landmarks (L-sBMDS). Figure 12 demonstrates that very few landmarks are necessary to achieve high accuracy relative to the number of data points. Figure 13 plots the raw speed-ups, varying the number of landmarks as the number of data points increases. Figure 14 illustrates the posterior distribution of the strain-specific diffusion rates under the B-sBMDS/HMC model using 100 and 200 bands. When the number of bands is 200, we see no apparent difference from the full BMDS plot in Fig. 9. In Figs. 15, 16 and 17, we explore the sensitivity of our sparse methods to modeling assumptions such as the choice of prior, number of bands relative to dimensionality and likelihood.

**Funding** This work was supported by the NIH (K25 AI153816) and the NSF (DMS 2152774 and DMS 2236854).

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

Bakker R, Poole KT (2013) Bayesian metric multidimensional scaling. *Polit Anal* 21(1):125–140

Bedford T, Suchard MA, Lemey P, Dudas G, Gregory V, Hay AJ, McCauley JW, Russel CA, Smith DJ, Rambaut A (2014) Integrating influenza antigenic dynamics with molecular evolution. *Elife* 3:e01914

Beskos A, Pillai NS, Roberts GO, Sanz-Serna JM, Stuart AM (2013) Optimal tuning of the hybrid monte-carlo algorithm. *Bernoulli* 19(5A):1501–1534

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

Campello RJ, Moulavi D, Zimek A, Sander J (2015) Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Trans Knowl Discov Data (TKDD)* 10(1):5

Christensen R, Johnson W, Branscum A, Hanson TE (2011) Bayesian ideas and data analysis: an introduction for scientists and statisticians. CRC Press, Boca Raton

Ester M, Kriegel H-P, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD), pages 226–231. AAAI Press

Geyer CJ (2011) Introduction to MCMC, chapter 1. Campman & Hall, New York

Green PJ (1995) Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika* 82(4):711–732

Hahsler M, Piekenbrock M, Doran D (2019) dbscan: Fast density-based clustering with R. *J Stat Softw* 91(1):1–30

Hastings W (1970) Monte carlo sampling methods using markov chains and their applications. *Biometrika* 57(1):97–109

Holbrook AJ, Lemey P, Baele G, Dellicour S, Brockmann D, Rambaut A, Suchard MA (2021) Massive parallelization boosts big bayesian multidimensional scaling. *J Comput Graph Stat* 30(1):11–24

Jones T (2021) textmineR: Functions for Text Mining and Topic Modeling. R Package Version 3:5

Lemey P, Rambaut A, Drummond AJ, Suchard M (2009) Bayesian phylogeography finds its roots. *PLoS Comput Biol* 5(9):e1000520

Li Y, Ghafari M, Holbrook A, Boonen I, Amor N, Catalano S, Webster J, Li Y, Li H, Vergote V, Maes P, Chong Y, Laudisoit A, Baele P, Ngoy S, Mbalitini S, Gembu G, Musaba A, Goüy de Bellocq J, Leirs H, Verheyen E, Pybus O, Katzourakis A, Alagaili A, Gryseels S, Li Y, Suchard M, Bletsas M, Lemey P (2023) The evolutionary history of hepaciviruses. *BioRxiv: the Preprint Server for Biology*

Lin L, Fong DK (2019) Bayesian multidimensional scaling procedure with variable selection. *Comput Stat Data Anal* 129:1–13

Liu B, Lubold S, Raftery AE, McCormick TH (2024) Bayesian hyperbolic multidimensional scaling. *J Comput Graph Stat* 33(3):869–882

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. *J Chem Phys* 21(6):1087–1092

Neal RM (2012) MCMC using Hamiltonian dynamics, chapter 5. Campman & Hall, New York

Oh M-S, Raftery AE (2001) Bayesian multidimensional scaling and choice of dimension. *J Am Stat Assoc* 96(455):1031–1044

Oh M-S, Raftery AE (2007) Model-based clustering with dissimilarities: A bayesian approach. *J Comput Graph Stat* 16(3):559–585

Plummer M, Best N, Cowles K, Vines K (2006) Coda: Convergence diagnosis and output analysis for mcmc. *R News* 6(1):7–11

R Core Team (2023) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria

Raftery AE, Niu X, Hoff PD, Yeung KY (2012) Fast inference for the latent sapce network model using a case-control approximate likelihood. *J Comput Graph Stat* 21(4):909–919

Reimers N, Gurevych I (2019) 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

Roberts GO, Rosenthal JS (2001) Optimal scaling for various metropolis-hastings algorithms. *Stat Sci* 16(4):351–367

Rolfes MA, Foppa IM, Garg S, Flannery B, Brammer L, Singleton JA, Burns E, Jernigan D, Olsen SJ, Bressee J, Reed C (2018) Annual estimates of the burden of seasonal influenza in the united states: A tool for strengthening influenza surveillance and preparedness. *Influenza Other Respir Viruses* 12(1):132–137

Sadjadi M (2017) arxivscraper

Silva VD, Tenenbaum JB (2004) Sparse multidimensional scaling using landmark points. Technical Report (Stanford University)

Song K, Tan X, Qin T, Lu J, Liu T-Y (2020) Mpnet: Masked and permuted pre-training for language understanding. arXiv preprint [arXiv:2004.09297](https://arxiv.org/abs/2004.09297)

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

Wickham H (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer, Berlin

Wong WH, Shen X (1995) Probability inequalities for likelihood ratios and convergence rates of sieve MLES. *Ann Stat* 23(2):339–362

Zhang J, Cao J, Wang L (2025) Generalized bayesian multidimensional scaling and model comparison. arXiv preprint [arXiv:2306.15908](https://arxiv.org/abs/2306.15908)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Authors and Affiliations

Ami Sheth<sup>1</sup>  · Aaron Smith<sup>2</sup> · Andrew J. Holbrook<sup>1</sup>

 Ami Sheth  
amisheth26@g.ucla.edu

<sup>1</sup> Department of Biostatistics, University of California, Los Angeles, CA, USA

<sup>2</sup> Department of Mathematics and Statistics, University of Ottawa, Ottawa, Ontario, Canada