JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2022, VOL. 31, NO. 2, 337–350
https://doi.org/10.1080/10618600.2021.2000420

Learning Subspaces of Different Dimensions

Brian St. Thomas[a], Kisung You[b], Lizhen Lin[b], Lek-Heng Lim[c], and Sayan Mukherjee[d]

[a]Department of Statistical Science, Duke University, Durham, NC; [b]Department of Applied and Computational Mathematics and Statistics, The University of Notre Dame, Notre Dame, IN; [c]Computational and Applied Mathematics Initiative, Department of Statistics, University of Chicago, Chicago, IL; [d]Departments of Statistical Science, Mathematics, Computer Science, and Biostatistics & Bioinformatics, Duke University, Durham, NC

ABSTRACT

We introduce a Bayesian model for inferring mixtures of subspaces of different dimensions. The model allows flexible and efficient learning of a density supported in an ambient space which in fact can concentrate around some lower-dimensional space. The key challenge in such a mixture model is specification of prior distributions over subspaces of different dimensions. We address this challenge by embedding subspaces or Grassmann manifolds into a sphere of relatively low dimension and specifying priors on the sphere. We provide an efficient sampling algorithm for the posterior distribution of the model parameters. We illustrate that a simple extension of our mixture of subspaces model can be applied to topic modeling. The utility of our approach is demonstrated with applications to real and simulated data.

ARTICLE HISTORY
Received February 2020
Revised July 2021

KEYWORDS
Conway embedding;
Mixtures of subspaces;
Subspaces of different dimensions

1. Introduction

The problem of modeling manifolds has been of great interest in a variety of statistical problems including dimension reduction (Belkin and Niyogi 2003; Donoho and Grimes 2003; Roweis and Saul 2000), characterizing the distributions of statistical models as points on a Riemannian manifold (Rao 1945; Efron 1978; Amari 1982), and the extensive literature in statistics and machine learning on manifold learning (Giné and Koltchinskii 2006; Cook 2007; Mukherjee, Zhou, and Wu 2010). A generalization of the manifold setting is to model unions and intersections of manifolds (of possibly different dimensions), formally called stratified spaces (Goresky and MacPherson 1988; Geiger et al. 2001; Bendich, Mukherjee, and Wang 2012). Stratified spaces arise when data or parameter spaces are characterized by combinations of manifolds such as the case of mixture models. One of the most important special cases arises when the manifolds involved are all affine subspaces or linear subspaces. There is large body of work on learning this important special case of mixture of subspaces, see (Huang, Ma, and Vidal 1999; Zheng 2011; Pimentel-Alarcón et al. 2017; Chen, De, and Vijayaraghavan 2021). Mixtures of linear subspaces have been suggested in applications such as tracking images (Vidal, Ma, and Sastry 2005; Haro, Randall, and Sapiro 2007), representation learning of high-dimensional visual objects, such as faces, handwritten digits, or general targets (Laaksone and Oja 1996; Zheng 2011), quantitative analysis of evolution or artificial selection (Lande 1979; Hansen and Houle 2008), applications in communication and coding theory (Zheng and Tse 2002; Ashikhmin and Calderbank 2003), and is relevant for text modeling (Reisinger et al. 2010; Blei, Ng, and Jordan 2003). In this paper we provide a model for the simplest instance of inferring stratified spaces: estimating mixtures of linear subspaces of different dimensions.

A recent work Chandra, Canale, and Dunson (2021) proposed a latent mixture model for Bayesian clustering of high-dimensional data where a common factor loading matrix (subspace) is assumed across the clusters, thus a fixed subspace is learnt. Recent work on estimating mixtures of subspaces with theoretical guarantees has been limited to *equidimensional* subspaces (Lerman and Zhang 2010; Page, Bhattacharya, and Dunson 2013). A Bayesian procedure for inference of mixtures of subspaces of equal dimensions was developed in Page, Bhattacharya, and Dunson (2013). A penalized loss-based procedure was introduced in Lerman and Zhang (2010) to learn mixtures of subspaces, called K-flats. The conceptual difficulty in extending either approach to subspaces of different dimensions is the loss of Riemannian structure. A natural parameterization of subspaces of fixed dimension is as points on the Grassmann manifold where there is a natural geodesic. The problem with considering subspaces of different dimensions is that there is no clear Riemannian structure and no well-defined geodesic distance, and moving across dimensions introduces singularities that result in a loss of the Riemannian structure. The complication due to the lack of Riemannian structure manifests in the Bayesian approach by requiring the posterior samples drawn from subspaces of different dimensions which is a difficult sampling problem via reversible-jump MCMC (Green 1995). In the penalized loss model, the problem is immediate as the loss function requires the computation of distances between subspaces which is problematic for subspaces of different dimensions.

There are many applications where variable dimensions of subspaces would be of great utility. In the quantitative genetics

CONTACT Lizhen Lin *lizhen.lin@nd.edu* University of Notre Dame, Applied and Computational Mathematics and Statistics, Notre Dame, IN 46556.

setting (Lande 1979; Hansen and Houle 2008), subspaces spanned by the additive genetic variance-covariance matrix may be of different intrinsic dimensions, which necessitates to compare subspaces of varying dimensions. In the setting of shape statistics and Procrustes analysis (Kendall 1984), the number of dimension is related to the number of landmarks used to model a surface. Shape statistics with a variable number of dimensions or landmarks would have practical implications.

The key idea we develop in this article is that subspaces of different dimensions $1, 2, \ldots, m$ can be embedded into a sphere of relatively low dimension $\mathbb{S}^{(m-1)(m+2)/2}$ where chordal distances on the sphere can be used to compute distances between subspaces of differing dimensions (Conway, Hardin, and Sloane 1996). This embedding removes the discontinuity that occurs in moving between subspaces of different dimensions when one uses the natural metric for a Grassmann manifold. It allows us to embed subspaces of different dimensions into a sphere where we can sample efficiently. The idea of sampling subspaces of different dimensions extends the work of Hoff (2009) where efficient methodology was developed to sample from a space of orthonormal matrices with fixed intrinsic dimension using the matrix Bingham–von Mises–Fisher distribution. From a geometric perspective, the method developed in Hoff (2009) simulated from a Stiefel manifold with fixed intrinsic dimension. In this article, we provide a methodology to simulate over Stiefel manifolds of varying intrinsic dimensions.

There exists an extensive frequentist literature on subspace learning such as Sparse Subspace Clustering (Elhamifar and Vidal 2009), Low-rank Subspace Clustering and variations thereof, and Ensemble K-Subspaces (Lipor et al. 2021; Liu, Lin, and Yu 2010; Lu et al. 2012; Soltanolkotabi and Candes 2012; Vidal and Favaro 2014), many of which do not require the subspace dimensions be known or equal. We highlight some distinctive features of our method in comparison with the existing literature: (i) our model is fully Bayesian which allows the learning of all the parameters including the subspace dimension as well as the representation of the subspaces in the same framework while most of the existing algorithms are not able to recover the subspaces or the associated rich geometric information such as subspace dimensions; (ii) our model-based approach makes out of sample prediction possible and easy while most of other algorithms apply the spectral clustering algorithms to an affinity matrix of data, which returns only clustering information of the dataset making prediction of the cluster information of an out of sample difficult; (iii) our model does not require hyperparameter tuning or solving any optimization problem which can in general be computationally expensive and, (iv) our model allows uncertainty qualifications of all the parameters in the models. Numerical comparisons with these methods are carried out in our simulation section. A key contribution of our work is employing the Conway embedding for subspaces to a higher-dimensional sphere which allows us to impose a natural prior on subspaces of different dimensions through a prior on the sphere. The proposed algorithm is implemented in an R package, which is available at *https://kisungyou.com/T4cluster*.

The structure of the paper is as follows. In Section 2, we state a likelihood model for a mixture of k subspaces each of dimension d_k. In Section 2.3, we define the embedding procedure we use to model subspaces of different dimensions and specify the model with respect to the likelihood and prior. In Section 3, we provide an algorithm for sampling from the posterior distribution. For some of the parameters standard methods will not be sufficient for efficient sampling and we use a Gibbs posterior for efficient sampling. In Section 4, we extend the mixture of subspaces model to topic modeling. In Section 5, we use simulated data to provide an empirical analysis of the model and then we use real data to show the utility of the model. We close with a discussion.

2. Model Specification

2.1. Notation

We first specify notation for the geometric objects used throughout this paper. The Grassmann manifold or Grassmannian of a d-dimensional subspace in $\mathbb{R}^m$ will be denoted $\mathrm{Gr}(d, m)$, $d \leq m$. The Stiefel manifold of $m \times d$ matrices with orthonormal columns will be denoted $\mathrm{V}(d, m)$ and when $d = m$ we write $\mathrm{O}(d)$ for the orthogonal group. We use boldfaced uppercase letters, for example, $\mathbf{U}$, to denote subspaces and the corresponding letter in normal typeface, for example, U, to denote the matrix whose columns form an orthonormal basis for the respective subspace. Note that $\mathbf{U} \in \mathrm{Gr}(d, m)$ and $U \in \mathrm{V}(d, m)$. A subspace has infinitely many different orthonormal bases, related to one another by the equivalence relation $U' = UX$ where $X \in \mathrm{O}(d)$. We identify a subspace $\mathbf{U}$ with the equivalence class of all its orthonormal bases $\{UX \in \mathrm{V}(m, d) : X \in \mathrm{O}(d)\}$ thereby allowing the identification $\mathrm{Gr}(d, m) = \mathrm{V}(d, m)/\mathrm{O}(d)$.

In this article, the dimension of the ambient space m will always be fixed but our discussions will often involve multiple copies of Grassmannians $\mathrm{Gr}(d, m)$ with different values of d. We will use the term "Grassmannian of dimension d" when referring to $\mathrm{Gr}(d, m)$ even though as a manifold, $\dim \mathrm{Gr}(d, m) = d(m - d)$.

2.2. Likelihood Specification

We consider data $X = (x_1, \ldots, x_n)$ drawn independently and identically from a mixture of K subspaces where each observation x_i is measured in the ambient space $\mathbb{R}^m$. We assume each point in the population is concentrated near a linear subspace $\mathbf{U}_k$ which we represent with an orthonormal basis U_k, $\mathbf{U}_k = \mathrm{span}(U_k)$, $k = 1, \ldots, K$.

We first state the likelihood of a sample conditional on the mixture component. Each mixture component is modeled using a d_k-dimensional normal distribution to capture the subspace and a $m - d_k$-dimensional normal distribution to model the residual error or null space:

$$U_k^{\mathsf{T}} x \sim \mathcal{N}_{d_k}(\mu_k, \Sigma_k), \quad V_k^{\mathsf{T}} x \sim \mathcal{N}_{m-d_k}(V_k^{\mathsf{T}}\theta_k, \sigma_k^2 I),$$

where U_k is the orthonormal basis for the kth component and $U_k^{\mathsf{T}} x$ is modeled by a multivariate normal with mean μ_k and covariance Σ_k. V_k is the basis for the null space $\ker(U_k)$, which models the residual error as multivariate normal with variance $\sigma_k^2 I$. We are estimating affine subspaces so the parameter θ_k serves as a location parameter for the component. Also note that without loss of generality we can assume that Σ_k is diagonal

since we may diagonalize the covariance matrix $\Sigma_k = Q_k D_k Q_k^\mathsf{T}$ and rotate U_k by Q_k resulting in a parameterization that depends on U_k and a diagonal matrix. The distributions for the null space and the subspace can be combined and specified by either of the following parameterizations:

$$x \sim \begin{cases} \mathcal{N}_m\left(U_k\mu_k + \theta_k, U_k\Sigma_k U_k^\mathsf{T} + \sigma_k^2 V_k V_k^\mathsf{T}\right) \\ \mathcal{N}_m\left(U_k\mu_k + \theta_k, U_k(\Sigma_k - \sigma_k^2 I_{d_k})U_k^\mathsf{T} + \sigma_k^2 I_m\right). \end{cases} \tag{1}$$

It will be convenient to use the second parameterization for our likelihood model.

Given the above likelihood model for a component, we can specify the following mixture model:

$$x \sim \sum_{k=1}^{K} w_k \mathcal{N}_m\left(U_k\mu_k + \theta_k, U_k(\Sigma_k - \sigma_k^2 I_{d_k})U_k^\mathsf{T} + \sigma_k^2 I_m\right), \tag{2}$$

where $w = (w_1, \ldots, w_K)$ is a probability vector and we assume K components. We will use a latent or auxiliary variable approach to sample from the above mixture model and specify a K-dimensional vector δ with a single entry of 1 and all other entries of zero, $\delta \sim \text{Mult}(1, w)$. The conditional probability of x given the latent variable is

$$x|\delta \sim \sum_{k=1}^{K} \delta_k \mathcal{N}_m\left(U_k\mu_k + \theta_k, U_k(\Sigma_k - \sigma_k^2 I_{d_k})U_k^\mathsf{T} + \sigma_k^2 I_m\right).$$

2.3. Prior Specification and the Spherical Embedding

The parameters in the likelihood for each component are $(\theta_k, \Sigma_k, \sigma_k^2, U_k, \mu_k, d_k)$ and the mixture parameters are weights w. Again we fix the number of mixtures at K. Prior specification for some of these parameters are straightforward. For example, the location parameter θ_k is normal, the variance terms Σ_k and σ_k^2 are inverse-gamma, and the mixture weights are Dirichlet. A prior distribution for each triple (U_k, μ_k, d_k) is less obvious.

The inherent difficulty in sampling this triple is that we do not want to fix the dimension of the subspace d_k, but consider d_k as random. We can state the following joint prior on the triple (U_k, μ_k, d_k):

$$\pi(U_k, \mu_k, d_k) = \pi(U_k|d_k)\,\pi(\mu_k|d_k)\,\pi(d_k).$$

Given d_k we can specify $\mu_k|d_k$ as a multivariate normal of dimension d_k. For $\pi(U_k|d_k)$, one possible choice is to specify a conjugate distribution for U_k as the von Mises–Fisher (MF) distribution (Fisher 1953)

$$\text{MF}(U_k|A) \propto \text{etr}(A^\mathsf{T}U_k),$$

where etr is the exponential trace operator. The matrix von Mises–Fisher distribution is a spherical distribution over the set of all $m \times d_k$ matrices, also known as the Stiefel manifold which we denote as $\text{V}(d_k, m)$. In addition to the requirement on fixed dimensionality, this specification faces many other issues. As d_k changes, the dimension of the matrix A also needs to change and one cannot simply add columns of zeros since columns need to be orthonormal. An additional constraint on the prior is that a small change in dimension d_k should only change the prior on U_k slightly. This constraint is to avoid model fitting inconsistencies. This constraint highlights the key difficulty in prior specification over subspaces of different dimensions: how to measure the distance between subspaces of different dimensions.

We will use the geometry of the subspace $\mathbf{U}_k$ to specify an appropriate joint prior on (U_k, d_k). Recall that the set of all d_k-dimensional linear subspaces in $\mathbb{R}^m$ is the Grassmann manifold $\text{Gr}(d_k, m)$ and that we represent a subspace $\mathbf{U}_k \in \text{Gr}(d_k, m)$ with an orthonormal matrix $U_k \in \text{V}(d_k, m)$ from an equivalence class $\{U_k \in \text{V}(d_k, m) : \text{span}(U_k) = \mathbf{U}_k\}$. We need to place priors on Grassmanians of different dimension d_k. The key tool we use to specify such a prior is the embedding of $\text{Gr}(d_k, m)$ into $\mathbb{S}^{(m-1)(m+2)/2}$, an appropriately chosen sphere[1] in $\mathbb{R}^{m(m+1)/2}$ as proposed in Conway, Hardin, and Sloane (1996). This embedding allows us to embed subspaces of different dimensions into the same space and measure distances between the embedded subspaces as a function of only the ambient (embedded) space. We will use this embedding to place a prior on U_k which implicitly specifies a prior on d_k. This embedding will have some very nice properties in terms of prior specification and computation.

The following theorem states that embedding the Grassmanian into a sphere allows us to measure distances between subspaces.

Theorem 1 (Conway–Hardin–Sloane 1996). The representation of a subspace $\mathbf{U} \in \text{Gr}(d, m)$ by its projection matrix $P_\mathbf{U}$ gives an isometric embedding of $\text{Gr}(d, m)$ into a sphere of radius $\sqrt{d(m-d)/m}$ in $\mathbb{R}^{m(m+1)/2}$, with $d_p(\mathbf{U}, \mathbf{V}) = \frac{1}{\sqrt{2}}\|P_\mathbf{U} - P_\mathbf{V}\|_F$, where $P_\mathbf{V}$ is the projection matrix onto $\mathbf{V}$.

The embedding procedure can be described in the following steps: (i) given a basis U_k compute the projection matrix $P_k = U_k U_k^\mathsf{T}$, (ii) take all the entries of P_k in the upper triangle (or lower triangle) as well as all the elements in the diagonal except for one as a vector in $\mathbb{R}^{m(m+1)/2-1}$. The sum of all the entries on the vector will be a constant. This is a result of the orthogonality of U_k, which means that all the subspaces of dimension d_k lie on the same sphere. The key observation by Conway, Hardin, and Sloane (1996) was that if the extra coordinate is included, thus embedding into $\mathbb{R}^{m(m+1)/2}$, the subspaces are still embedded into spheres and each of these spheres are cross-sections of a higher-dimensional sphere which we denote as $\mathbb{S}^{(m-1)(m+2)/2}$. The sphere $\mathbb{S}^{(m-1)(m+2)/2}$ is centered at $\varphi\left(\frac{1}{2}I_m\right) = \text{vech}\left(\frac{1}{2}I_m\right)$ where $\varphi(A)$ denotes the embedding of the projection matrix A and vech is the half-vectorization operation

$$\text{vech}\left(\begin{bmatrix} a & b \\ b & d \end{bmatrix}\right) = \begin{bmatrix} a \\ b \\ d \end{bmatrix},$$

for an example of 2 by 2 matrix. The 0-dimensional subspace is embedded at the origin $\mathbf{0} \in \mathbb{R}^{m(m+1)/2}$. The radius of $\mathbb{S}^{(m-1)(m+2)/2}$ is $\sqrt{m(m+1)/8}$. In summary,

$$\mathbb{S}^{(m-1)(m+2)/2} = \{x \in \mathbb{R}^{m(m+1)/2} : \|x - c\|^2 = m(m+1)/8\},$$
$$\text{where } c = \text{vech}\left(\frac{1}{2}I_m\right).$$

[1] Note that the dimension of a sphere in $\mathbb{R}^d$ is $d-1$ and that $m(m+1)/2 - 1 = (m-1)(m+2)/2$.

Grassmann manifolds are embedded into cross-sections of $\mathbb{S}^{(m-1)(m+2)/2}$ where the projection matrix corresponding to the pre-image has an integer-valued trace. The geodesic distance along the surface of the sphere, $d_{\mathbb{S}^{(m-1)(m+2)/2}}$, corresponds to the projective distance $d_p(\cdot,\cdot)$ between two subspaces $\mathbf{U}_1, \mathbf{U}_2 \in \mathrm{Gr}(d,m)$,

$$d_p(\mathbf{U}_1,\mathbf{U}_2) = \left[\sum\nolimits_{j=1}^{d} \sin^2(\theta_j)\right]^{1/2},$$

where $\theta_1,\ldots,\theta_d$ are the principal angles between the subspaces.

The representation of Grassmannians as points on $\mathbb{S}^{(m-1)(m+2)/2}$ has several useful properties.

Sphere interpretation. The sphere $\mathbb{S}^{(m-1)(m+2)/2}$ provides an intuitive way to sample subspaces of different dimensions by sampling from $\mathbb{S}^{(m-1)(m+2)/2}$. Under the projective distance, the sphere also has an intuitive structure. For example, distances between subspaces of different dimensions can also be computed as the distance between points on the sphere where the points will be located on different cross-sections. Under the projective distance, the orthogonal complement of a subspace $\mathbf{U}$ is the point on $\mathbb{S}^{(m-1)(m+2)/2}$ that maximizes the projective distance. Further, the projection matrix is always invariant to the representation U.

Differentiable. The projective distance, however, is square differentiable everywhere, making it more suitable in general for optimization problems. This is not the case for distances like geodesic distance or the Asimov–Golub–Van Loan distance where maximizing the distance between a set of subspaces will result in distances that lie near non-differentiable points (Conway, Hardin, and Sloane 1996). This numerical instability can lead to sub-optimal solutions.

Ease of computation. The projective distance is easy to compute via principal angles, which are readily computable with singular value decomposition (Golub and Van Loan 2013). Working with the embedding requires only a relatively small number of coordinates—in fact only quadratic in m or $m(m+1)/2$. Furthermore one can exploit many properties of a sphere in Euclidean space in our computations. For example, sampling from a sphere is simple. The number of required coordinates is small compared to alternative embeddings of the Grassmannian, see Hamm and Lee (2005). In contrast, the usual Plücker embedding requires a number of coordinates that is $\binom{m}{d}$, that is, exponential in m. Moreover, the Plücker embedding does not reveal a clear relationship between Grassmannians of different dimensions, as there is using the spherical embedding.

We will place a prior on projection matrices by placing a distribution over the lower half of $\mathbb{S}^{(m-1)(m+2)/2}$, points on $\mathbb{S}^{(m-1)(m+2)/2}$ corresponding to cross-sections where the subspace corresponding to the pre-image has dimension $d < m(m+1)/4$. We only consider the lower half since we assume the model to be low-dimensional. The prior over projection matrices implies a prior over U_k and d_k. A point drawn from $\mathbb{S}^{(m-1)(m+2)/2}$ may not correspond to a subspace since only a point with integer-valued trace qualifies as a pre-image. We address this problem by the following procedure: given a sampled point $q \in \mathbb{S}^{(m-1)(m+2)/2}$ we return the closest point $p \in \mathbb{S}^{(m-1)(m+2)/2}$ that is the pre-image of a subspace. The following theorem states the procedure.

Theorem 2. Given a point $q \in \mathbb{S}^{\ell}$, the point p that minimizes the geodesic distance on $\mathbb{S}^{\ell}$, $d_{\mathbb{S}^{\ell}}(q,p)$, subject to

$$\varphi^{-1}(p) \in \bigcup_{d=0}^{\ell} \mathrm{Gr}(d,\ell)$$

can be found by the following procedure:

1. Compute $Q = \varphi^{-1}(q)$.
2. Set the dimension of p to $d = \mathrm{tr}(Q)$.
3. Compute the eigendecomposition $Q = A\Lambda A^{-1}$.
4. Set B an $\ell \times d$ matrix equal to the columns of A corresponding to the top d eigenvalues.
5. Let $p = \varphi(BB^{\mathsf{T}})$.

Proof. We can consider two cases dichotomously. Suppose the point $q \in \mathbb{S}^{\ell}$ is already on a cross-section of the sphere corresponding to d-dimensional subspaces $\mathrm{Gr}(d,\ell)$. The eigendecomposition of $Q = \varphi^{-1}(q)$ returns exactly d non-zero eigenvalues and the corresponding eigenvectors form a basis for the subspace that is embedded into the point q. On the other hand, suppose the point q lies between cross-sections of the sphere that correspond to Grassmannians. The above algorithm minimizes the Euclidean distance between the point p and q since the procedure to use top d eigenpairs is the best low-rank approximation of any symmetric matrix (Golub and Van Loan 2013) and therefore minimizes the distance on $\mathbb{S}^{\ell}$ induced by the embedding. □

The full model is specified as follows for each x_i, $i = 1,\ldots,n$,

$$\begin{aligned} w &\sim \mathrm{Dir}_K(\alpha),\ \delta_i \sim \mathrm{Mult}(w), \\ P_k &\sim \mathcal{P}(\mathbb{S}^{(m-1)(m+2)/2}), \quad U_k U_k^{\mathsf{T}} = P_k, \\ d_k &= \mathrm{tr}(P_k), \qquad (3) \\ \mu_k | d_k &\sim \mathcal{N}_{d_k}(0, \lambda I), \\ \theta_k | U_k &\sim \mathcal{N}_m(0, \phi I), \quad U_k^{\mathsf{T}}\theta_k = 0, \qquad (4) \\ \sigma_k^{-2} &\sim \mathrm{Ga}(a,b), \\ \Sigma_{k(j)}^{-1} | d_k &\sim \mathrm{Ga}(c,d), \quad j = 1,\ldots,d_k, \\ x_i | \delta_i &\sim \sum\nolimits_{k=1}^{K} \delta_{ik} \mathcal{N}_m\Big(U_k\mu_k + \theta_k, U_k(\Sigma_k - \sigma_k^2 I_{d_k})U_k^{\mathsf{T}} \\ &\qquad + \sigma_k^2 I_m\Big), \end{aligned}$$

where Equation (3) corresponds to sampling from a distribution $\mathcal{P}$ supported on the lower half of the sphere $\mathbb{S}^{(m-1)(m+2)/2}$ a projection matrix P_k that corresponds to a subspace, computing the dimension d_k as the trace of the subspace, and computing the subspace U_k from the projection. Equation (4) corresponds to sampling from a normal distribution subject to the projection constraint $U_k^{\mathsf{T}}\theta_k = 0$.

3. Posterior Sampling

In this section, we provide an efficient algorithm for sampling the model parameters from the posterior distribution. Sampling

directly from a joint posterior distribution of all the parameters is intractable and we will use Markov chain Monte Carlo methods for sampling. For most of the parameters, we can sample from the posterior using a Gibbs sampler. This is not the case for sampling from the posterior distribution over projection matrices with prior $\mathcal{P}$ on the sphere $\mathbb{S}^{(m-1)(m+2)/2}$. The prior $\mathcal{P}$ should place more mass on cross-sections of the sphere corresponding to lower dimensions d_k. Sampling efficiently from a joint distribution on d_k, P_k is difficult. We will address this problem by using a Gibbs posterior (Jiang and Tanner 2008) to sample the projection matrices. We first state the Gibbs posterior we use to sample U_k and θ_k efficiently and the rationale for this form of the posterior. We then close with the sampling algorithm for all the model parameters.

It is not obvious how to place a prior on the sphere $\mathbb{S}^{(m-1)(m+2)/2}$ that will allow for efficient sampling. We follow the idea of a Gibbs posterior to design an efficient sampler. While a standard posterior takes the form of

$$\text{posterior} \propto \text{prior} \times \text{likelihood},$$

a Gibbs posterior replaces the likelihood with a loss or risk function that depends on both the data as well as the parameter of interest. In our case, the loss function is given by

$$\begin{aligned} &L(P_{[1:K]}, \theta_{[1:K]}, X) \qquad (5)\\ &= \frac{1}{n}\sum_{i=1}^{n}\Big[\min_{k=1,\ldots,K}\Big(\|P_k(x_i-\theta_k)-(x_i-\theta_k)\|^2+\operatorname{tr}(P_k)\Big)\Big]\\ &= \frac{1}{n}\sum_{i=1}^{n}\Big[\min_{k=1,\ldots,K}(e_{ik}+d_k)\Big], \end{aligned}$$

where e_{ik} is the residual error for the ith sample given by the kth subspace with the error defined by our likelihood model. The above loss function corresponds to computing for each sample the residual error to the closest subspace weighted by the dimension of the subspace. The penalty weighting the dimension of the subspace enforces a prior that puts more mass on subspaces of lower dimension. Given the likelihood or loss function, we state the following Gibbs posterior:

$$\begin{aligned} g(P_{[1:K]}, \theta_{[1:K]}|X) \propto \exp\big(-n\psi L(P_{[1:K]}, \theta_{[1:K]}, X)\big)\\ \pi(P_{[1:K]})\pi(\theta_{[1:K]}), \qquad (6) \end{aligned}$$

where ψ is a chosen temperature parameter. Gibbs posterior is simply a loss-oriented alternative to the likelihood based posterior distribution. It is traditionally used to account for model misspecification. Here, the Gibbs posterior is used to avoid overfitting by arbitrarily increasing the dimension of the subspace and for computational efficiency in sampling.

3.1. Sampling $U_{[1:K]}$ and $\theta_{[1:K]}$ from the Gibbs Posterior

In this subsection, we outline our procedure for sampling the model parameters $U_{[1:K]}$ and $\theta_{[1:K]}$ using a Metropolis–Hastings algorithm with a modified version of random walk on the sphere. We first state a few facts that we will use. First, there is a deterministic relation between U_k and P_k, so given P_k we can compute U_k. Also recall that not every point sampled from $\mathbb{S}^{(m-1)(m+2)/2}$ qualifies as an image of a subspace. Given a point $s_k^0 \in \mathbb{S}^{(m-1)(m+2)/2}$ we denote the subspace corresponding to this point as $P_k = \varphi^{-1}(s_k^0)$, which is the closest projection matrix to s_k^0 corresponding to a subspace. The procedure to compute P_k from s_k^0 is given in Theorem 2. We obtain U_k corresponding to the top d_k eigenvectors of P_k where d_k is the trace of P_k.

We now state two procedures. The first procedure initializes the parameters $U_{[1:K]}$ and $\theta_{[1:K]}$. The second procedure computes the ℓth sample of the parameters.

The first procedure which we denote as **Initialize**$(U_{[1:K]}, \theta_{[1:K]})$ proceeds as follows:

1. Draw $\sigma \sim \mathfrak{S}_K$, the symmetric group of permutations on K elements.
2. For $i = 1, \ldots, K$,
   (a) draw $z^0_{\sigma(i)} \sim \mathcal{N}_{m(m+1)/2}(0, \tau I)$;
   (b) compute $s^0_{\sigma(i)} = (\sqrt{m(m+1)/8})z^0_{\sigma(i)}/\|z^0_{\sigma(i)}\| + \varphi(\frac{1}{2}I_m)$;
   (c) compute $P^0_{\sigma(i)} = \varphi^{-1}(s^0_{\sigma(i)})$;
   (d) compute $d^0_{\sigma(i)} = \operatorname{tr}(P^0_{\sigma(i)})$;
   (e) compute $U^0_{\sigma(i)}$ as the top $d^0_{\sigma(i)}$ eigenvectors of $P^0_{\sigma(i)}$;
   (f) draw $\beta^0_{\sigma(i)} \sim \mathcal{N}(0, I_m)$;
   (g) compute $\theta^0_{\sigma(i)} = (I_m - P^0_{\sigma(i)})\beta^0_{\sigma(i)}$.

The first step permutes the order we initialize the K components. Step (a) samples a point from a multivariate normal with the dimension of the sphere. In Step (b) we normalize the sampled point, recenter it, and embed it onto the sphere $\mathbb{S}^{m(m+1)/2}$. In Step (c), we compute the projection matrix by computing the closest subspace to the embedded point computed in Step (b). Given the projection matrix we compute the dimension in Step (d) and the basis of the subspace in Step (e). Steps (f) and (g) we compute the θ parameters.

The second procedure which we denote as **Update**$\big(U^{(\ell)}_{[1:K]}, \theta^{(\ell)}_{[1:K]}\big)$ computes the ℓth sample as follows:

1. Draw $\sigma \sim \mathfrak{S}_K$, the symmetric group of permutations on K elements.
2. For $i = 1, \ldots, K$,
   (a) draw $z_{\sigma(i)} \sim \mathcal{N}_{m(m+1)/2}(z^{(\ell-1)}_{\sigma(i)}, \tau I)$;
   (b) compute $s_{\sigma(i)} = (\sqrt{m(m+1)/8})z_{\sigma(i)}/\|z_{\sigma(i)}\| + \varphi(\frac{1}{2}I_m)$;
   (c) compute $P_{\sigma(i)} = \varphi^{-1}(s_{\sigma(i)})$;
   (d) compute $d_{\sigma(i)} = \operatorname{tr}(P_{\sigma(i)})$;
   (e) compute $U_{\sigma(i)}$ as the top $d_{\sigma(i)}$ eigenvectors of $P_{\sigma(i)}$;
   (f) draw $u \sim \text{Unif}[0, 1]$;
   (g) set
   $$P_{[1:K]} = \big[P^{(\ell-1)}_{[1:K]-\sigma(i)}, P_{\sigma(i)}\big];$$
   (h) set
   $$\theta_{[1:K]} = \big[\theta^{(\ell-1)}_{[1:K]-\sigma(i)}, (I_m - U_{\sigma(i)}U^T_{\sigma(i)})\theta^{(\ell-1)}_{\sigma(i)}\big];$$
   (i) compute the acceptance probability
   $$\alpha = \min\left(1, \frac{\exp\big(-n\psi L(P_{[1:K]}, \theta_{[1:K]}, X)\big)}{\exp\big(-n\psi L(P^{(\ell-1)}_{[1:K]}, \theta^{(\ell-1)}_{[1:K]}, X)\big)}\right);$$
   (j) set
   $$\big(U^{(\ell)}_{\sigma(i)}, z^{(\ell)}_{\sigma(i)}\big) = \begin{cases} \big(U_{\sigma(i)}, z_{\sigma(i)}\big) & \text{if } \alpha > u,\\ \big(U^{(\ell-1)}_{\sigma(i)}, z^{(\ell-1)}_{\sigma(i)}\big) & \text{otherwise;} \end{cases}$$

(k) draw $\beta_{\sigma(i)} \sim \mathcal{N}_m(\beta_{\sigma(i)}^{(\ell-1)}, I_m)$;
(l) compute $\theta_{\sigma(i)} = (I_m - P_{\sigma(i)}^{\ell-1})\beta_{\sigma(i)}$;
(m) draw $u \sim \text{Unif}[0,1]$;
(n) set

$$\theta_{[1:K]} = \left[\theta_{[1:K]-\sigma(i)}^{(\ell-1)}, \theta_{\sigma(i)}\right];$$

(o) compute the acceptance probability

$$\alpha = \min\left(1, \frac{\exp\left(-n\psi L(P_{[1:K]}^{(\ell-1)}, \theta_{[1:K]}, X)\right)}{\exp\left(-n\psi L(P_{[1:K]}^{(\ell-1)}, \theta_{[1:K]}^{(\ell-1)}, X)\right)}\right);$$

(p) set

$$\left(\theta_{\sigma(i)}^{(\ell)}, \beta_{\sigma(i)}^{(\ell)}\right) = \begin{cases} \left(\theta_{\sigma(i)}, \beta_{\sigma(i)}\right) & \text{if } \alpha > u, \\ \left(\theta_{\sigma(i)}^{(\ell-1)}, \beta_{\sigma(i)}^{(\ell-1)}\right) & \text{otherwise.} \end{cases}$$

Many steps of this procedure are the same as the first procedure with the following exceptions. In Steps (a) and (k) we are centering the random walk to the previous values of $z_{\sigma(i)}$ and $\beta_{\sigma(i)}$ respectively. Step (g) updates the set of K projection matrices by replacing the ith projection matrix in the set with the proposed new matrix. Step (h) is analogous to Step (g) but for the set of θ vectors. In Step (j) we update the subspace and in Step (p) we update the θ vector.

3.2. Sampling Algorithm

We now state the algorithm we use to sample from the posterior. To simplify notation we work with precision matrices $J_k = \Sigma_k^{-1}$ instead of the inverse of covariance matrices for each mixture component. Similarly, we work with the precision of the k-th component γ_k instead of the inverse of the variance, $\gamma_k = \sigma_k^{-2}$.

The following procedure provides posterior samples:

1. Draw $U_{[1:K]}^{(0)}, \theta_{[1:K]}^{(0)}, d_{[1:K]}^{(0)} \sim \textbf{Initialize}(U_{[1:K]}, \theta_{[1:K]})$.
2. Draw $J_{k(j_k)} \sim \text{Ga}(a, b)$ for $k = 1, \ldots, K$ and $j_k = 1, \ldots, d_k^{(0)}$.
3. For $t = 1, \ldots, T$,

(a) for $i = 1, \ldots, n$ and $k = 1, \ldots, K$, compute

$$e_{ik} = \left\| P_k^{(t-1)}\left(x_i - \theta_k^{(t-1)}\right) - \left(x_i - \theta_k^{(t-1)}\right)\right\|^2;$$

(b) for $i = 1, \ldots, n$, set

$$w_i = \left(\frac{\exp(-\kappa e_{i1})}{\sum_{j'=1}^K \exp(-\kappa e_{ij'})}, \ldots, \frac{\exp(-\kappa e_{iK})}{\sum_{j'=1}^K \exp(-\kappa e_{ij'})}\right);$$

(c) for $i = 1, \ldots, n$, draw $\delta_i \sim \text{Mult}(w_i)$;
(d) update for $k = 1, \ldots, K$ each $\mu_k^{(t)} \sim \mathcal{N}(m_k^*, S_k^*)$ where

$$S_k^* = \left(n_k J_k^{(t-1)} + \lambda^{-1} I\right)^{-1},$$
$$m_k^* = S_k^*\left(J_k^{(t-1)} U_k^{(t-1)\mathsf{T}} \sum_{\delta_i = k} x_i\right),$$

and $n_k = \#\{i : \delta_i = k\}$;
(e) update for $k = 1, \ldots, K$, and each $\gamma_k^{(t)} \sim \text{Ga}(a_k^*, b_k^*)$,

$$a_k^* = n_k(m - d_k) + a,$$
$$b_k^* = b + \frac{n_k}{2}(\theta_k^{(t-1)})^{\mathsf{T}}\theta_k^{(t-1)} + \sum_{\delta_i = k}\left(\frac{1}{2}x_i^{\mathsf{T}}x_i - x_i^{\mathsf{T}}U_k^{(t-1)}U_k^{(t-1)\mathsf{T}}x_i\right) - \theta_k^{(t-1)\mathsf{T}}\sum_{\delta_i = k} x_i;$$

(f) update for $k = 1, \ldots, K$, and $j_k = 1, \ldots, d_k^{(t-1)}$,

$$J_{k(j_k)}^{(t)} \sim \text{Ga}\left(\frac{n_k}{2} + a, b + \frac{1}{2}\sum_{\delta_i = k}\left(U_k^{(t-1)\mathsf{T}}x_i - \mu_k^{(t)}\right)_{j_k}^2\right),$$

where $(u)_j$ denotes the jth element of the vector u;
(g) draw

$$U_{[1:K]}^{(t)}, \theta_{[1:K]}^{(t)}, d_{[1:K]}^{(t)} \sim \textbf{Update}\left(U_{[1:K]}^{(t-1)}, \theta_{[1:K]}^{(t-1)}\right).$$

The update steps for μ, σ^2, and Σ are (d), (e), and (f), respectively, and are given by the conditional probabilities given all other variables. Steps (a), (b), and (c) assign the latent membership variable to each observation based on the distance to the K subspaces. We set the parameter κ very large which effectively assigns membership of each x_i to the subspace with the least residual error.

Here, we suggest some heuristic guidelines for implementation of the algorithm. Two algorithmic parameters of a Metropolis–Hastings (MH) algorithm for the Gibb's posterior distribution are the proposal distribution and temperature, which are often abbreviated or omitted in the presentation of algorithmic details. We opted for a strategy to sequentially determine the two parameters through a burn-in period to make our MH sampler draw a moderate number of informative samples. In the first stage of burn-in, the proposal variance parameter τ is fixed and temperature is chosen via a line search on a log-scale grid 10^{-20} to 10^{20} until the acceptance ratio is bound in a wide range. Once the temperature is determined, the second stage is to adjust proposal variance τ until the acceptance ratio falls in a narrower range during the burn-in period. We observed that the two-stage parameter tuning does not heavily depend on initialization of τ so we fixed the initial value $\tau = 1$ in our experiments. We used the 10%–90% as a wide range for tuning the temperature at the first stage and the 25%–45% range for tuning the step-size parameter in the latter stage, where the choice of these values are subject to a user's decision.

4. Specification of a Topic Model: A Generative Model on the Stiefel Manifold

The idea behind topic modeling is to specify a generative model for documents where the model parameters provide some intuition about a collection of documents. A common representation for documents is what is called a "bag of words" model where the grammar and structure of a document are ignored and a document is just a vector of counts of words (Blei, Ng, and Jordan 2003; Deerwester et al. 1990; Hoffman 1999). A natural generative model for collections of documents is an admixture of topics where each topic is a multinomial distribution over words. This model is called a latent Dirichlet allocation (LDA) model (Pritchard, Stephens, and Donnelly 2000; Blei, Ng, and Jordan 2003). We will propose a slight variation of the LDA model later in this section which is a direct extension/application of a mixture of subspaces.

The generative model specified by LDA considers a vocabulary of size V and each of D documents is a mixture of multinomials where each multinomial corresponds to one of K topics and each topic has a different multinomial distribution over the words in the vocabulary. In a spherical admixture model

(SAM) (Reisinger et al. 2010) the vector of word counts in each document is transformed by centering at zero and normalizing to unit length. The idea behind a SAM is to represent word frequencies as directional distributions on a hypersphere. The advantage of a SAM is that one can simultaneously model both frequency as well as presence/absence of words, whereas an LDA model can only model frequency. There is empirical evidence of greater accuracy in using a SAM for sparse data such as text (Banerjee and Dhillon 2005; Zhong and Ghosh 2005).

We extend the SAM model in two important ways, first by ensuring all the topics are orthogonal. The logic behind orthogonality constraints in the topics is to avoid the empirically observed problem of redundant topics. Removing stop words can mitigate redundancy but it is not always known a priori what words should be stop words. For example, the word "topic" should be a stop word in a corpus of papers on topic modeling. In Hoff (2009), a Bayesian model for an orthogonal SAM is specified and a posterior sampling procedure is developed. A key insight of this paper was how to efficiently simulate from the set of orthonormal matrices using the matrix Bingham–von Mises–Fisher distribution. For an orthogonal SAM model, the K topics on V words were modeled using the matrix Bingham–von Mises–Fisher distribution which is on the Stiefel manifold, $\mathcal{V}(V,K)$. The second extension is to infer the number of topics K. While the orthogonality constraints help with the interpretation of topics and remove redundant topics, topics with low posterior mixture probabilities and low coherence can still appear. This is mainly driven by misspecification of the number of topics.

We now state a novel SAM model that enforces orthonormal columns as well as allows for the inference of the number of topics. Using the Conway embedding allows us to place a joint prior over the number of topics as well as the word frequencies in each topic. We are able to sample from Stiefel manifolds of variable intrinsic dimension K by coupling draws from the von Mises–Fisher distribution with the map given by inverting the Conway embedding. This allows us to avoid using the matrix Bingham–von Mises–Fisher distribution.

We provide some intuition for our SAM with orthogonality constraints as well as useful notation before we specify the model. We will simulate a topic matrix ϕ where the columns of the matrix are topics $\{\phi_k\}_{k=1}^K$ and the number of topics K is random. Orthogonal matrices of the topics are sampled from a distribution over Stiefel manifolds $\mathcal{V}(V,K)$, with fixed ambient dimension V (the number of words) and variable embedding dimension K. For each document probability vector of topic proportions θ_d over the K topics is generated. The ℓ_2-normalized unit vector v_d representing normalized word frequencies for each document is generated from the topic proportions θ_d and the topic matrix ϕ. The following notation and concepts will be used in the generative model. We denote $\mathbb{S}$ as the Conway sphere $\mathbb{S}^{(V-1)(V+2)/2}$ which is the collection of orthogonal subspaces of variable dimension embedded into a sphere. We denote $\varphi(\cdot)$ and $\varphi^{-1}(\cdot)$ as the embedding function and its inverse respectively. We denote $\text{vMF}_{\mathbb{S}}$ as the von Mises–Fisher distribution over the Conway sphere $\mathbb{S}^{(V-1)(V+2)/2}$ and vMF_V as the von Mises–Fisher distribution over the unit sphere $\mathbb{S}^V$. Given the topics matrix and the topic proportions θ_d for a document, a spherical average of the topics with respect to the proportions is the admixed parameter that models the combination of topics in document and is computed by $\text{avg}(\phi,\theta_d) = \frac{\phi\theta_d}{||\phi\theta_d||}$. We did not use the Buss-Fillmore spherical average due to computational considerations. Given a vocabulary of size V and D documents, the ℓ_2-normalized unit vector v_d for each document is specified by the following hierarchical model:

$$\begin{aligned}
\mu|\kappa_0 &\sim \text{vMF}_{\mathbb{S}}(m,\kappa_0) && \text{(corpus average)},\\
\eta|\mu,\xi &\sim \text{vMF}_{\mathbb{S}}(\mu,\xi) && \text{(embedded orthogonal topics)},\\
(\phi,K) &= \varphi^{-1}(\eta) && \text{(orthogonal topics and number of topics)}\\
\theta_d|\alpha &\sim \text{Dir}_K(\alpha) && \text{(topic proportions for each document)}\\
\bar{\phi}_d &= \text{avg}(\phi,\theta_d) && \text{(spherical average, admixed parameter)}\\
v_d|\bar{\phi}_d,\kappa &\sim \text{vMF}_V\left(\bar{\phi}_d,\tau\right) && \text{(generates a document vector)}.
\end{aligned} \tag{7}$$

The main difference in the above model and prior models on spheres (Hoff 2009; Reisinger et al. 2010) is that instead of sampling topics from the embedding on the Conway sphere, fixed-K topic vectors were each sampled from a von Mises–Fisher distribution over $\mathbb{S}^V$. In Reisinger et al. (2010), the vectors were not constrained to be orthogonal, and in Hoff (2009) an efficient procedure is given to simulate these K orthogonal vectors. The Conway sphere in this case is extremely high dimensional and there is computational utility in reducing the vocabulary size.

As in the mixture of subspaces model, we require a prior that will place greater weight on models with fewer topics as increasing the number of topics will result in a better fit with respect to the likelihood. In the same spirit as Section 3, we specify a Gibbs posterior to place a prior on the Conway sphere that can be efficiently be sampled from and favors models with fewer topics. We specify the following loss function for each document vector v_d:

$$L\left(\phi,K|\{v_d\}_{d=1}^D,\tau\right) = \frac{2DK}{V} - \sum_{d=1}^{D}\left(\tau\bar{\phi}_d^T v_d\right) \tag{8}$$

and corresponding Gibbs posterior

$$g(\phi,K|\{v_d\}_{d=1}^D,\tau) \propto \exp\Big(-\psi DL\times\left(\phi,K|\{v_d\}_{d=1}^D,\tau\right)\Big)\pi(\phi). \tag{9}$$

The maximum penalty above is D and would counterbalance a perfect fit to each document with a loss value of zero. Using the Gibbs posterior allows us to skip the step of estimating the corpus average parameter μ since the remaining parameters and μ are conditionally independent given the topics. We set the temperature parameter ψ using out-of-sample fits on a random search over $\log(\psi)\in[-10,10]$.

The remaining parameters of Model (7) are estimated using the same sampling steps as in a standard SAM once the topics and number of topics are sampled. The high-dimension of the Conway sphere can result in slower mixing of the topics and it is of interest to explore EM or Hamiltonian Monte Carlo approaches for computational gain.

5. Results on Real and Simulated Data

We illustrate the utility of the embedding through an extensive simulation study and real data analysis. The simulation study

involves synthetic data with simple geometric structure to contrast the performance of our method with k-means clustering, a Gaussian mixture model, and a Bayesian factor model. We also compare performance of non-Bayesian subspace clustering methods. In the second example, we compare the performance of our model with a logistic model, a Gaussian mixture model, as well as a factor model on three supervised classification problems from the UCI machine learning repository (Bache and Lichman 2013). The last example compares the spherical topic model we developed to a latent Dirichlet allocation model on a corpus of NSF award abstracts (NSF 2010).

5.1. Lines Intersecting a Plane

Possibly the simplest example of a mixture of subspaces is a line puncturing a plane. We will use this example to illustrate basic properties of the mixture of subspaces as well as explore comparisons to comparable models. We will study how well we can cluster the observed points into those sampled from the plane or line respectively. We perform clustering comparison in two-folds. Our method is first compared against standard clustering algorithms. Also, we compare with popular non-Bayesian methods in the context of subspace clustering.

The mixture model for a line intersecting a plane in $\mathbb{R}^3$ comprises two components: a subspace $\mathbf{U}_1$ corresponding to a line and a subspace $\mathbf{U}_2$ corresponding to the plane. Although simple, this example can be challenging inference problem. To understand the effect of uncertainty of subspace measurements on accuracy of models we add isotropic noise around the subspaces via a precision parameter.

The data are specified by the following distribution with the following five values for the precision parameter of the isotropic noise around the subspaces, $\nu = [10, 5, 1, 0.5, 0.2]$:

Line	Plane
$U_1 \sim \text{Unif}(\text{V}(1,3))$,	$U_2 \sim \text{Unif}(\text{V}(2,3))$,
$\mu_1 \sim \mathcal{N}_1(0,1)$,	$\mu_2 \sim \mathcal{N}_2(0,I)$,
$\Sigma_1^{-1} \sim \text{TGa}(1,1,\nu)$,	$\text{diag}(\Sigma_2^{-1}) \overset{\text{iid}}{\sim} \text{TGa}(1,1,\nu)$,
$(I_3 - U_1U_1^\text{T})^{-1}\theta_1 \sim \mathcal{N}_3(0,I)$,	$(I_3 - U_2U_2^\text{T})^{-1}\theta_2 \sim \mathcal{N}_3(0,I)$,

where $\text{TGa}(1,1,\nu)$ is a left truncated Gamma truncated at precision ν. Given these parameters for the two mixture components, we specify the following two conditional distributions:

$$x|\text{Line} \overset{\text{iid}}{\sim} \mathcal{N}_3(U_1\mu_1 + \theta_1, U_1(\Sigma_1 - \sigma_1^2)U_1^\text{T} + \sigma_1^2 I),$$
$$x|\text{Plane} \overset{\text{iid}}{\sim} \mathcal{N}_3(U_2\mu_2 + \theta_2, U_2(\Sigma_2 - \sigma_1^2 I)U_2^\text{T} + \sigma_1^2 I).$$

We generated 500 observations from both the line and the plane, see Figure 1. For each of the five variance levels, 10 datasets were generated.

Our initial comparison set of models includes K-means clustering (K-means), a mixture of normals (GMM), a mixture of nonparametric factor models (MFM) (Carvalho et al. 2008), our mixture of subspaces model with variable dimensions (MSM), and our mixture of subspaces model with dimension fixed to $d = 2$ (MSM $d_1 = d_2 = 2$). For the subspace model we set the temperature parameter for the Gibbs posterior to 10^{-6}, and acceptance rate between 38% and 48% was achieved for the subspace and affine mean parameters. For the model-based clustering algorithms in this experiment, a holdout set of 50 observations from the line and plane is used for testing. Under the settings described, a comparison of clustering accuracy of the five models is summarized in Table 1. We report the range in clustering accuracy for each method on the holdout set over the ten runs. We conclude from Table 1 that (i) K-means performs poorly, (i) as the precision parameter increases, the performance of the GMM improves and starts to approach the MFM and MSM results, (iii) the MSM with variable dimension

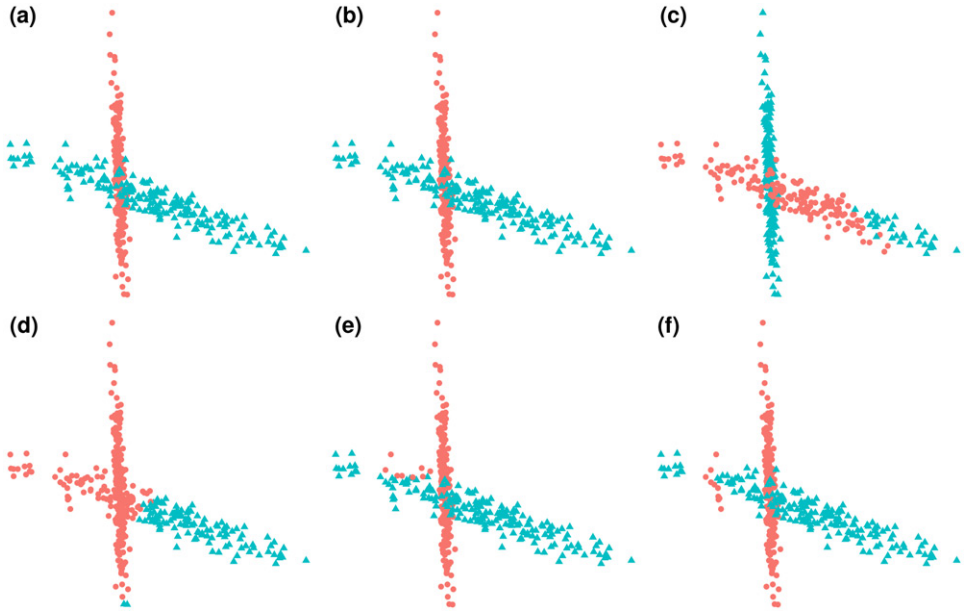


Figure 1. An example of one line and one plane in $\mathbb{R}^3$. The upper-left panel (a) displays the true cluster assignments and estimated clusterings are given for (b) MSM, (c) SSC, (d) K-means, (e) GMM, and (f) MFM algorithms.

Table 1. Range of cluster assignment accuracy of *K*-means and model-based algorithms for the simulated data.

Precision	MSM	MSM $d = (2, 2)$	GMM	MFM	*K*-means
0.2	(0.95, 0.99)	(0.92, 0.97)	(0.89, 0.97)	(0.95, 0.99)	(0.66, 0.77)
0.5	(0.90, 0.98)	(0.90, 0.96)	(0.87, 0.98)	(0.91, 0.99)	(0.64, 0.82)
1	(0.87, 0.98)	(0.86, 0.97)	(0.85, 0.98)	(0.88, 0.98)	(0.57, 0.70)
2	(0.87, 0.96)	(0.87, 0.95)	(0.87, 0.95)	(0.86, 0.98)	(0.64, 0.72)
5	(0.84, 0.97)	(0.84, 0.97)	(0.84, 0.97)	(0.83, 0.97)	(0.59, 0.80)

Table 2. Range of cluster assignment accuracy for subspace clustering algorithms.

Precision	MSM	EKSS	LRR	LRSC	LSR	SSC
0.2	(0.51, 0.59)	(0.51, 0.73)	(0.50, 0.57)	(0.49, 0.56)	(0.49, 0.55)	(0.49, 0.54)
0.5	(0.59, 0.87)	(0.49, 0.83)	(0.49, 0.63)	(0.49, 0.72)	(0.49, 0.70)	(0.49, 0.69)
1	(0.60, 0.87)	(0.49, 0.88)	(0.49, 0.69)	(0.57, 0.75)	(0.55, 0.75)	(0.50, 0.79)
2	(0.59, 0.98)	(0.51, 0.98)	(0.50, 0.79)	(0.53, 0.83)	(0.51, 0.82)	(0.50, 0.80)
5	(0.76, 1.00)	(0.50, 1.00)	(0.49, 0.99)	(0.52, 0.95)	(0.52, 0.95)	(0.51, 0.69)

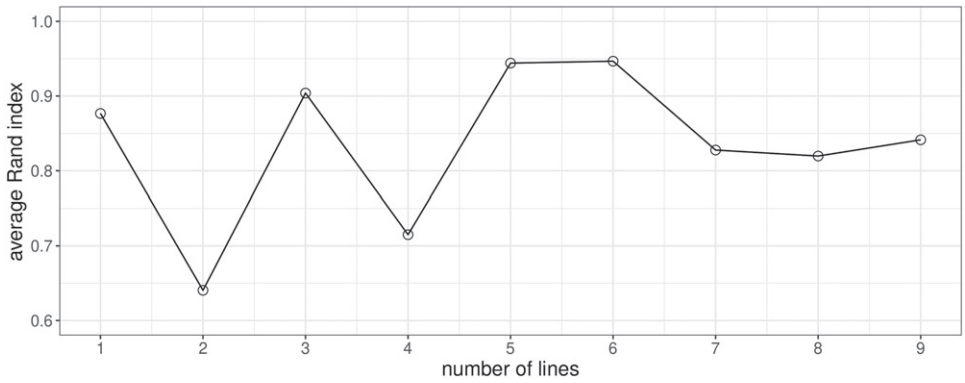


Figure 2. Average Rand index of 1 plane and *k* lines simulation.

outperforms the MSM with fixed dimension, and (iv) the MSM and MFM results are very similar. Note that the MSM provides richer geometric information including the intrinsic dimension of the subspace.

The second simulation examined recovering a line puncturing a plane and contrasts our method to popular subspace clustering methods, including Ensembles of *K*-Subspaces (EKSS) (Lipor et al. 2021), low-rank representation (LRR) (Liu, Lin, and Yu 2010), low-rank subspace clustering (LRSC) (Vidal and Favaro 2014), Least-squares regression (LSR) (Lu et al. 2012), and sparse subspace clustering (SSC) (Elhamifar and Vidal 2009). In this experiment, we use MSM with variable dimensions and the same settings as before. For the five subspace clustering algorithms compared, we use the simplest form of each model. For example, a cost function of the SSC algorithm has two regularization terms besides the sparse connectivity objective to penalize sparse outlying entries and noise. Since these penalties involve tuning hyperparameters, we use minimal formulation of cost functions. Furthermore, these methods are not model-based and out-of-sample prediction may not be straightforward so that we opt to measure and report clustering accuracy for the whole datasets, which are summarized in Table 2. The experiment shows that MSM performs well, comparable to EKSS which ensembles a large number of independent runs, while MSM enables to extract information about estimated dimensionalities of the fitted affine subspaces and predict cluster information for new observations. Except for $\nu = 0.2$, MSM outperforms all competing algorithms from which we can conclude that MSM handles noisy sample well unless it becomes extensive beyond an algorithm's capability. Furthermore, one noticeable observation is that it has the best worst-case performance across all levels of precision, leading to empirical guarantee of its worst-scenario performance.

Next, we examine beyond a line and a plane in $\mathbb{R}^3$—multiple subspaces and higher dimensional cases. We initially tried a specific example of two lines and one plane where 100 observations are drawn from each subspace. MSM reports mean Rand index of 0.9453 when $k = 3$ while $k = 2$ gives 0.7325 and $k = 4$ gives 0.7094, respectively, from 100 posterior samples. To observe the algorithm's capability to distinguish unions of lines from a plane, we performed another test where data consists of 1 plane and $(k-1)$ lines all randomly generated for $k = 2, 3, \ldots, 9$. For each subspace, 100 observations are drawn. Figure 2 shows the algorithm's capability to distinguish unions of lines from a plane where the increased number of lines does not hurt the performance much. One may wonder whether our model can recognize the union of two subspaces with dimension 1 as a single subspace of dimension 2 in this setup. Although any pair of perfect lines forms a plane in theory, we see that when the number of mixture components is properly chosen, the algorithm successfully distinguishes two lines.

Second, we assess how the model performs for a line and a plane example embedded in higher dimensions. To describe

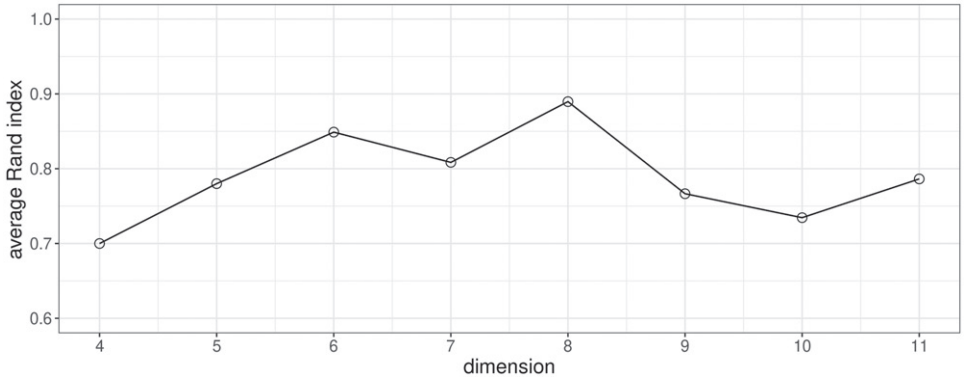


Figure 3. Performance of MSM on a line and a plane example with varying dimensions.

Table 3. Mean clustering assignment accuracy of a line and a plane example embedded in higher dimensions.

Dimension	Precision	MSM	EKSS	LRR	LRSC	LSR	SSC	GMM	*K*-means
5	0.2	0.538	0.592	0.513	0.510	0.509	0.510	0.565	0.531
	0.5	0.578	0.657	0.581	0.613	0.611	0.549	0.668	0.632
	1	0.792	0.631	0.595	0.711	0.705	0.623	0.604	0.651
	2	0.836	0.658	0.632	0.742	0.729	0.615	0.602	0.674
	5	0.878	0.676	0.608	0.716	0.706	0.650	0.661	0.625
10	0.2	0.572	0.566	0.511	0.510	0.511	0.505	0.568	0.578
	0.5	0.621	0.676	0.591	0.614	0.612	0.547	0.666	0.644
	1	0.632	0.656	0.597	0.687	0.682	0.602	0.690	0.684
	2	0.684	0.667	0.610	0.734	0.719	0.665	0.604	0.662
	5	0.866	0.623	0.605	0.736	0.728	0.683	0.694	0.630
25	0.2	0.604	0.604	0.516	0.519	0.518	0.505	0.566	0.607
	0.5	0.671	0.713	0.582	0.606	0.599	0.551	0.737	0.673
	1	0.683	0.707	0.626	0.691	0.674	0.638	0.682	0.673
	2	0.858	0.682	0.617	0.724	0.718	0.664	0.688	0.665
	5	0.911	0.683	0.636	0.771	0.761	0.694	0.755	0.617
50	0.2	0.638	0.610	0.518	0.512	0.510	0.510	0.560	0.629
	0.5	0.790	0.643	0.569	0.610	0.609	0.547	0.674	0.725
	1	0.845	0.696	0.606	0.677	0.671	0.592	0.723	0.688
	2	0.865	0.683	0.603	0.719	0.707	0.649	0.671	0.768
	5	0.887	0.682	0.619	0.755	0.742	0.686	0.658	0.695
100	0.2	0.624	0.603	0.524	0.526	0.526	0.510	0.613	0.604
	0.5	0.751	0.651	0.620	0.679	0.668	0.608	0.711	0.721
	1	0.748	0.657	0.594	0.721	0.696	0.617	0.692	0.674
	2	0.850	0.620	0.570	0.634	0.628	0.580	0.713	0.667
	5	0.913	0.633	0.621	0.823	0.810	0.779	0.694	0.657

an embedding procedure, given an arbitrary projection matrix $P \in \mathbb{R}^{p\times 3}$ where $P^\top P = I_3$ is an identity matrix, we map a generated dataset of one line and one plane in $\mathbb{R}^3$ into $\mathbb{R}^p$ by post-multiplying $P^\top(PP^\top)^+$ where A^+ is a pseudo-inverse of an arbitrary matrix A. Figure 3 shows that increasing the dimension of an ambient space p does not hurt the performance much as long as the underlying structure is low-dimensional.

Table 3 summarizes further experiments that show competitive performance of our model against non-Bayesian subspace clustering algorithms as well as GMM and K-means in higher dimensions $p = 5, 10, 25, 50, 100$. As the magnitude of white noise gets smaller, MSM shows a tendency to outperform competing algorithms. Along with the immediate benefit, one key feature of our model is its ability to detect the dimensionality of the subspaces unlike non-Bayesian algorithms that do not require explicit recovery of low-dimensional structures. When the number of cluster is properly chosen, we observed light level of errors in estimated dimensions for each subspace from the generated data although increased amount of additive noise may affect discovering true dimensionalities.

Remark 1. The model is empirically invariant to the scaling of the datasets. Figure 4 shows results from applying MSM algorithm on the same dataset (the line intersecting plane example) scaled by constants from 2 to 20, which concludes that scale of the data does not degrade performance of the algorithm as long as the assumption on the intrinsic structure of the data is valid.

5.2. Classification on UCI Data

To study the clustering performance on more realistic data, we examined the classification accuracy on three datasets from the UCI Data Repository: the Statlog Vehicle Silhouettes data (Siebert 1987), the Wisconsin Breast Cancer data (Mangasarian and Wolberg 1990), and the Statlog Heart data (Detrano et al. 1989). Our metric of success on all three data was holdout classification accuracy. We compared five models: a (multinomial) logit model (Logit), our mixture of subspaces model with variable dimensions (MSM), our mixture of subspaces model

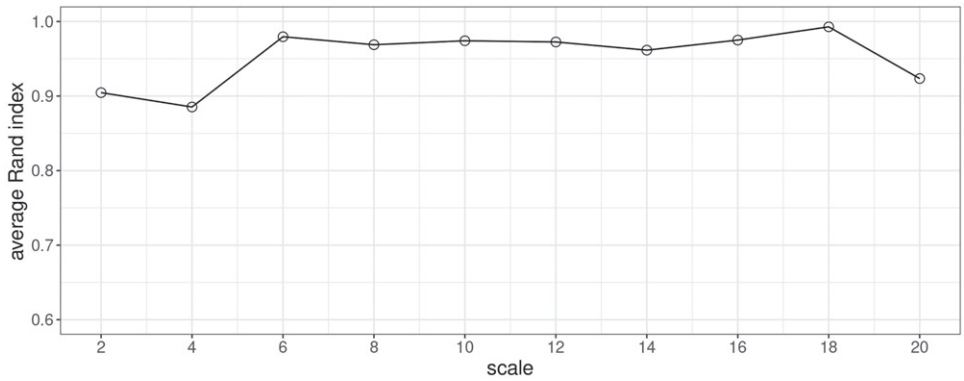


Figure 4. Performance of MSM with rescaled data.

Table 4. Range of cluster assignment accuracy for the three datasets using five models on holdout data.

Dataset	MSM	MSM $d = (5, 2, 2)$	EKSS	GMM	Logit	MFM
Breast	(0.89, 0.94)	(0.83, 0.89)	(0.79, 0.93)	(0.64, 0.70)	(0.78, 0.86)	(0.90, 0.96)
Heart	(0.77, 0.81)	(0.73, 0.77)	(0.63, 0.79)	(0.56, 0.60)	(0.72, 0.78)	(0.80, 0.82)
Vehicle	(0.77, 0.83)	(0.75, 0.80)	(0.70, 0.84)	(0.74, 0.79)	(0.46, 0.59)	(0.76, 0.85)

Table 5. Uncertainty in the dimension of the subspaces.

Dataset	Class 1	Class 2	Class 3	Class 4
Breast	{5, 0.45; 6, 0.53}	{4, 0.66; 5, 0.34}	–	–
Heart	{1, 0.10; 2, 0.78; 3, 0.12}	{1, 0.79; 2, 0.21}	–	–
Vehicle	{1, 0.69; 2, 0.31}	{1, 0.76; 2, 0.24}	{1, 0.78; 2, 0.22}	{1, 0.93; 2, 0.07}

with fixed dimensions (MSM $d = (5, 2, 2)$), and a mixture of nonparametric factor models (MFM) (Carvalho et al. 2008).

For the subspace models, the temperature parameter of the Gibbs posterior was set to obtain an acceptance ratio in the range of 20–40% during the burn-in period. We did not use a cross-validation criterion to set the temperature parameter due to computational burden. To compute predictive accuracy, we use the maximum a posterior estimate of our MCMC runs to classify a new point.

The Heart Data Set contains 270 observations of two classes with 13 covariates, the Vehicle Data Set contains 846 observations of four classes on 18 covariates, and the Breast Cancer Data Set contains 569 observations of two classes on 30 covariates. For each dataset, we measured the test error on a holdout set of 10% of the data. We repeated the test error estimates 10 times and report the range in test errors in Table 4 for the results.

We conclude from Table 4 that (i) the MSM with variable dimension outperforms the MSM with fixed dimension, (ii) the mixture of subspaces and mixture of factors perform as well or better than the logit model which is the only supervised method, (iii) MSM and MFM have very comparable performance.

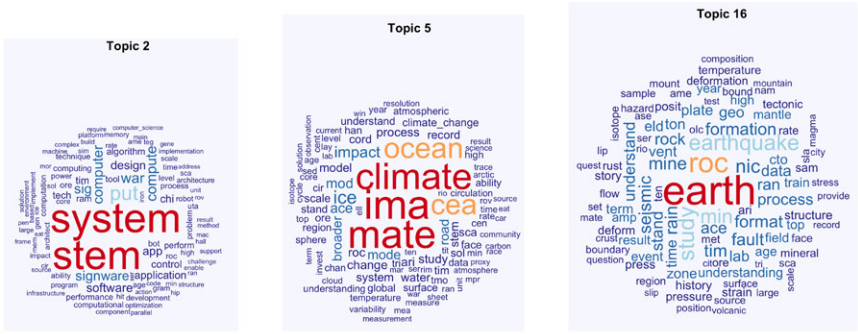
Our subspace model allows for an estimate of the dimension of the linear subspaces which is not possible in the the Bayesian mixture models proposed in Page, Bhattacharya, and Dunson (2013), the nonparametric mixture of factor models proposed in Carvalho et al. (2008), or the penalized cost-based mixture of subspaces model (Lerman and Zhang 2010). Table 5 states the uncertainty in estimates of the subspace dimension, the notation we use is $\{\dim_1, \Pr(\dim_1); \dim_2, \Pr(\dim_2); ...\}$.

5.3. Analysis of NSF Award Abstracts

In this subsection, we compare the topic model proposed in Section 4 to the standard latent Dirichlet allocation model. The corpus we use to compare the two methods consists of 13,092 abstracts from NSF awards in 2010 (NSF 2010). The vocabulary was constructed using the tokenizer from the Mallet package with bi-gram extraction (McCallum 2002). The vocabulary was then reduced to the terms that were within the top 10% term frequency inverse document frequency metric (Wu et al. 2008) and occurred in at least five documents. The resulting vocabulary consisted of 78,343 terms. The average length of the documents after trimming the vocabulary was 379 words.

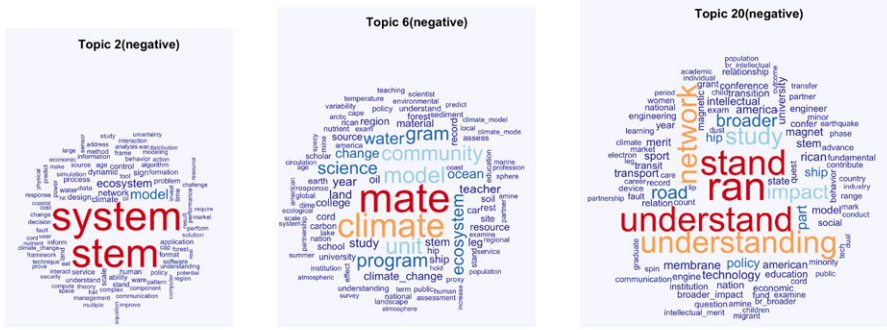
We compared the spherical topic model specified in Equation (7) to the a standard LDA model with 20 fixed topics. It was pointed out in Reisinger et al. (2010) that a direct comparison of topic models and spherical topic models is not possible/meaningful. We instead examine the topic coherence and most relevant words in each topic. Example word clouds are displayed in Figures 5(a) and 5(b,c).

In the spherical topic models, positive and negative parts of the topic vectors tend to be thematically coherent (also noted by Reisinger et al. 2010). This is interesting because of a richer representation of the topics. One interesting result of the model specified in Equation (7) on the NSF abstract data are that the positive and negative components of the topics tend to relate to broader impact terms and field or discipline terms, respectively. It may not be surprising that writers of grants from different disciplines use different goals in broader impacts. However, the spherical topic model structure gives us a tool for making thematic connections that would not be obvious from simply



(a) LDA results. The size of the terms indicate the relevance to the topic

(b) Spherical topic model results for positively weighed topics. The size of the terms indicate the relevance to the topic.



(c) Spherical topic model results for negatively weighed topics.

Figure 5. Comparison of world clouds for LDA versus the spherical topic model on the NSF abstract data.

looking at the top terms of a topic. Unlike standard mixture models, our spherical topic model allows for the inference on the number of topics with the support of the density over the number of topics ranging from 26 topics to 34 topics centered around 30 topics.

6. Discussion

We present a method for learning or inferring mixtures of linear subspaces of different dimensions. We show how this model can be trivially adapted for admixture modeling. The key idea in our procedure is based on the observation that subspaces of different dimensions can be represented as points on a sphere, which is very useful for inference. The utility of this representation is that sampling from a sphere is straightforward. There exists a distance between subspaces of different dimensions that is differentiable and can be computed using principal angles, allowing us to avoid MCMC algorithms that jump between models of different dimensions. We suspect that this idea of embedding or representing models of different dimensions by embedding them into a common space endowed with a distance metric that allows for ease of computation and sampling as well as nice analytic properties may also be of use in other settings besides subspace models.

One of the potential limitations of our model is that it cannot handle scenarios when the ambient dimension of the data is super huge. For example, in a gene expression dataset, it is not uncommon to collect data with more than 20k genes where the intrinsic dimension of the subspace is much smaller. Future work will be devoted to extending our model to this setup. One potential solution is to impose a prior that induces much stronger sparsity on the subspace dimensions. Scaling our estimation procedure to higher dimensions and more samples will also require greater computational efficiency and an EM-type algorithm for this model holds promise. It is also of interest to examine if we can replace the Gibbs posterior with an efficient fully Bayesian procedure.

Acknowledgments

We thank to the editor, the associate editor, and three reviewers for their valuable comments which have led to substantial improvement of our article. SM and BST thank to Robert Calderbank, Daniel Runcie, and Jesse Windle for useful discussions.

Funding

SM is pleased to acknowledge support from grants NIH (Systems Biology) 5P50-GM081883, AFOSR FA9550-10-1-0436, and NSF CCF-1049290. BST is pleased to acknowledge support from NSF grant DMS-1127914 to the Statistics and Applied Mathematics Institute. The work of LHL is partially supported by NSF IIS-1546413, NSF DMS-1854831, and DARPA HR00112190040. The contribution of LL was supported by IIS 1663870, DMS-1654579, DMS-2113642 and a grant R01ES017240 from the National Institute of Environmental Health Sciences (NIEHS) of the National Institute of Health (NIH).

ORCID

Kisung You http://orcid.org/0000-0002-8584-459X
Sayan Mukherjee http://orcid.org/0000-0002-6715-3920

References

Amari, S. (1982), "Differential Geometry of Curved Exponential Families — Curvatures and Information Loss," *Annals of Statistics*, 10, 357–385. [337]

Ashikhmin, A., and Calderbank, A. R. (2003), " Grassmannian Packings From Operator Reed-Muller Codes," *IEEE Transactions on Information Theory*, 56, 5689–5714. [337]

Bache, K., and Lichman, M. (2013), "UCI Machine Learning Repository." Available at: *http://archive.ics.uci.edu/ml*. [344]

Banerjee, A., Dhillon, I. S., Ghosh, J., Sra, S. (2005), "Clustering on the Unit Hypersphere Using von Mises-Fisher Distributions," *Journal of Machine Learning Research*, 6, 1345–1382. [343]

Belkin, M., and Niyogi, P. (2003), "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," *Neural Computation*, 15, 1373–1396. [337]

Bendich, P., Mukherjee, S., and Wang, B. (2012), "Local Homology Transfer and Stratification Learning," *ACM-SIAM Symposium on Discrete Algorithms*. [337]

Blei, D. M., Ng, A.Y., and Jordan, M. I. (2003), Latent Dirichlet Allocation *Journal of Machine Learning Research*, 3, 993– 1022. [337,342]

Carvalho, C.M., Chang, J., Lucas, J.E., Nevins, J.R., Wang, Q., West, M. (2008), "High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics," *Journal of the American Statistical Association*, 103, 1438–1456. [344,347]

Chandra, K., Canale, A., and Dunson, D. (2020), "Escaping the Curse of Dimensionality in Bayesian Model Based Clustering," arXiv e-prints 2006.02700. [337]

Chen, A., De, A., and Vijayaraghavan, A. (2021), "Learning a Mixture of Two Subspaces Over Finite Fields," *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, PMLR , pp. 481–504. [337]

Conway, J.H., Hardin, R.H., and Sloane, N.J.A. (1996), "Packing Lines, Planes, Etc.: Packings in GRassmannian Spaces," *Experiment. Math.*, 5, 83–159. [338,339,340]

Cook, R. (2007), "Fisher Lecture: Dimension Reduction in Regression," *Statistical Science*, 22, 1–26. [337]

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., Froelicher, V. (1989), "International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease," *American Journal of Cardiology*. 604, 304–310. [346]

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T.K., Harshman, R. (1990), Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41, 391–407. [342]

Donoho, D., and Grimes, C. (2003), "Hessian Eigenmaps: New Locally Linear Embedding Techniques for High-Dimensional Data," *Proceedings of the National Academy of Sciences*, 100, 5591–5596. [337]

Elhamifar, E., and Vidal, R. (2009), "Sparse Subspace Clustering," *IEEE Conference on Computer Vision and Pattern Recognition*, 2790–2797. [338,345]

Efron, B. (1978), "The Geometry of Exponential Families," *Annals of Statistics*, 6, 362–376. [337]

Fisher, R. A. (1953), "Dispersion on a Sphere," *Proceedings of the Royal Society of London*, Series A, 217, 295–305. [339]

Dan Geiger, D. H., King, H., and Meek, C. (2001), "Stratified Exponential Families: Graphical Models and Model Selection," *Annals of Statistics*, 29, 505–529. [337]

Giné, E., and Koltchinskii, V. (2006), "Empirical Graph Laplacian Approximation of Laplace-Beltrami Operators: Large Sample Results," in *High-Dimensional Probability*, Vol. 51 of *IMS Lecture Notes Monogr. Ser.*, 238–259. Beachwood, OH: Inst. Math. Statist. [337]

Golub, G., and Van Loan, C. (2013), *Matrix Computations*, (4th ed.), Baltimore, MD: John Hopkins University Press. [340]

Goresky, M., and MacPherson, R. (1988), *Stratified Morse Theory*, Springer-Verlag. [337]

Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732. [337]

Hamm, J., and Lee, D. D. (2005), "Grassmann Discriminant Analysis: A Unifying View on Subspace-Based Learning," *Advances in NIPS*, 17. [340]

Hansen, T. F., and Houle, D. (2008), "Measuring and Comparing Evolvability and Constraint in Multivariate Characters," *Journal of Evolutionary Biology*, 21,1201–1219. [337,338]

Haro, G., Randall, G., and Sapiro, G. (2007), "Stratification Learning: Detecting Mixed Density and Dimensionality in High Dimensional Point Clouds," *Advances in Neural Information Processing Systems*, 19, 553–560. [337]

Hoff, P. D. (2009), "Simulation of the Matrix Bingham–von Mises–Fisher Distribution, With Applications to Multivariate and Relational Data," *The Journal of Computational and Graphical Statistics*, 18, 438–456. [338,343]

Hoffman, T. (1999), "Probabilistic Latent Semantic Indexing," *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50–57. [342]

Huang, K., Ma, Y., and Vidal, R. (1999), "Minimum Effective Dimension for Mixtures of Subspaces: A Robust GPCA Algorithm and Its Applications," *IEEE Conference on Computer Vision and Pattern Recognition,* Vol. II, pp. 631–638. [337]

Jiang, W., and Tanner, M. A. (2008), "Gibbs Posterior for Variable Selection in High-Dimensional Classification and Data Mining," *Annals of Statistics*, 36, 2025–2550. [341]

Kendall, D. G. (1984), "Shape Manifolds, Procrustean Metrics, and Complex Projective Spaces," *Bulletin of the London Mathematical Society*, 16, 81–121. [338]

Laaksone, J., and Oja, E. (1996), "Subspace Dimension Selection and Averaged Learning Subspace Method in Handwritten Digit Classification," in: *Proceedings of International Conference on Artificial Neural Networks*, pp. 227–232. [337]

Lande, R. (1979), "Quantitative Genetic-Analysis of Multivariate Evolution, Applied to Brain-Body Size Allometry," *Evolution*, 33, 402–416. [337,338]

Lerman, G., and Zhang, T. (2010), "Probabilistic Recovery of Multiple Subspaces in Point Clouds by Geometric l_p Minimization," *Annals of Statistics*, 39, 2686–2715. [337,347]

Lipor, J., Hong, D., Tan, Y. S., and Balzano, L. (2021), "Subspace Clustering Using Ensembles of K-Subspaces," arXiv: 1709.04744. [338,345]

Liu, G., Lin, Z., and Yu, Y. (2010), "Robust Subspace Segmentation by Low-Rank Representation," *International Conference on Machine Learning*, pp. 663–670. [338,345]

Lu, C., Min, H., Zhao, Z., Zhu, L., Huang, D., and Yan, S. (2012), "Robust and Efficient Subspace Segmentation Via Least Squares Regression," *European Conference on Computer Vision*, 7578, 347–360. [338,345]

Mangasarian, O. L., and Wolberg, W. H. (1990), "Cancer Diagnosis Via Linear Programming," *Siam News*, 23, 1 and 18. [346]

McCallum, A. K. (2002), MALLET: A Machine Learning for Language Toolkit. Available at: *http://mallet.cs.umass.edu*. [347]

Mukherjee, S., Zhou, D-X., and Wu, Q. (2010), "Learning Gradients and Feature Selection on Manifolds," *Bernoulli*, 16, 181–207. [337]

NSF 2010 Awards. Available at: *http://www.nsf.gov/awardsearch/download.jsp*. [344,347]

Page, G., Bhattacharya, A., and Dunson, D. B. (2013), "Classification Via Bayesian Nonparametric Learning of Affine Subspaces," *Journal of American Statistical Association*, 108, 187–201. [337,347]

Pimentel-Alarcón, D., Balzano, L., Marcia, R., Nowak, R., and Willett, R. (2017), "Mixture Regression as Subspace Clustering," *2017 International Conference on Sampling Theory and Applications (SampTA), Tallin,* pp. 456–459. [337]

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000), "Inference of Population Structure Using Multilocus Genotype Data," *Genetics*, 155, 945–959. [342]

Rao, C. R. (1945), "Information and Accuracy Obtainable in the Estimation of Statistical Parameters," *Bulletin Calcutta Mathematical Society*, 37, 81–91. [337]

Reisinger, J., Waters, A., Silverthorn, B., and Mooney, R. J. (2010), "Spherical Topic Models," *Proceedings of the 27th ICML*. [337,343,347]

Roweis, S., and Saul, L. (2000), "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science*, 290, 2323–2326. [337]

Schwartz, L. (1965), "On Bayes Procedures," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4, 10–26.

Siebert, J. P. (1987), "Vehicle Recognition Using Rule Based Methods," *Turing Institute Research Memorandum TIRM-87-018*, Glasgow, Scotland: Turing Institute. [346]

Soltanolkotabi, M., and Candés, E. (2012), "A Geometric Analysis of Subspace Clustering With Outliers," *Annals of Statistics*, 40, 2195–2238. [338]

Vidal, R., Ma, Y., and Sastry, S. (2005), "Generalized Principal Component Analysis (GPCA)," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27, 1945–1959. [337]

Vidal, R., and Favaro, P. (2014), "Low Rank Subspace Clustering (LRSC)," *Pattern Recognition Letters*, 43, 47–61. [338,345]

Wu, H. C., Luk, R. W. P., Wong, K. F., Kwok, K. L. (2008), "Interpreting TF-IDF Term Weights As Making Relevance Decisions," *ACM Transactions on Information Systems*, 26, 1–37. [347]

Zheng, L., and Tse, D. N. C. (2002), "Communication on the Grassmann Manifold: A Geometric Approach to the Noncoherent Multiple-Antena Channel," *IEEE Transactions on Information Theory*, 48, 359–383. [337]

Zheng, H. (2011), "Mixture of Subspace Learning With Adaptive Dimensionality: A Self-Organizing Approach," in: *Foundations of Intelligent Systems. Advances in Intelligent and Soft Computing*, eds. Y. Wang and T. Li, Vol. 122. Berlin: Springer. [337]

Zhong, S., and Ghosh, J. (2005), "Generative Model-Based Document Clustering: A Comparative Study," *Knowledge and Information Systems*, 8, 374–384. [343]