LEARNING SUBSPACES OF DIFFERENT DIMENSIONS

BRIAN ST. THOMAS, LIZHEN LIN, LEK-HENG LIM, AND SAYAN MUKHERJEE

Abstract. We introduce a Bayesian model for inferring mixtures of subspaces of different dimensions. 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. We also prove posterior consistency for the mixture of subspaces model. The utility of our approach is demonstrated with applications to real and simulated data.

1. Introduction

The problem of modeling manifolds has been of great interest in a variety of statistical problems including dimension reduction [7, 15, 42], characterizing the distributions of statistical models as points on a Riemannian manifold [2, 17, 40], and the extensive literature in statistics and machine learning on manifold learning [12, 22, 36]. A generalization of the manifold setting is to model unions and intersections of manifolds (of possibly different dimensions), formally called stratified spaces [8, 19, 24]. 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. Mixtures of linear subspaces have been suggested in applications such as tracking images [27, 45], quantitative analysis of evolution or artificial selection [26, 32], applications in communication and coding theory [3, 48], and is relevant for text modeling [41, 9].In this paper we provide a model for the simplest instance of inferring stratified spaces, estimating mixtures of linear subspaces of different dimensions.

The idea of dimension reduction via projections onto low-dimensional subspaces goes back at least to Adcock [1] and Edgworth [16], with methodological and foundational contributions by R. A. Fisher [18]; see [12] for an excellent review. It is very interesting that in 1922 Fisher suggested that the statistical setting where the number of variables is greater than the number of observations, $p \gg n$ could be addressed by reducing the dimension of p to very few p^* summaries of the data where $p^* < n$. The summaries in this setting were linear combinations of the variables. This idea of dimension reduction has been extensively used statistics ranging from classical methods such as principal components analysis (PCA) [30] to a variety of recent methods, some algorithmic and some likelihood based, that fall under the category of nonlinear dimension reduction and manifold learning [7, 12, 15, 22, 36, 42]. A challenging setting for both algorithmic and probabilistic models in this setting is where the data are being generated from multiple populations inducing a mixture distribution. It is particularly challenging when the mixtures are of different dimensions.

In many applications a useful model for the observed high-dimensional data assumes the data is concentrated around a lower-dimensional structure in the high-dimensional ambient space. In addition, it is often the case that the data is generated from multiple processes or populations each of which has low-dimensional structure. In general, the degrees of freedom or number of parameters of the processes capturing the different populations need not be equal. In this paper, we address this problem of modeling data arising from a mixture of manifolds of different dimensions for the restricted case where the manifolds are linear subspaces.

The most recent work that offers both estimators and provides guarantees on estimates for inferring mixtures of subspaces has been limited to *equidimensional* subspaces [33, 38]. A Bayesian procedure for mixtures of subspaces of equal dimensions was developed in Page et al. [38]. A penalized loss based procedure was introduced in Lerman and Zhang [33] to learn mixtures of K-flats. There are significant difficulties in extending either approach to subspaces of different dimensions. The key difficulty in extending either approach is addressing the singularity introduced in moving between subspaces of different dimensions

when one parameterizes a subspace as a point on the Grassmann manifold and uses the natural geodesic on this manifold. This difficulty appears in the Bayesian approach as requiring the posterior samples to come from models of different dimensions which will require methods such as reversible jump MCMC which may cause mixing problems. The difficulty is immediate in the penalized loss model as the loss is based on a distance to subspaces and if the dimensions of the subspaces vary the loss based procedure becomes very difficult. A related line of work appears in Hoff [28] where efficient methodology is developed to sample from a space of orthonormal matrices with fixed intrinsic dimension using the matrix Bingham–von Mises–Fisher distribution. One of the results in this paper is a procedure to sample from orthonormal matrices of varying intrinsic dimensions. From a geometric perspective a method was developed in [28] to simulate from a Stiefel manifold with fixed intrinsic dimension, in this paper we provide a methodology to simulate over Stiefel manifolds of varying intrinsic dimensions.

The key idea we develop in this paper 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 [11]. This embedding removes the discontinuity that occurs in moving between subspaces of different dimensions when one uses the natural metric for a Grassmann manifold. The other tool we make use of is a Gibbs posterior [31] which allows us to efficiently obtain posterior samples of the model parameters.

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 a frequentist analysis of the Bayesian procedure is given that proves posterior consistency of the procedure. In

Section 6 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 d-dimensional subspaces in $\mathbb{R}^m$ will be denoted $\mathrm{Gr}(d, m)$. 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, e.g., $\mathbf{U}$, to denote subspaces and the corresponding letter in normal typeface, e.g., 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 $X = (x_1, \dots, x_n)$ drawn independent and identically from a mixture of K subspaces where each observation x_i is measured in the ambient space $\mathbb{R}^m$. We assume that each population is concentrated near a linear subspace $\mathbf{U}_k$ which we represent with an orthonormal basis U_k, $\mathbf{U}_k = \mathrm{span}(U_k)$, $k = 1, \dots, K$.

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 is modeled by a multivariate normal with mean μ_k and covariance Σ_k and 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 and by construction $\theta_k \in V_k$. 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 and subspace can be combined and specified by either of the following parameterizations

$$x \sim \begin{cases} \mathcal{N}_m\left(U_k\mu_k + \theta_k, U_k\Sigma_k U_k^\mathsf{T} + \sigma_k^2 V_k V_k^\mathsf{T}\right) \\ \mathcal{N}_m\left(U_k\mu_k + \theta_k, U_k(\Sigma_k - \sigma_k^2 I_{d_k})U_k^\mathsf{T} + \sigma_k^2 I_m\right). \end{cases} \tag{2.1}$$

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

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

$$x \sim \sum_{k=1}^{K} w_k \mathcal{N}_m\left(U_k\mu_k + \theta_k, U_k(\Sigma_k - \sigma_k^2 I_{d_k})U_k^\mathsf{T} + \sigma_k^2 I_m\right), \tag{2.2}$$

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 z 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

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

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 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, we want to consider d_k as random. We can state the following joint prior on the triple (U_k, μ_k, d_k)

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

Given d_k we can specify $\mu_k \mid d_k$ as a multivariate normal of dimension d_k. Given d_k we can also specify a conjugate distribution for U_k as the von Mises–Fisher (MF) distribution

$$\mathrm{MF}(U_k \mid A) \propto \mathrm{etr}(A^{\mathsf{T}} U_k),$$

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 $\mathrm{V}(d_k, m)$. A prior on d_k would take values over $[0, \ldots, m]$ and for each value the conditional distributions $\pi(U_k \mid d_k)$ and $\pi(\mu_k \mid d_k)$ need to be specified. For μ_k a prior distribution of $\mathcal{N}_{d_k}(0, \lambda I)$ seems reasonable since we can assume the mean is zero and the entries independent for any d_k. Specifying the conditional distribution for $\pi(U_k \mid d_k)$ is not as clear. As d_k changes the dimension of the matrix A needs to change and one cannot simply add columns of zeroes 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. Note that we can not simply integrate out d_k or U_k as nuisance parameters since we have no prior specification.

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 $\mathrm{Gr}(d_k, m)$ and that we represent a subspace $\mathbf{U}_k \in \mathrm{Gr}(d_k, m)$ with an orthonormal matrix $U_k \in \mathrm{V}(d_k, m)$ from an equivalence class $\{U_k \in \mathrm{V}(k, m) : \mathrm{span}(U_k) = \mathbf{U}_k\}$. 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 $\mathrm{Gr}(d_k, m)$ into $\mathbb{S}^{(m-1)(m+2)/2}$, an appropriately chosen sphere[1] in $\mathbb{R}^{m(m+1)/2}$, as proposed in Conway et al. [11]. 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 2.1 (Conway–Hardin–Sloane 1996)**.** *The representation of a subspace* $\mathbf{U} \in \mathrm{Gr}(d, m)$ *by its projection matrix* $P_{\mathbf{U}}$ *gives an isometric embedding of* $\mathrm{Gr}(d, m)$ *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 proceeds in the following steps: (1) given a basis U_k compute the projection matrix $P_k = U_k^{\mathsf{T}} U_k$, (2) 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 k lie on the same sphere. The key observation by Conway et al. [11] 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 $\varphi\left(\frac{1}{2}I_m\right) = \mathrm{vech}\left(\frac{1}{2}I_m\right)$ where $\varphi(A)$ denotes the embedding of the projection matrix A and vech is the half-vectorization operation

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

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

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,

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

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 $\mathbf{U}_1, \mathbf{U}_2 \in \operatorname{Gr}(d, m)$,

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

where $\theta_1, \ldots, \theta_d$ are the principal angles between the subspaces. We illustrate the embedding for two projection matrices in Figure 2.1.

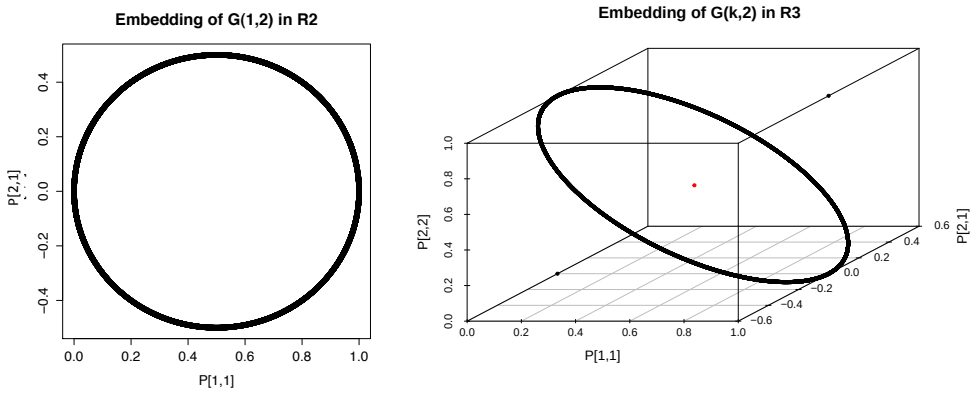


FIGURE 2.1. An illustration of the spherical embedding for subspaces from $\operatorname{Gr}(1, 2)$ into $\mathbb{R}^2$ on the left, and of $\operatorname{Gr}(k, 2)$, $k = 0, 1, 2$ on the right. Images of the embedding are in black, and the center of the sphere is in red. The coordinates for the embedding into $\mathbb{R}^2$ is the first column of the projection matrix. By including the last entry in the diagonal of the projection matrix, we obtain coordinates for the embedding into $\mathbb{R}^3$. $\operatorname{Gr}(0, 2)$ and $\operatorname{Gr}(2, 2)$ are trivial sets giving $\mathbf{0}_3$ and I_3 as their projection matrices. They act as poles on the sphere with coordinates $(0, 0, 0)$ and $(1, 0, 1)$.

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, these points will be 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 [11]. This numerical instability can lead to sub-optimal solutions.

Ease of computation: The projective distance is easy to compute via principal angles, which are in turn readily computable with singular value decomposition [23]. 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 [25]. In contrast the usual Plücker embedding requires a number of coordinates that is $\binom{m}{d}$, i.e., 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

imples 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, recall only points with integer trace have subspaces 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 2.2. *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*

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

can be found by the following procedure

(i) *Compute $Q = \varphi^{-1}(q)$.*

(ii) *Set the dimension of p to $d = \operatorname{tr}(Q)$.*

(iii) *Compute the eigendecomposition $Q = A\Lambda A^{-1}$.*

(iv) *Set B an $\ell \times d$ matrix equal to the columns of A corresponding to the top d eigenvalues.*

(v) *Let $p = \varphi(BB^{\mathsf{T}})$.*

Proof. In the case where the point $q \in \mathbb{S}^\ell$ is already on a cross-section of the sphere corresponding to $\operatorname{Gr}(d, \ell)$, the eigendecomposition will return exactly d non-zero eigenvalues. The eigenvectors give a basis for the subspace that is embedded into the point q. Similarly when the point q is between cross sections corresponding to Grassmannians, the above algorithm minimizes the Euclidean distance between the point p and q, and therefore minimizes the distance on $\mathbb{S}^\ell$. □

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

$$\begin{aligned} w &\sim \operatorname{Dir}_K(\alpha), \\ \delta_i &\sim \operatorname{Mult}(w), \\ P_k &\sim \mathcal{P}(\mathbb{S}^{(m-1)(m+2)/2}), \quad U_k U_k^{\mathsf{T}} = P_k, \quad d_k = \operatorname{tr}(P_k), \\ \mu_k \mid d_k &\sim \mathcal{N}_{d_k}(0, \lambda I), \end{aligned} \tag{2.3}$$

$$
\begin{aligned}
\theta_k \mid U_k &\sim \mathcal{N}_m(0, \phi I), \quad U_k^{\mathsf{T}}\theta_k = 0, \\
\sigma_k^{-2} &\sim \mathrm{Ga}(a, b), \\
\Sigma_{k(j)}^{-1} \mid d_k &\sim \mathrm{Ga}(c, d), \quad j = 1, \ldots, d_k, \\
x_i \mid \delta_i &\sim \sum\nolimits_{k=1}^{K} \delta_{ik} \mathcal{N}_m\Big(U_k\mu_k + \theta_k, U_k(\Sigma_k - \sigma_k^2 I_{d_k})U_k^{\mathsf{T}} + \sigma_k^2 I_m\Big),
\end{aligned}
\tag{2.4}
$$

where equation (2.3) 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 and computing the dimension d_k as the trace of the subspace and computing the subspace U_k from the projection and equation (2.4) 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 [31] 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 efficiently sampling. We can however follow the idea of a Gibbs posterior to design an efficient sampler. The idea behind a Gibbs posterior is to replace the standard posterior which takes the form of

$$\textit{posterior} \propto \textit{prior} \times \textit{likelihood}$$

with a distribution based on a loss or risk function that depends on both the data as well the parameter of interest in our case the loss function is given by

$$L(P_{[1:K]}, \theta_{[1:K]}, X) = \frac{1}{n}\sum\nolimits_{i=1}^{n}\Big[\min_{k=1,\dots,K}\Big(\|P_k(x_i-\theta_k)-(x_i-\theta_k)\|^2+\operatorname{tr}(P_k)\Big)\Big], \tag{3.1}$$
$$= \frac{1}{n}\sum\nolimits_{i=1}^{n}\Big[\min_{k=1,\dots,K}(e_{ik}+d_k)\Big],$$

where e_{ik} is the residual error for the ith sample given by the k-th 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

$$g(P_{[1:K]}, \theta_{[1:K]} \mid X) \propto \exp\big(-n\psi L(P_{[1:K]}, \theta_{[1:K]}, X)\big)\pi(P_{[1:K]})\pi(\theta_{[1:K]}), \tag{3.2}$$

where ψ is a chosen temperature parameter. A Gibbs posterior is simply a loss oriented alternative to the likelihood based posterior distribution. Traditionally it is 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 from the model parameters $U_{[1:K]}$ and $\theta_{[1:K]}$ using a Metropolis–Hastings algorithm which is effectively a random walk on the sphere. We first state a few facts that we will use. First recall that there is a deterministic relation between U_k and P_k, so given a P_k we can compute U_k. Also recall that a point sampled from $\mathbb{S}^m$ is not the pre-image of a subspace. Given a point $s_k^0 \in \mathbb{S}^m$ we denote the subspace corresponding to this point as $P_k = \varphi^{-1}(s_k^0)$, this 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 2.2. We obtain U_k correspond 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 ℓ-th 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 $s^0_{\sigma(i)} = (\sqrt{m(m+1)/8}) z^0_{\sigma(i)} / \|z^0_{\sigma(i)}\| + \varphi(I_m)$;
   (c) compute $P^0_{\sigma(i)} = \varphi^{-1}(s^0_{\sigma(i)})$;
   (d) compute $d^0_{\sigma(i)} = \mathrm{tr}(P^0_{\sigma(i)})$;
   (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 (e) and (f) 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 ℓ-th 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 $s_{\sigma(i)} = (\sqrt{m(m+1)/8}) z_{\sigma(i)} / \|z_{\sigma(i)}\| + \varphi(I_m)$;
   (c) compute $P_{\sigma(i)} = \varphi^{-1}(s_{\sigma(i)})$;
   (d) compute $d_{\sigma(i)} = \mathrm{tr}(P_{\sigma(i)})$;
   (e) compute $U_{\sigma(i)}$ as the top $d_{\sigma(i)}$ eigenvectors of $P_{\sigma(i)}$;
   (f) draw $u \sim \mathrm{Unif}[0, 1]$;

(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

$$\alpha = \frac{\exp\big(-n\psi L(P_{[1:K]}, \theta_{[1:K]}, X)\big)}{\exp\big(-n\psi L(P^{(\ell-1)}_{[1:K]}, \theta^{(\ell-1)}_{[1:K]}, X)\big)};$$

(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 $\beta_{\sigma(i)} \sim \mathcal{N}_m(\beta^{(\ell-1)}_{\sigma(i)}, I_m)$;

(l) compute $\theta_{\sigma(i)} = (I_m - P^{\ell-1}_{\sigma(i)})\beta_{\sigma(i)}$;

(m) draw $u \sim \text{Unif}[0, 1]$;

(n) set

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

(o) compute the acceptance probability

$$\alpha = \frac{\exp\big(-n\psi L(P^{(\ell-1)}_{[1:K]}, \theta_{[1:K]}, X)\big)}{\exp\big(-n\psi L(P^{(\ell-1)}_{[1:K]}, \theta^{(\ell-1)}_{[1:K]}, X)\big)};$$

(p) set

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

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 i-th 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 follow procedure provides posterior samples:

1. Draw $U_{[1:K]}^{(0)}, \theta_{[1:K]}^{(0)}, d_{[1:K]}^{(0)} \sim$ **Initialize**$(U_{[1:K]}, \theta_{[1:K]})$.
2. Draw $J_{k(j_k)} \sim \mathrm{Ga}(a, b)$ 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

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

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

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

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

$$a_k^* = n_k(m - d_k) + a,$$

$$b_k^* = b + \frac{n_k}{2}(\theta_k^{(t-1)})^{\mathsf{T}}\theta_k^{(t-1)} + \sum\nolimits_{\delta_i = k}\Big(\frac{1}{2}x_i^{\mathsf{T}}x_i - x_i^{\mathsf{T}}U_k^{(t-1)}U_k^{(t-1)\mathsf{T}}x_i\Big) - \theta_k^{(t-1)\mathsf{T}}\sum\nolimits_{\delta_i = k} x_i;$$

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

$$J_{k(j_k)}^{(t)} \sim \mathrm{Ga}\Big(\frac{n_k}{2} + a, b + \frac{1}{2}\sum\nolimits_{\delta_i = k} \big(U_k^{(t-1)\mathsf{T}} x_i - \mu_k\big)_{j_k}^2\Big),$$

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

(g) draw

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

The update steps for μ, σ^2, Σ are (d), (e), (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 effective assigns membership of each x_i to the subspace with least residual error.

When drawing from the Gibb's posterior distribution via a Metropolis–Hastings algorithm, the proposal distribution and temperature are adjusted through a burn-in period. In the first stage of burn-in, the proposal variance parameter $\tau = 1$ is fixed, while temperature is selected by a decreasing line search on a log-scale grid, from 10^{-20} to 10^{20} until the acceptance ratio reaches the 10%–90% range. With temperature fixed, the proposal variance τ is adjusted until the acceptance ratio falls in the 25%–45% range during the burn-in period. Thinning was applied in that every third draw of the sampler was kept, this was determined from autocorrelation analysis.

4. Specification of a Topic Model

4.1. **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 is ignored and a document is just a vector of counts of words [9, 14, 29]. 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 [9, 39].

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.

We first state the standard LDA model. Given D documents, k topics, and a vocabulary of size V the counts of the i-th word in the d-th document is specified by the following hierarchical model

$$\begin{aligned}
\theta_d \mid \alpha &\sim \mathrm{Dir}_K(\alpha) && \text{(topic probabilities for each document)},\\
\phi_k \mid \beta &\sim \mathrm{Dir}_V(\beta) && \text{(word probabilities for each topic)},\\
z_{i,d} \mid \theta_d &\sim \mathrm{Mult}_K(\theta_d) && \text{(topic assignment for each word in each document)},\\
w_{i,d} \mid \phi_{z_{i,d}} &\sim \mathrm{Mult}(\phi_{z_{i,d}}) && \text{(word counts for each word in each document)}.
\end{aligned}$$

In a spherical admixture model (SAM) [41] the vector of word counts in each document transformed by centering at zero and normalizing to unit length. The idea behind a SAM is to represent data as direction distributions on a hypersphere. The advantage of a SAM is that one can simultaneously model both frequency as well as presence/absence of words, an LDA model can only model frequency. There is empirical evidence of greater accuracy in using a SAM for sparse data such as text [5, 49]. 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. Strategies to eliminate this problem include removing what are called stop words from the corpus, for example words including "such", "as", and "and." However, it is not always the case that stop words for a particular corpus are known a priori. For example the word "topic" should be a stop word in a corpus of papers on topic modeling. Introducing an orthogonality constraint on the topics can enforce the prior knowledge that they should be interpretable as distinct. In [28] a Bayesian model for an orthogonal SAM is specified and a posterior sampling procedure is developed. A key insight 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, $\mathcal{V}(V, K)$.

Our second extension is to infer the number of topics K. While the orthogonality constraints help with interpretation of topics and removes redundant topics there are still topics with low posterior mixture probabilities and low coherence can still occur. This is mainly driven by misspecification of the number of topics. We now state our novel SAM model that enforces orthonormal columns as well as allows for the inference of the number of topics. The novel contributions of our model are inference of the number of topics by using the geometry of the Conway embedding to place a joint prior over the number of topics and topic. We are able to sample a from Stiefel manifolds of variable intrinsic dimension K by coupling draws from the von Mises–Fisher distribution with inversion of 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, this orthogonal matrix is 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 $\mathbb{S}^{(V-1)(V+2)/2}$ this 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, we wanted to avoid iterative procedures. 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

$$
\begin{aligned}
\mu \mid \kappa_0 &\sim \text{vMF}_{\mathbb{S}}(m, \kappa_0) && \text{(corpus average)},\\
\eta \mid \mu, \xi &\sim \text{vMF}_{\mathbb{S}}(\mu, \xi) && \text{(embedded orthogonal topics)},\\
(\phi, K) &= \varphi^{-1}(\eta) && \text{(orthogonal topics and number of topics)},\\
\theta_d \mid \alpha &\sim \text{Dir}_K(\alpha) && \text{(topic proportions for each document)},\\
\bar{\phi}_d &= \text{avg}(\phi, \theta_d) && \text{(spherical average, admixed parameter)},\\
v_d \mid \bar{\phi}_d, \kappa &\sim \text{vMF}_V\left(\bar{\phi}_d, \tau\right) && \text{(generates a document vector)}.
\end{aligned}
\tag{4.1}
$$

The main difference in the above model and prior models on spheres [28, 41] is that instead of sampling topics from the embedding on the Conway sphere a fixed K topic vectors were each sampled from a von Mises–Fisher distribution over $\mathbb{S}^V$. In [41] the vectors were not constrained to be orthogonal, in [28] 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

$$
L\big(\phi, K \mid \{v_d\}_{d=1}^D, \tau\big) = \frac{2DK}{V} - \sum_{d=1}^{D} \big(\tau \bar{\phi}_d^T v_d\big)
\tag{4.2}
$$

and corresponding Gibbs posterior

$$
g(\phi, K \mid \{v_d\}_{d=1}^D, \tau) \propto \exp\Big(- \psi DL \times \big(\phi, K \mid \{v_d\}_{d=1}^D, \tau\big)\Big)\pi(\phi)
\tag{4.3}
$$

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 $\log(\psi) \in [-10, 10]$.

Inference of the remaining parameters of Model (4.1) 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. Posterior consistency

In this section we show that the mixture of subspaces model specified in Section 2 has good frequentist properties. We provide an asymptotic analysis of our model and state some theoretical guarantees. We will prove posterior consistency—the posterior distribution specified by our model contracts to any arbitrary neighborhood of the true data generating density. Adapting proof techniques from the extensive literature on posterior consistency of Bayesian models [21] to our model is non-trivial.

Before we state our results we provide an explanation of the relation between our consistency result, the estimation procedure, and ideally what proof statements would be of interest. The MCMC algorithm stated in Section 3 uses a Gibbs posterior while our analysis in this section is for a standard posterior. Providing an argument for posterior consistency and proper calibration of the Gibbs posterior is of great interest but beyond the scope of this paper. A natural question is the asymptotic analysis of the convergence of mixture components and weights, basically an analysis of the clustering performance. The inference of mixture components is significantly more complicated that convergence in density and involves subtle identifiability issues.

Let $\mathcal{M}$ be the space of all the densities in $\mathbb{R}^m$ and f_0 be the true data generating density. We first define some notion of distances and neighborhoods in $\mathcal{M}$. A weak neighborhood of

f_0 with radius ϵ is defined as

$$W_\epsilon(f_0) = \left\{ f : \left| \int gf\, dx - \int gf_0\, dx \right| \le \epsilon \text{ for all } g \in C_b(\mathbb{R}^m) \right\}, \tag{5.1}$$

where $C_b(\mathbb{R}^m)$ is the space of all continuous and bounded functions on $\mathbb{R}^m$. The Hellinger distance $d_H(f, f_0)$ is defined as

$$d_H(f, f_0) = \left(\frac{1}{2} \int \left[\sqrt{f(x)} - \sqrt{f_0(x)} \right]^2 dx \right)^{1/2}.$$

Denote $U_\epsilon(f_0)$ an ϵ-Hellinger neighborhood around f_0 with respect to d_H. The Kullback–Leibler (KL) divergence between f_0 and f is defined to be

$$d_{\mathrm{KL}}(f_0, f) = \int f_0(x) \log \frac{f_0(x)}{f(x)}\, dx, \tag{5.2}$$

with $K_\epsilon(f_0)$ denoting an ϵ-KL neighborhood of f_0.

One of the key geometric insights of the Conway embedding given in Section 2 is that the geometric embedding allows for the specification of prior distributions of the parameters on a smooth space. Let Π_s be a prior on the sphere $\mathbb{S}^{(m-1)(m+2)/2}$ which can be taken to be the uniform distribution or the von Mises–Fisher distribution. By projecting the samples from Π_s onto the cross-sections of the sphere, Π_s induces a prior distribution on the subspaces basis $\mathbf{U}$ which we denote by $\Pi_{\mathbf{U}}$.

Note that our model induces a prior Π on $\mathcal{M}$. Assume the true density f_0 follows the following the regularity conditions

(i) $f_0(x)$ is bounded away from zero and bounded above by some constant M for all $x \in \mathbb{R}^m$;
(ii) $|\int \log(f_0(x)) f_0(x)\, dx| < \infty$;
(iii) for some $\delta > 0$, $\int [\log(f_0(x)/f_\delta(x))] f_0(x)\, dx < \infty$, where $f_\delta(x) = \inf_{y:|y-x|<\delta} f_0(y)$;
(iv) there exists $\alpha > 0$ such that $\int |x|^{2(1+\alpha)m} f_0(x)\, dx < \infty$.

We will show that the posterior distribution $\Pi(\,\cdot \mid x_1, \ldots, x_n)$ concentrates around any true density f_0 as $n \to \infty$. The following theorem is on weak consistency.

Theorem 5.1. *The posterior distribution* $\Pi(\cdot \mid x_1, \dots, x_n)$ *is weakly consistent. That is, for all* $\epsilon > 0$,

$$\Pi(W_\epsilon(f_0) \mid x_1, \dots, x_n) \to 1 \quad Pf_0^\infty\text{-almost surely as } n \to \infty, \tag{5.3}$$

where $W_\epsilon(f_0)$ *is a weak neighborhood of* f_0 *with radius* ϵ *and* Pf_0^∞ *represents the true probability measure for* $(x_1, x_2, \dots)$.

Proof. A result due to [43] states that if Π assumes positive mass to any Kullback–Leibler neighborhood of f_0, then the resulting posterior is weakly consistent. Therefore, one needs to show for all $\epsilon > 0$,

$$\Pi(\{f : d_{\mathrm{KL}}(f_0, f) \le \epsilon\}) > 0. \tag{5.4}$$

Note that $\Pi(d_i = m) > 0$ for any $i = 1, \dots, K$. Then one has

$$\Pi(K_\epsilon(f_0)) \ge \tag{5.5}$$
$$\int_{\mathrm{O}(m)\times\cdots\times\mathrm{O}(m)} \Pi(K_\epsilon(f_0) \mid U_1, \dots, U_K,\ d_1 = \cdots = d_k = m)\, d\Pi(U_1, \dots, U_K \mid d_1 = \cdots = d_k = m).$$

Therefore it suffices to show that $\Pi(K_\epsilon(f_0) \mid U_1, \dots, U_K) > 0$ where $U_1, \dots, U_K$ are the bases of the respective m-dimensional subspaces $\mathbf{U}_1, \dots, \mathbf{U}_K$.

We will show that there exists K large enough such that given m-dimensional subspaces $\mathbf{U}_1, \dots, \mathbf{U}_K$, the following mixture model assigns positive mass to any KL neighborhood of f_0,

$$f(x, \mathbf{U}, \Sigma) = \sum\nolimits_{j=1}^{K} w_j \mathcal{N}(\phi(\mu_j), \widetilde{\Sigma}_j), \tag{5.6}$$

with $\phi(\mu_j) = U_j\mu_j + \theta_j$ and $\widetilde{\Sigma}_j = U_j(\Sigma_j - \sigma^2 I_{d_j})U_j^{\mathsf{T}} + \sigma_j^2 I_m$.

If $\mathbf{U}_1, \dots, \mathbf{U}_K$ have the same dimension m and $U_1, \dots, U_K$ are a choice of orthonormal bases on the respective subspaces, then an infinite-dimensional version of our model can be

given by

$$X \sim g(x,\Gamma) = \int_{\mathbb{R}^m} \mathcal{N}(x;\phi(\mu),\Sigma)P(d\mu), \phi(\mu) = U\mu + \theta, \quad \Sigma = U(\Sigma_0 - \sigma^2 I_m)U^\mathsf{T} + \sigma^2 I_m$$

with parameters $\Gamma = (U, \theta, \Sigma_0, \sigma, P)$. The prior for P can be given by a Dirichlet process prior whose base measure has full support in $\mathbb{R}^m$ while the priors of the rest parameters $(U, \theta, \Sigma_0, \sigma)$ can be given the same as those of our model. By Theorem 3.1 of [38] or Theorem 2 of [47], there exists an open subset $\mathcal{P}$ of the space of all the probability measures on $\mathbb{R}^m$ such that for for all $\epsilon > 0$, any $P \in \mathcal{P}$ such that

$$\int_{\mathbb{R}^m} f_0(x) \log \frac{f_0(x)}{g(x,\Gamma)}\, dx \leq \epsilon. \tag{5.7}$$

We will first show that for any $\epsilon' > 0$ there exists K large enough and $w_1, \ldots, w_K$, $\theta_1, \ldots, \theta_K$, $\mu_1, \ldots, \mu_K$, $U_1, \ldots, U_K$, $\Sigma_1, \ldots, \Sigma_K$, $\sigma_1, \ldots, \sigma_K$, and σ^2 such that

$$\left| g(x,\Gamma) - \sum\nolimits_{j=1}^{K} w_j \mathcal{N}(\phi(\mu_j), \widetilde{\Sigma}_j) \right| \leq \epsilon',$$

for any $P \in \mathcal{P}$.

Let L be some large enough number. We first partition $\mathbb{R}^m$ into $L^m + 1$ sets. Let

$$A_{i_1,\ldots,i_m} = \prod_{j=1}^{m} \left(-L + (i_j - 1)\frac{\log L}{L}, -L + i_j \frac{\log L}{L} \right]$$

and

$$A_0^L \cup \left(\bigcup\nolimits_{i_1,\ldots,i_m=1}^{L} A_{i_1,\ldots,i_m} \right) = \mathbb{R}^m.$$

Pick $\mu_j \in A_{i_1,\ldots,i_m}$ where $j = 1, \ldots, L^m$ and let $w_j = P(A_{i_1,\ldots,i_m})$. Approximating the integral by the finite sum over the cubes in $\mathbb{R}^m$, for all $\epsilon'' > 0$, there exists L large enough such that

$$\left| g(x,\Gamma) - \sum\nolimits_{j=1}^{L^m} w_j \mathcal{N}(\mu_j U + \theta_j, \Sigma) \right| \leq \epsilon''.$$

Let $K = L^m$. Now let $U_1, \ldots, U_K$ be K points in the ϵ'' neighborhood of U, $\theta_1, \ldots, \theta_K$ in the ϵ'' neighborhood of θ, $\Sigma_1, \ldots, \Sigma_K$ be in the ϵ'' neighborhood of Σ_0, and $\sigma_1, \ldots, \sigma_K$ in the ϵ''

neighborhood of σ, then by the continuity of $\sum_{j=1}^K w_j \mathcal{N}(\mu_j U + \theta, \Sigma)$, one can show that

$$\begin{aligned}\left| g(x,\Gamma) - \sum_{j=1}^K w_j \mathcal{N}(\mu_j U_j + \theta_j, \Sigma_j) \right| &\leq \left| g(x,\Gamma) - \sum_{j=1}^K w_j \mathcal{N}(\mu_j U + \theta, \Sigma) \right| \\ &\quad + \left| \sum_{j=1}^K w_j \mathcal{N}(\mu_j U + \theta, \Sigma) - \sum_{j=1}^K w_j \mathcal{N}(\mu_j U_j + \theta_j, \widetilde{\Sigma}_j) \right| \\ &\leq \epsilon'' + \delta(\epsilon''),\end{aligned}$$

where $\delta(\epsilon'') \to 0$ when $\epsilon'' \to 0$. For the above choices of $w_1, \ldots, w_K$, $\mu_1, \ldots, \mu_K$, $\theta_1, \ldots, \theta_K$, $U_1, \ldots, U_K$, $\Sigma_1, \ldots, \Sigma_K$, $\sigma_1, \ldots, \sigma_K$ corresponding to any $P \in \mathcal{P}$, one looks at

$$\begin{aligned}\int_{\mathbb{R}^m} f_0(x) \log \frac{f_0(x)}{\sum_{j=1}^K w_j \mathcal{N}(\mu_j U_j + \theta_j, \widetilde{\Sigma}_j)} \, dx &= \int_{\mathbb{R}^m} f_0(x) \log \frac{f_0(x)}{g(x,\Gamma)} \, dx \\ &\quad + \int_{\mathbb{R}^m} f_0(x) \log \frac{g(x,\Gamma)}{\sum_{j=1}^K w_j \mathcal{N}(\mu_j U_j + \theta_j, \widetilde{\Sigma}_j)} \, dx.\end{aligned}$$

Take $\epsilon'' = \epsilon$. By the continuity of the log function, one has

$$\int_{\mathbb{R}^m} f_0(x) \log \frac{f_0(x)}{\sum_{j=1}^K w_j \mathcal{N}(\mu_j U_j + \theta_j, \widetilde{\Sigma}_j)} \, dx \leq \epsilon + \delta'(\epsilon) \tag{5.8}$$

where $\delta'(\epsilon) \to 0$ when $\epsilon \to 0$. Note that our prior assigns positive mass to arbitrary neighborhood of $w_1, \ldots, w_K$, $\theta_1, \ldots, \theta_K$, $\mu_1, \ldots, \mu_K$, $U_1, \ldots, U_K$, $\Sigma_1, \ldots, \Sigma_K$, $\sigma_1, \ldots, \sigma_K$, σ^2. By (5.7), (5.8) and the continuity of the model, our assertion follows. □

In proving the following strong consistency theorem, we assume that the parameters σ_i^2, σ^2 and the diagonal elements of Σ_i $(i = 1, \ldots, K)$ follow i.i.d. truncated Gamma priors supported on some bounded interval $[0, M]$ for some large enough constant M.

Theorem 5.2. *The posterior distribution $\Pi(\cdot \mid x_1, \ldots, x_n)$ is strongly consistent. That is, for all $\epsilon > 0$,*

$$\Pi\left(U_\epsilon(f_0) \mid x_1, \ldots, x_n\right) \to 1 \quad P f_0^\infty - \textit{almost surely as } n \to \infty, \tag{5.9}$$

where $U_\epsilon(f_0)$ is a neighborhood of f_0 with radius ϵ with respect to the Hellinger distance.

Proof. By Theorem 5.1, the true density f_0 is in the weak support of our model. Then by a result due to [6] (also see Theorem 2 in [20]), for all $\epsilon > 0$ if we can construct sieves $D_{\delta,n} \subseteq \mathcal{M}$ with $\delta < \epsilon$ such that the metric entropy $\log N(\delta, D_{\delta,n}) \leq n\beta$ for some $\beta < \epsilon^2/2$ and $\Pi(D_{\epsilon,n}^c) \leq C_1 \exp(-n\beta_1)$ with some constants C_1 and β_1. Then the posterior distribution $\Pi(\cdot \mid x_1, \ldots, x_n)$ is strongly consistent at f_0.

Denote

$$\Theta = (K, \theta_1, \ldots, \theta_K, w_1, \ldots, w_K, \mu_1, \ldots, \mu_K, U_1, \ldots, U_K, \Sigma_1, \ldots, \Sigma_K, \sigma_1, \ldots, \sigma_K, \boldsymbol{\sigma})$$

and

$$\boldsymbol{\sigma} = (\sigma_{11}, \ldots, \sigma_{1d_1}, \ldots, \sigma_{K1}, \ldots, \sigma_{Kd_K}, \sigma, \sigma_1, \ldots, \sigma_K)$$

where σ_{ij}^2 is the jth diagonal element of Σ_i which we assume to be diagonal in our model. Let

$$\boldsymbol{\Theta}_n = \{\Theta : K = K_n,\ \|\theta_i\| \leq \bar{\theta}_n,\ \|\mu_i\| \leq \bar{\mu}_n,\ \max(\boldsymbol{\sigma}) \leq M,\ \min(\boldsymbol{\sigma}) \geq h_n,\ i = 1, \ldots, K\}$$

where K_n, $\bar{\theta}_n$, $\bar{\mu}_n$, M and h_n are some sequences depending on n.

Let $D_{\delta,n} = \{f(x, \Theta) : \Theta \in \boldsymbol{\Theta}_n\}$ where $f(x, \Theta)$ is given by (5.6) for any Θ. We need to verify the metric entropy and the prior mass of $D_{\delta,n}$. Let

$$\Theta_{K_n}^i = (d_1^i, \ldots, d_K^i, \theta_1^i, \ldots, \theta_K^i, w_1^i, \ldots, w_K^i, \mu_1^i, \ldots, \mu_K^i, U_1^i, \ldots, U_K^i, \boldsymbol{\sigma}^i), \quad i = 1, 2.$$

Note that posterior consistency with respect to the Hellinger distance is equivalent to posterior consistency with respect to the L^1-distance due to the equivalence of the two distances. For $i = 1, 2$, let $\phi^i(\mu_j) = U_j^i \mu_j^i + \theta_j^i$ and $\widetilde{\Sigma}_j^i = U_j^i(\Sigma_j^i - (\sigma^i)^2 I_{d_j^i})(U_j^i)^{\mathsf{T}} + (\sigma_j^i)^2 I_m$. One has

$$\begin{aligned}
&\int_{\mathbb{R}^m} |f(x, \Theta_{K_n}^1) - f(x, \Theta_{K_n}^2)|\, dx \\
&\quad = \int_{\mathbb{R}^m} \Big| \sum\nolimits_{j=1}^{K_n} w_j^1 \mathcal{N}_m(\phi^1(\mu_j), \widetilde{\Sigma}_j^1) - \sum\nolimits_{j=1}^{K_n} w_j^2 \mathcal{N}_m(\phi^2(\mu_j), \widetilde{\Sigma}_j^2) \Big|\, dx \\
&\quad = \int_{\mathbb{R}^m} \Big| \sum\nolimits_{j=1}^{K_n} w_j^1 \mathcal{N}_m(\phi^1(\mu_j), \widetilde{\Sigma}_j^1) - \sum\nolimits_{j=1}^{K_n} w_j^2 \mathcal{N}_m(\phi^2(\mu_j), \widetilde{\Sigma}_j^2)
\end{aligned}$$

$$
\begin{aligned}
&\quad + \sum_{j=1}^{K_n} w_j^1 \mathcal{N}_m(\phi^2(\mu_j), \widetilde{\Sigma}_j^2) - \sum_{j=1}^{K_n} w_j^1 \mathcal{N}_m(\phi^2(\mu_j), \widetilde{\Sigma}_j^2) \Big| \, dx \\
&\leq \int_{\mathbb{R}^m} \Big| \sum_{j=1}^{K_n} w_j^1 \left(\mathcal{N}_m(\phi^1(\mu_j), \widetilde{\Sigma}_j^1) - \mathcal{N}_m(\phi^2(\mu_j), \widetilde{\Sigma}_j^2) \right) \Big| \, dx \\
&\qquad + \int_{\mathbb{R}^m} \Big| \sum_{j=1}^{K_n} (w_j^1 - w_j^2) \mathcal{N}_m(\phi^2(\mu_j), \widetilde{\Sigma}_j^2) \Big| \, dx \\
&\leq \int_{\mathbb{R}^m} \Big| \sum_{j=1}^{K_n} w_j^1 \left(\mathcal{N}_m(\phi^1(\mu_j), \widetilde{\Sigma}_j^1) - \mathcal{N}_m(\phi^2(\mu_j), \widetilde{\Sigma}_j^2) \right) \Big| \, dx + \sum_{j=1}^{K_n} |w_j^1 - w_j^2| \\
&\leq \sum_{j=1}^{K_n} w_j^1 \int_{\mathbb{R}^m} \Big| \left(\mathcal{N}_m(\phi^1(\mu_j), \widetilde{\Sigma}_j^1) - \mathcal{N}_m(\phi^2(\mu_j), \widetilde{\Sigma}_j^2) \right) \Big| \, dx + \sum_{j=1}^{K_n} |w_j^1 - w_j^2|.
\end{aligned}
$$

Note that

$$
\begin{aligned}
&\int_{\mathbb{R}^m} \Big| \mathcal{N}_m(\phi^1(\mu_j), \widetilde{\Sigma}_j^1) - \mathcal{N}_m(\phi^2(\mu_j), \widetilde{\Sigma}_j^2) \Big| \, dx \\
&\quad = \int_{\mathbb{R}^m} \Big| \mathcal{N}_m(\phi^1(\mu_j), \widetilde{\Sigma}_j^1) - \mathcal{N}_m(\phi^2(\mu_j), \widetilde{\Sigma}_j^1) + \mathcal{N}_m(\phi^2(\mu_j), \widetilde{\Sigma}_j^1) - \mathcal{N}_m(\phi^2(\mu_j), \widetilde{\Sigma}_j^2) \Big| \, dx \\
&\quad \leq \int_{\mathbb{R}^m} \Big| \mathcal{N}_m(\phi^1(\mu_j), \widetilde{\Sigma}_j^1) - \mathcal{N}_m(\phi^2(\mu_j), \widetilde{\Sigma}_j^1) \Big| \, dx + \int_{\mathbb{R}^m} \Big| \mathcal{N}_m(\phi^2(\mu_j), \widetilde{\Sigma}_j^1) - \mathcal{N}_m(\phi^2(\mu_j), \widetilde{\Sigma}_j^2) \Big| \, dx.
\end{aligned}
$$

By the proof of Lemma 5 of [47], one has for the first term of the above expression

$$
\begin{aligned}
\int_{\mathbb{R}^m} \Big| \mathcal{N}_m(\phi^1(\mu_j), \widetilde{\Sigma}_j^1) - \mathcal{N}_m(\phi^2(\mu_j), \widetilde{\Sigma}_j^1) \Big| \, dx &\leq \sqrt{\frac{2}{\pi}} \frac{\|\phi^1(\mu_j) - \phi^2(\mu_j)\|}{\lambda_1(\widetilde{\Sigma}_j^1)^{1/2}} \\
&= \sqrt{\frac{2}{\pi}} \frac{\|(U_j^1 \mu_j^1 + \theta_j^1) - (U_j^2 \mu_j^2 + \theta_j^2)\|}{\lambda_1(\widetilde{\Sigma}_j^1)^{1/2}} \\
&\leq \sqrt{\frac{2}{\pi \lambda_1(\widetilde{\Sigma}_j^1)}} \left(\|\theta_j^1 - \theta_j^2\| + \|U_j^1 \mu_j^1 - U_j^2 \mu_j^2\| \right),
\end{aligned}
$$

where $\lambda_1(\widetilde{\Sigma}_j^1)$ is the smallest eigenvalue of $\widetilde{\Sigma}_j^1$. Therefore combining all the terms above

$$
\begin{aligned}
&\int_{\mathbb{R}^m} |f(x, \Theta_{K_n}^1) - f(x, \Theta_{K_n}^2)| \, dx \\
&\quad \leq \sum_{j=1}^{K_n} |w_j^1 - w_j^2| + \max_{j=1,\ldots,K_n} \left\{ \sqrt{\frac{2}{\pi \lambda_1(\widetilde{\Sigma}_j^1)}} \left(\|\theta_j^1 - \theta_j^2\| + \|U_j^1 \mu_j^1 - U_j^2 \mu_j^2\| \right) \right\}
\end{aligned}
$$

$$+\sum_{j=1}^{K_n} w_j \int_{\mathbb{R}^m} \left|\mathcal{N}_m(\phi^2(\mu_j), \widetilde{\Sigma}_j^1) - \mathcal{N}_m(\phi^2(\mu_j), \widetilde{\Sigma}_j^2)\right| \, dx.$$

Without loss of generality, assume $\det(\widetilde{\Sigma}_j^2) \geq \det(\widetilde{\Sigma}_j^1)$. One has

$$\begin{aligned}
&\int_{\mathbb{R}^m} \left|\mathcal{N}_m(\phi^2(\mu_j), \widetilde{\Sigma}_j^1) - \mathcal{N}_m(\phi^2(\mu_j), \widetilde{\Sigma}_j^2)\right| \, dx \\
&\quad = \frac{1}{(2\pi)^{m/2}} \int_{\mathbb{R}^m} \bigg| \frac{1}{\det(\widetilde{\Sigma}_j^1)^{1/2}} \exp\big[-\tfrac{1}{2}(x-\phi^2(\mu_j))^{\mathsf{T}}(\widetilde{\Sigma}_j^1)^{-1}(x-\phi^2(\mu_j))\big] \\
&\qquad\qquad - \frac{1}{\det(\widetilde{\Sigma}_j^2)^{1/2}} \exp\big[-\tfrac{1}{2}(x-\phi^2(\mu_j))^{\mathsf{T}}(\widetilde{\Sigma}_j^2)^{-1}(x-\phi^2(\mu_j))\big] \bigg| \, dx \\
&\quad \leq \frac{2^{m+1}}{(2\pi)^{m/2}} \int_{[0,\infty)^m} \max\bigg\{0, \ \frac{1}{\det(\widetilde{\Sigma}_j^1)^{1/2}} \exp\big[-\tfrac{1}{2}(x-\phi^2(\mu_j))^{\mathsf{T}}(\widetilde{\Sigma}_j^1)^{-1}(x-\phi^2(\mu_j))\big] \\
&\qquad\qquad - \frac{1}{\det(\widetilde{\Sigma}_j^2)^{1/2}} \exp\big[-\tfrac{1}{2}(x-\phi^2(\mu_j))^{\mathsf{T}}(\widetilde{\Sigma}_j^2)^{-1}(x-\phi^2(\mu_j))\big] \bigg\} \, dx.
\end{aligned} \tag{5.10}$$

We take $\delta = \epsilon/2$. We first partition $[0,1]^{K_n}$ into N_w grid points. With a choice of the grid points given in Lemma 1 in [20], the number grid points needed for $\sum_{j=1}^{K_n} |w_j^1 - w_j^2| \leq \epsilon/8$ is bounded by

$$\log N_w \leq K_n \left[1 + \log\left(\frac{1+\epsilon/8}{\epsilon/8}\right)\right].$$

Note that the number of balls of radius δR used to cover a Euclidean ball centered at the origin of radius R in $\mathbb{R}^m$ is bounded by $(3/\delta)^m$. Therefore, the number of balls of radius $\epsilon[2\pi\lambda_1(\widetilde{\Sigma}_1)]^{1/2}/32$ needed to cover $\|\theta\| \leq \bar{\theta}_n$ is bounded by $(96\bar{\theta}_n/\epsilon[2\pi\lambda_1(\widetilde{\Sigma}_1)]^{1/2})^m$. Then one can always find $\theta_1^i, \ldots, \theta_{K_n}^i$, $i = 1, 2$, such that

$$\max_{j=1,\ldots,K_n} \sqrt{\frac{2}{\pi\lambda_1(\widetilde{\Sigma}_j^1)}} \|\theta_j^1 - \theta_j^2\| \leq \frac{\epsilon}{8}.$$

Note that μ_j^1 and μ_j^2 do not necessarily have the same dimension. We view them as elements in $\mathbb{R}^m$ by filling the last $m - d_j^i$, $i = 1, 2$, elements with 0. Then one can show that $\|U_j^1\mu_j^1 - U_j^2\mu_j^2\| \leq \|\mu_j^1 - \mu_j^2\|$. Then the covering number for μ is bounded by $(96\bar{\mu}_n/\epsilon[2\pi\lambda_1(\widetilde{\Sigma}_1)]^{1/2})^m$

with which one can find $\mu_1^i, \dots, \mu_{K_n}^i$, $i = 1, 2$, such that

$$\max_{j=1,\dots,K_n} \sqrt{\frac{2}{\pi\lambda_1(\widetilde{\Sigma}_j^1)}} \|U_j^1 \mu_j^1 - U_j^2 \mu_j^2\| \leq \frac{\epsilon}{8}.$$

Note that for $\Theta \in \mathbf{\Theta}_n$, all the eigenvalues of $\widetilde{\Sigma}$ lie in the interval $[h_n^2, M^2]$. We divide the m-dimensional cube $[h_n^2, M^2]^m$ into smaller cubes such that there are $N_{\boldsymbol{\lambda}}$ grid points for the m eigenvalues $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m)$ of $\widetilde{\Sigma}$. Note that for any $\widetilde{\Sigma}_1$ whose eigenvalues fall into one of the cubes in the grid, one can always find $\widetilde{\Sigma}_2$ such that $\widetilde{\Sigma}_1^{-1} - \widetilde{\Sigma}_2^{-1}$ is positive definite. For example, one may take $\widetilde{\Sigma}_2^{-1} = \widetilde{\Sigma}_1^{-1} - \tilde{\epsilon} I_m$ where $\tilde{\epsilon}$ is small enough and $0 < \tilde{\epsilon} < \min\{1/\lambda_1(\widetilde{\Sigma}_1), \dots, 1/\lambda_m(\widetilde{\Sigma}_1)\}$. Then

$$\exp\big[-\tfrac{1}{2}(x - \phi^2(\mu_j))^{\mathsf{T}} (\widetilde{\Sigma}_j^1)^{-1} (x - \phi^2(\mu_j))\big] \leq \exp\big[-\tfrac{1}{2}(x - \phi^2(\mu_j))^{\mathsf{T}} (\widetilde{\Sigma}_j^2)^{-1} (x - \phi^2(\mu_j))\big].$$

Then from the above inequality and (5.10), one has

$$\begin{aligned}
&\int_{\mathbb{R}^m} \left| \mathcal{N}_m(\phi^2(\mu_j), \widetilde{\Sigma}_j^1) - \mathcal{N}_m(\phi^2(\mu_j), \widetilde{\Sigma}_j^2) \right| dx \\
&\quad \leq \frac{2^{m+1}}{(2\pi)^{m/2}} \int_{[0,\infty)^m} \left(\frac{1}{\det(\widetilde{\Sigma}_j^1)^{1/2}} - \frac{1}{\det(\widetilde{\Sigma}_j^2)^{1/2}} \right) \exp\big[-\tfrac{1}{2}(x - \phi^2(\mu_j))^{\mathsf{T}} (\widetilde{\Sigma}_j^2)^{-1} (x - \phi^2(\mu_j))\big] \, dx \\
&\quad \leq 2^m \frac{\det(\widetilde{\Sigma}_j^2)^{1/2} - \det(\widetilde{\Sigma}_j^1)^{1/2}}{\det(\widetilde{\Sigma}_j^1)^{1/2}}.
\end{aligned}$$

We divide the range of each of the eigenvalues $[h_n^2, M^2]$ into L equidistant intervals and let $\lambda_{jl} = h_n^2(1 + \epsilon/2^{m+3})^{2l_j/m}$ where $j = 1, \dots, m$ and $1 \leq l_j \leq L$. We pick L to be the smallest integer which satisfies $h_n^2(1 + \epsilon/2^{m+3})^{2l_j/m} \geq M^2$. We pick the jth eigenvalue of $\widetilde{\Sigma}_j^1$ and $\widetilde{\Sigma}_j^2$ to be in some interval $[\lambda_{j(l-1)}, \lambda_{jl}]$ and satisfying the ordering on the eigenvalues. Then one has

$$2^m \frac{\det(\Sigma_j^2)^{1/2} - \det(\Sigma_j^1)^{1/2}}{\det(\Sigma_j^1)^{1/2}} \leq 2^m \frac{\prod_{j=1}^m h_n^2(1 + \epsilon/2^{m+3})^{2l_j/m} - \prod_{j=1}^m h_n^2(1 + \epsilon/2^{m+3})^{2(l_j-1)/m}}{\prod_{j=1}^m h_n^2(1 + \epsilon/2^{m+3})^{(2l_j-1)/m}} \leq \frac{\epsilon}{8},$$

and the metric entropy of $\boldsymbol{\lambda}$ in $[h_n^2, M^2]^m$ is bounded by

$$\log N_{\boldsymbol{\lambda}} \leq m \left[\frac{\log(M^2/h_n^2)}{\log(1 + \epsilon/2^{m+3})} + 1 \right].$$

Letting $\bar{\theta}_n = \bar{\mu}_n$ and combining all the terms above on the entropy numbers, the metric entropy number of $D_{n,\epsilon}$ is bounded by

$$\log N(\delta, D_{\delta,n}) \leq K_n \left[1+\log\left(\frac{1+\epsilon/8}{\epsilon/8}\right)\right] + 2m \log\left[\frac{96\bar{\theta}_n}{\epsilon(2\pi\lambda_1(\widetilde{\Sigma}_1))^{1/2}}\right] + m\left[\frac{2\log(M/h_n)}{\log(1+\epsilon/2^{m+3})}+1\right].$$

Let $K_n = c_1\sqrt{n}$, $\bar{\theta}_n = c_2\sqrt{n}$, and $h_n = c_3 n^{-1/b}$ for some constants $b > 0$ and c_1, c_2 and c_3 small enough. Then one has

$$\log N(\delta, D_{\delta,n}) \leq n\beta.$$

It remains to verify the condition that the prior mass outside $D_{\delta,n}$ is exponentially small given our priors on the parameters and the above choice of K_n, $\bar{\theta}_n$, h_n and M. We assume multivariate normal priors (with diagonal covariance matrices) for θ and μ, thus with a choice of $\bar{r}_n$, $\bar{\mu}_n$, one can show that using changing of variables and Mill's inequality, the tail $P(\|\theta\| \geq \bar{\theta}_n)$ decays exponentially. The elements $\sigma_1^2, \ldots, \sigma_K^2, \sigma^2$ and the variance (i.e., diagonal) terms of Σ_j, $j = 1, \ldots, m$, are assumed to follow the i.i.d. Gamma priors with density $b^a/\Gamma(a)x^{-a-1}\exp(-b/x)$ and hyperparameters a and b. Then by direct calculation of the integrals,

$$P(\min(\boldsymbol{\sigma}) \leq h_n) \leq c_4 n \exp(-\tilde{C}_4 h_n^{-b}) \leq c_4 \exp(-\tilde{c}_4 n + \log n),$$

which decays exponentially fast. □

6. Results on real and simulated data

We illustrate the utility of the embedding using three data analysis examples. The first example 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. In the second example, 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 [4]. The last example compares the spherical topic

model we developed to a latent Dirichlet allocation model on a corpus of NSF award abstracts [37].

6.1. **Line 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. The models we compare are K-mans clustering (K-means), a mixture of normals (GMM), a mixture of non-parametric factor models (MFM) [10], our mixture of subspaces model with variable dimensions (MSM), and our mixture of subspaces model with dimension fixed to to $d = 2$ (MSM $d = 2.2$). For the subspace model we set the temperature parameter for the Gibbs posterior to 10^{-6}, and acceptance rates between 38% and 48% were achieved for the subspace and affine mean parameters.

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 situation to infer. To understand the effect of uncertainty of subspace measurements on accuracy of models we add a isotropic noise around the subspaces via a precision parameter.

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

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

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

We generated 500 observations from both the line and the plane, see Figure 6.1. For each of the five variance levels, ten data sets were generated, and a holdout set of 50 observations from the line and plane.

A comparison of clustering accuracy of the five models is summarized in Table 6.1. We report the range in clustering accuracy for eaach method on the holdout set over the ten runs. We conclude from Table 6.1 that: (1) K-means performs poorly, (2) as the precision parameter increases the performance of the GMM improves and starts to approach the MFM and MSM results, (3) the MSM with variable dimension outperforms the MSM with fixed dimension, and (4) the MSM and MFM results are very similar. Note that the MSM provides more geometric information inculding the dimension of the subspace.

TABLE 6.1. Range of cluster assignment accuracy for the simulated data for the five methods.

SYNTHETIC DATA

Precision	K-means	MSM	MSM $d = 2, 2$	GMM	MFM
0.1	$(0.66, 0.77)$	$(0.95, 0.99)$	$(0.92, 0.97)$	$(0.89, 0.97)$	$(0.95, 0.99)$
0.5	$(0.64, 0.82)$	$(0.90, 0.98)$	$(0.90, 0.96)$	$(0.87, 0.98)$	$(0.91, 0.99)$
1	$(0.57, 0.70)$	$(0.87, 0.98)$	$(0.86, 0.97)$	$(0.85, 0.98)$	$(0.88, 0.98)$
2	$(0.64, 0.72)$	$(0.87, 0.96)$	$(0.87, 0.95)$	$(0.87, 0.95)$	$(0.86, 0.98)$
5	$(0.59, 0.80)$	$(0.84, 0.97)$	$(0.84, 0.97)$	$(0.84, 0.97)$	$(0.83, 0.97)$

6.2. **Classification on UCI data.** To study the clustering performance on more realistic data we examined the classification accuracy on three data sets from the UCI Data Repository: the Statlog Vehicle Silhouettes data [44], the Wisconsin Breast Cancer data [34], and the Statlog Heart data [13]. 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 with

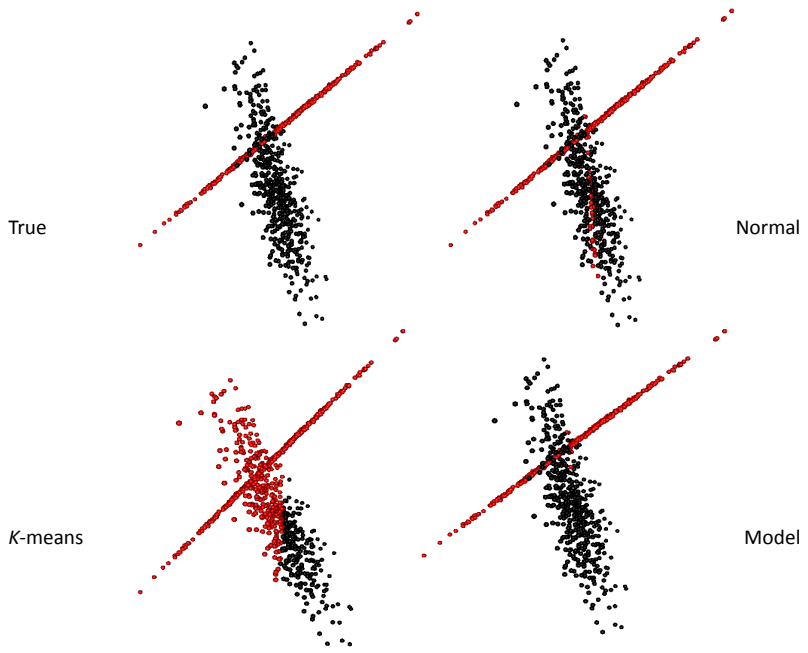


FIGURE 6.1. The upper-left panel displays the true cluster assignments, the upper-right panel displays the assignemnts given by the GMM, the lower-left panel displays the the K-means assignment, and the lower-right panel displays the assignmets by the MSM.

fixed dimensions (MSM $d = (5, 2, 2)$), and a mixture of non-parametric factor models (MFM) [10].

For the subspace models the temperature parameter of the Gibbs posterior was set to obtain an acceptance ration in in the range of 20-40% during the burn-in period. We did not use a cross-validation criteria 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 ten times and report the range in test errors in Table 6.2 for the results.

We conclude from Table 6.2 that: the MSM with variable dimension outperforms the MSM with fixed dimension, the mixture of subspaces and mixture of factors perform as well or better than the logit model which is the only supervised method, MSM and MFM have very comparable performance.

TABLE 6.2. Range of cluster assignment accuracy for the three data sets using five models on holdout data.

REAL DATA

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

Our subspace models allows for an estimate of the dimension of the linear subspace. This is not possible for either the Bayesian mixture models proposed in [38], the non-parametric mixture of factor models proposed in [10], or the penalized cost based mixture of subspaces model [33]. In Table 6.3 we state the posterior probabilities estimates of the dimension of the subspaces.

TABLE 6.3. Posterior probabilities for the dimension of the subspace of clusters in each data set.

BREAST

Class	Dim.	Post. Prob.
1	5	0.45
1	6	0.53
2	4	0.66
2	5	0.34

HEART

Class	Dim.	Post. Prob.
1	1	0.10
1	2	0.78
1	3	0.12
2	1	0.79
2	2	0.21

VEHICLE

Class	Dim.	Post. Prob.
1	1	0.69
1	2	0.31
2	1	0.76
2	2	0.24
3	1	0.78
3	2	0.22
4	1	0.93
4	2	0.07

6.3. **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 [37]. The vocabulary was constructed using the tokenizer from the Mallet package with bi-gram

extraction [35]. The vocabulary was then reduced to the to terms that were within the top 10% term frequencyinverse document frequency metric [46] 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 (4.1) to the a standard LDA model with 20 topics. We fixed the LDA model to have 20 topics. It was pointed out in [41] 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 6.2 and 6.3.

One thing to notice about the output of the spherical topic models is that when investigating the topics, the positive and the negative parts of the topic vectors tend to be thematically coherent (also noted by [41]). This is interesting because in a sense it allows a denser representation of the topics. One interesting result of the model specified in (4.1) on the NSF abstract data is 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 looking at the top terms of a topic.

Unlike standard admixture models, our spherical topic model allows for the inference on the number of topics. In Figure 6.4 we display the posterior distribution over the number of topics which is centered around 30.

7. 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 was using the observation that subspaces of different dimensions can be represented as points on a sphere is very useful for inference. The utility of this representation is that sampling from a sphere is straightforward, there exists a distance between

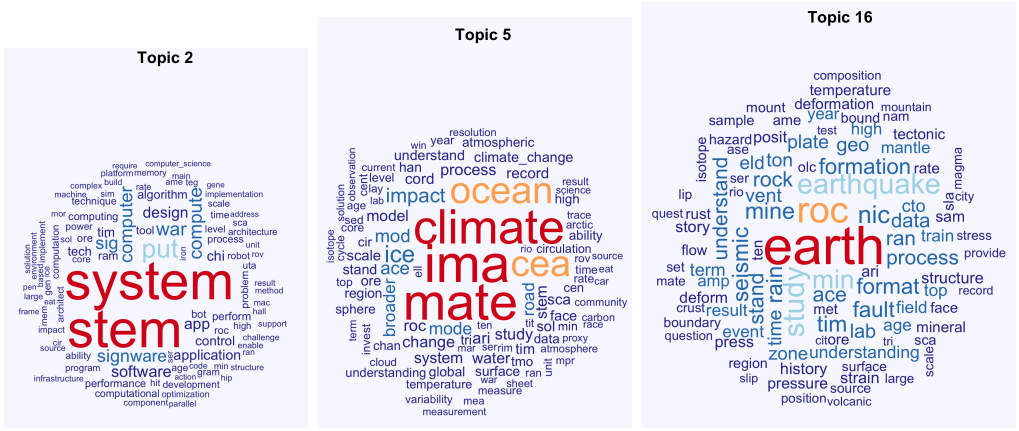


FIGURE 6.2. Topics from LDA applied to 13,092 NSF awards. The size of the terms correspond to the value of the term in the topic.

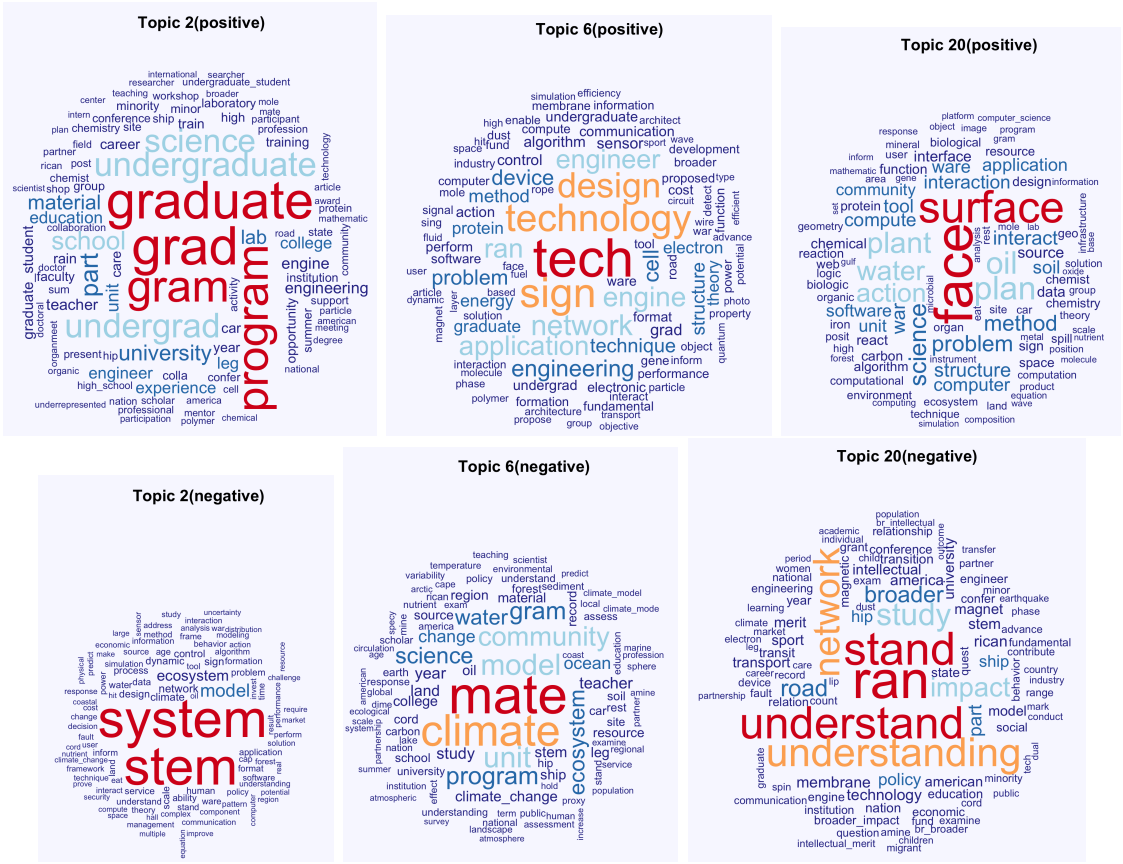


FIGURE 6.3. Topics for the model specified in (4.1). The size of the terms correspond to the absolute value of the terms in the topics. We display both the positive and negatively weighted topic values.

subspaces of different dimensions that is differentiable and can be computed using principal angles, we avoid MCMC algorithms that jump between models of different dimensions. We

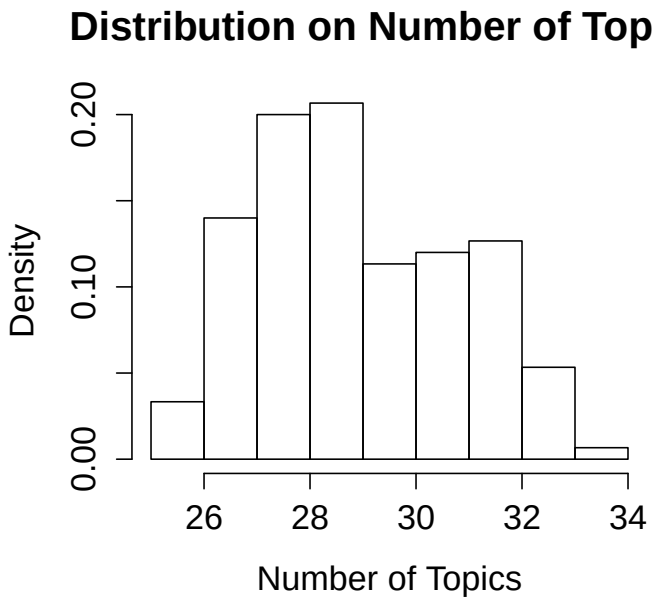


FIGURE 6.4. Distribution on the number of topics inferred by the model. The uncertainty in the number of topics is greater than in the mixture of subspaces applications in the previous subsections.

suspect that this idea of embedding or representing models of different dimensions by embedding them into a common space 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 subspaces.

Scaling our estimation procedure to higher dimensions and more samples will require greater computational efficiency and an EM-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.

ACKNOWLEDGEMENTS

SM and BST would like to thank Robert Calderbank, Daniel Runcie, and Jesse Windle for useful discussions. 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 AFOSR FA9550-13-1-0133,

NSF DMS-1209136, and NSF DMS-1057064. This work of LL is supported by a DUKE iiD grant and grant R01ES017240 from the National Institute of Environmental Health Sciences (NIEHS) of the National Institute of Health (NIH).

References

[1] R.J. Adcock. A problem in least squares. *The Analyst*, 5:53–54, 1878.

[2] S. Amari. Differential geometry of curved exponential families — curvatures and information loss. *Annals of Statistics*, 10(2):357–385, 1982.

[3] A. Ashikhmin and A.R. Calderbank. Grassmannian packings from operator Reed-Muller codes *IEEE Trans. Info. Theory*, 56(11):5689–5714, 2003.

[4] K. Bache and M. Lichman. UCI Machine Learning Repository `http://archive.ics.uci.edu/ml`, 2013.

[5] A. Banerjee, I.S. Dhillon, J. Ghosh, S. Sra. Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research*, 6:1345–1382, 2005.

[6] A.R. Barron. The expoential convergence of posterior probabilities with implications for Bayes estimators of density functions. *Technical Report 7*, 1988.

[7] M. Belkin and P. Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[8] P. Bendich, S. Mukherjee, and B. Wang. Local homology transfer and stratification learning. *ACM-SIAM Symposium on Discrete Algorithms*, 2012.

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

[10] C.M. Carvalho, J. Chang, J.E. Lucas,J.R. Nevins, Q. Wang,M. West. High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics. *Journal of the American Statistical Association*, 103(484):1438–1456, 2008.

[11] J.H. Conway, R.H. Hardin, and N.J.A. Sloane. Packing lines, planes, etc.: packings in Grassmannian spaces. *Experiment. Math.*, 5(2):83–159, 1996.

[12] R.D. Cook. Fisher lecture: Dimension reduction in regression. *Statistical Science*, 22(1):1–26, 2007.

[13] R. Detrano., A. Janosi, W. Steinbrunn, M. Pfisterer, J. Schmid, S. Sandhu, K. Guppy, S. Lee, V. Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *American Journal of Cardiology.* 604: 304-310, 1989.

[14] S. Deerwester, S.T. Dumais, G.W. Furnas,T.K. Landauer, R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science.* 41(6): 391-407, 1990.

[15] D. Donoho and C. Grimes. Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100:5591–5596, 2003.

[16] F.Y. Edgeworth. On the reduction of observations. *Philosophical Magazine*, pages 135–141, 1884.

[17] B. Efron. The geometry of exponential families. *Annals of Statistics*, 6:362–376, 1978.

[18] R.A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Statistical Society A*, 222:309–368, 1922.

[19] Dan Geiger, David Heckerman, Henry King, and Christopher Meek. Stratified exponential families: Graphical models and model selection. *Annals of Statistics*, 29(2):505–529, 2001.

[20] S. Ghoshal, J.K. Ghosh, and R.V. Ramamoorthi. Posterior consistency of Dirichlet mixtures in density estimation. *Annals of Statistics*, 27, 143–158, 1999.

[21] S. Ghoshal. Dirichlet process, related priors and posterior asymptotics. *In Bayesian Nonparametrics*, Editors: N. L. Hjort, C. Holmes, P. Müller, S.G. Walker, 36–83, 2010.

[22] E. Giné and V. Koltchinskii. Empirical graph Laplacian approximation of Laplace-Beltrami operators: large sample results. In *High-dimensional probability*, volume 51 of *IMS Lecture Notes Monogr. Ser.*, pages 238–259. Inst. Math. Statist., Beachwood, OH, 2006.

[23] G. Golub and C. Van Loan. Matrix Computations. 4th Ed, John Hopkins University Press, Baltimore, MD, 2013.

[24] Mark Goresky and Robert MacPherson. *Stratified Morse Theory*. Springer-Verlage, 1988.

[25] J. Hamm and D. D. Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning *Advances in NIPS*, 17, 2005.

[26] T.F. Hansen and D. Houle. Measuring and comparing evolvability and constraint in multivariate characters *Journal of Evolutionary Biology*, 21:1201–1219, 2008.

[27] G. Haro, G. Randall, and G. Sapiro. Stratification learning: Detecting mixed density and dimensionality in high dimensional point clouds. *International Conference on Machine Learning*, 25, 2008.

[28] P.D. Hoff. Simulation of the Matrix Bingham–von Mises–Fisher Distribution, with Applications to Multivariate and Relational Data. *J. Compt. Graph. Stat.*, 18(2):438–456, 2009.

[29] T. Hoffman. Probabilistic Latent Semantic Indexing. *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 50–57, 1999.

[30] H. Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educ. Psych.*, 24(6):417–441, 1933.

[31] W. Jiang and M.A. Tanner. Gibbs posterior for variable selection in high-dimensional classification and data mining. *Ann. Statist.*, 36(5):2025–2550, 2008.

[32] R. Lande. Quantitative genetic-analysis of multivariate evolution, applied to brain-body size allometry *Evolution*, 33(1):402–416, 1979.

[33] G. Lerman and T. Zhang. Probabilistic recovery of multiple subspaces in point clouds by geometric l_p minimization. *Annals of Statistics*, 39(5):2686–2715, 2010.

[34] O.L. Mangasarian and W.H. Wolberg. Cancer diagnosis via linear programming. *Siam News* 23(5):1&18, 1990.

[35] A.K. McCallum. MALLET: A Machine Learning for Language Toolkit. `http://mallet.cs.umass.edu`, 2002.

[36] S. Mukherjee, D-X. Zhou, and Q. Wu. Learning gradients and feature selection on manifolds. *Bernoulli*, 16(1):181–207, 2010.

[37] NSF 2010 Awards. `http://www.nsf.gov/awardsearch/download.jsp`.

[38] G. Page, A. Bhattacharya, and D.B. Dunson. Classification via Bayesian nonparametric learning of affine subspaces. *J. Am. Statist. Assoc.*, 108(501):187–201, 2013.

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

[40] C. R. Rao. Information and accuracy obtainable in the estimation of statistical parameters. *Bulletin Calcutta Math. Soc.*, 37:81–91, 1945.

[41] J. Reisinger, A. Waters, B. Silverthorn, and R.J. Mooney. Spherical Topic Models. *Proceedings of the 27th ICML*. 2010.

[42] S. Roweis and L. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290:2323–2326, 2000.

[43] L. Schwartz. On Bayes procedures. *Z. Wahrsch. Verw. Gebiete*, 4:10–26, 1965.

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

[45] R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1945–1959, 2005.

[46] H.C. Wu, R.W.P. Luk,K.F. Wong, K.L. Kwok. Interpreting TF-IDF Term Weights As Making Relevance Decisions. *ACM Trans. Inf. Syst.*, 26(3):1–37, 2008.

[47] Y. Wu and S. Ghosal. The L_1-consistency of Dirichlet mixtures in multivariate Bayesian density estimation. *J. Multivar. Anal.*, 101(10):2411–2419, 2010.

[48] L. Zheng and D.N.C. Tse. Communication on the Grassmann manifold: A geometric approach to the noncoherent multiple-antena channel. *IEEE Trans. Info. Theory*, 48(2):359–383, 2002.

[49] S. Zhong, J. Ghosh. Generative model-based document clustering: A comparative study.. *Knowledge and Information Systems*, 8(3):374–384, 2005.

Department of Statistical Science, Duke University, Durham, NC 27708-0251.

E-mail address: brian.st.thomas@duke.edu

Department of Statistics and Data Sciences, The University of Texas at Austin, Austin, TX 78746.

E-mail address: lizhen.lin@austin.utexas.edu

Computational and Applied Mathematics Initiative, Department of Statistics, University of Chicago, Chicago, IL 60637-1514.

E-mail address: lekheng@galton.uchicago.edu

Departments of Statistical Science, Mathematics, Computer Science, Duke University, Durham, NC 27708-0251.

E-mail address: sayan@stat.duke.edu