On a minimum enclosing ball of a collection of linear subspaces

Tim Marrinan[a,*], P.-A. Absil[b], Nicolas Gillis[a]

[a] *Department of Mathematics and Operational Research, University of Mons, Mons, Belgium*
[b] *ICTEAM Institute, UCLouvain, Louvain-la-Neuve, Belgium*

Abstract

This paper concerns the minimax center of a collection of linear subspaces. For k-dimensional subspaces of an n-dimensional vector space, this can be cast as finding the center of a minimum enclosing ball on a Grassmann manifold. For subspaces of differing dimension, the setting becomes a disjoint union of Grassmannians rather than a single manifold, and the problem is no longer well-defined. However, natural geometric maps exist between these manifolds with a well-defined notion of distance for the images of the subspaces under the mappings. Solving the problem in this context leads to a candidate minimax center on each of the constituent manifolds, but does not provide intuition about which candidate is the best representation of the data. Additionally, the solutions of different rank are generally not nested so a deflationary approach will not suffice, and the problem must be solved independently on each manifold. We propose an optimization problem parametrized by the rank of the minimax center. The solution is computed with a subgradient algorithm applied to the dual problem. By scaling the objective and penalizing the information lost by the rank-k minimax center, we jointly recover an optimal dimension, k^*, and a subspace at the center of the minimum enclosing ball, $\mathbf{U}^*$, that best represents the data.

Keywords: Grassmann manifold; minimum enclosing ball; minimax center; subgradient; low-rank; order-selection
2010 MSC: 90C47, 14M15, 49J35

1. Introduction

Finding the minimum enclosing ball (MEB) of a finite collection of points in a metric space, or the ℓ_∞-center of mass, is a topic of broad interest in the mathematical community [1, 2, 3, 4, 5, 6, 7]. For Euclidean data, the problem has been well studied, and research has transitioned towards finding approximate solutions efficiently when computing the MEB exactly is impractical [2, 6]. A breakthrough technique of Bădoiu and Clarkson [2] finds an optimal subset of the data, called

[★] This work is supported by the Fonds de la Recherche Scientifique (FNRS) and the Fonds Wetenschappelijk Onderzoek - Vlanderen (FWO) [EOS Project no O005318F-RG47]. N.G. acknowledges the support of the European Research Council [ERC starting grant no 679515].

[*] Corresponding author

Email addresses: timothy.marrinan@umons.ac.be (Tim Marrinan), pa.absil@uclouvain.be (P.-A. Absil), nicolas.gillis@umons.ac.be (Nicolas Gillis)

a core-set, such that finding the exact MEB of the core-set is computationally tractable. They show that the radius of this core-set will be bounded by $(1+\epsilon)$ times the radius of the entire data set, where ϵ depends only on the number of points in the core-set [2]. That is, the minimum enclosing ball can be approximated to any desired accuracy by increasing the number of points in the core-set, and the number of points needed for the radius of the core-set to be at most ϵ percent larger than the true radius is $\lceil \frac{2}{\epsilon} \rceil$. This solution represents efforts to make ℓ_∞-averaging possible for complex data sets.

The difficulty in computing the MEB of Euclidean data is due to the massive size of data sets to be averaged, however in less traditional settings other difficulties arise and contribute to the complexity of this task. Many modern problems are formulated on manifolds instead of Euclidean space in situations where the manifold geometry better represents the natural structure of the data model [8, 9, 10]. Afsari provided existence and uniqueness conditions for Riemannian ℓ_p centers of mass [11], and with this type of structure in mind, Arnaudon and Nielsen [1] adapted the efficient MEB algorithm of Bădoiu and Clarkson to Riemannian manifolds. For linear subspace data, a subclass of data addressed by [1], this work was further generalized by Renard, Gallivan, and Absil [3, 12]. They created a technique that applies to points lying on a disjoint union of Grassmann manifolds, that is, a collection of p_i-dimensional subspaces of $\mathbb{R}^n$ where p_i is not necessarily equal for all i. Although the data comes from a collection of manifolds, the MEB must be computed on one individual Grassmannian and the choice of which is not obvious. Determining which Grassmannian provides the best center for a collection of subspaces is one of the tasks of this manuscript, and we provide a geometrically motivated criteria for automatically selecting this manifold.

With subspace data, it is natural to think of the center of the Grassmannian minimum enclosing ball (GMEB) as the common information in the data set. Common subspace extraction can be found in subspace clustering [13], domain adaptation, and subspace alignment. These tools can be used in a plethora of tasks in pattern recognition including subspace tracking [14], face recognition [15, 16], video action recognition [17, 16], infected patient diagnosis [18], adaptive sorting [19], model reduction [20], and many more. Common subspace extraction is frequently done by finding the ℓ_2- or ℓ_1-center in cases where outliers are present in the data collection, but if the data are drawn from a uniform distribution whose support is a ball, the ℓ_∞-center gives the maximum likelihood estimator for the center of the support and thus may be preferred when all the subspaces have been drawn from a single uniform distribution [11]. Furthermore, techniques have been developed to prune outliers from data sets using the ℓ_∞-norm, with theoretical guarantees in some circumstances [21].

In this paper, we present a novel technique to accurately estimate the GMEB for a collection of linear subspaces of possibly differing dimension, and a geometrically inspired order-selection rule to identify the Grassmannian that best represents the shared information in the data. Choosing the ideal manifold on which to perform the ℓ_∞-averaging is inherently related to finding a common subspace of optimal rank, and thus the numerical experiments explore the relationships between different rank-adaptive subspace averaging methods.

The main contributions of the paper are summarized as follows. We propose

- a subgradient approach to solve the dual of the GMEB problem for subspaces of differing dimensions. A duality gap of zero certifies the solution as optimal.

- an unsupervised order-selection rule for the dimension of the center of the GMEB.

- a warm-start initialization for the subgradient algorithm that reduces the number of iterations needed for the subgradient algorithm to converge.
- a hybrid method for order-selection which modifies the existing rule of [22] for use with the center of the GMEB.
- a synthetic data model that allows us to measure the accuracy of an estimate for the center of the GMEB, and demonstrate the effectiveness of the proposed technique using data generated with this model.

Finally, we compare the proposed order-selection rules to existing methods for automatic order selection in subspace averaging with numerical experiments.

2. Mathematical background: Grassmannian minimum enclosing ball

In this section we provide the mathematical background necessary to formulate the GMEB problem for subspaces of differing dimension. We begin by stating the relevant properties of invariant metrics, a standard reference on this topic is [23]. We recall the maps defined in [24] that associate a subset of points on a single manifold with each subspace from the collection and the point-to-set distance that measures the dissimilarity of these sets. Finally, we explicitly state the minimax optimization problem that defines this GMEB.

Denote by $\mathrm{Gr}(k,n)$ the Grassmann manifold of k-dimensional subspaces in $\mathbb{R}^n$. If A is an $n \times k$ matrix with full column rank, the column space of A, $\mathrm{col}(A)$, defines a subspace that can be identified with a point $\mathbf{A} \in \mathrm{Gr}(k,n)$. Any matrix in the $\mathrm{GL}(k)$ orbit of A will have the same column space, so we assume without loss of generality that the chosen representative for $\mathbf{A} \in \mathrm{Gr}(k,n)$ is an orthonormal basis, $A \in \mathbb{R}^{n\times k}$ with $A^TA = I$. This assumption simplifies notation, however the choice of basis is not unique so a measure of distance between points must be orthogonally invariant. To see this, let $\mathrm{O}(k)$ denote the set of $k \times k$ orthogonal matrices. If $Q_k \in \mathrm{O}(k)$ then AQ_k is another orthonormal basis for $\mathbf{A}$. For any two points, $\mathbf{A}, \mathbf{B} \in \mathrm{Gr}(k,n)$, there exists a set of k principal angles, $0 \leq \theta_1(\mathbf{A},\mathbf{B}) \leq \cdots \leq \theta_k(\mathbf{A},\mathbf{B}) \leq \pi/2$, defined recursively as

$$\begin{aligned}
\theta_1(\mathbf{A},\mathbf{B}) &\doteq \min_{\mathbf{a}_1\in\mathbf{A},\mathbf{b}_1\in\mathbf{B}} \cos^{-1}\left(\frac{\mathbf{a}_1^T\mathbf{b}_1}{\|\mathbf{a}_1\|_2\|\mathbf{b}_1\|_2}\right), \text{ and for } i = 2,\dots,k \\
\theta_i(\mathbf{A},\mathbf{B}) &\doteq \min_{\mathbf{a}_i\in\mathbf{A},\mathbf{b}_i\in\mathbf{B}} \cos^{-1}\left(\frac{\mathbf{a}_i^T\mathbf{b}_i}{\|\mathbf{a}_i\|_2\|\mathbf{b}_i\|_2}\right) \\
&\quad \text{s.t. } \mathbf{a}_j^T\mathbf{a}_i = 0 \text{ for } j < i \\
&\quad\quad\ \mathbf{b}_j^T\mathbf{b}_i = 0 \text{ for } j < i.
\end{aligned} \tag{1}$$

The vectors that form these angles, $\{\mathbf{a}_1,\dots,\mathbf{a}_k\}$ and $\{\mathbf{b}_1,\dots,\mathbf{b}_k\}$, are called the left and right principal vectors, respectively, and when normalized, these vectors form orthonormal bases $A, B \in \mathbb{R}^{n\times k}$, for the spaces $\mathbf{A}$ and $\mathbf{B}$. The principal angles and principal vectors can be computed via the singular value decomposition (SVD) [25]. Let $A^TB = V\Sigma W^T$ be a thin SVD with the singular values sorted in nonincreasing order, so that $V \in \mathbb{R}^{k\times k}$ with $V^TV = I$, Σ is a $k \times k$ diagonal matrix, and $W \in \mathbb{R}^{k\times k}$ with $W^TW = I$. Then $\Sigma_{ii} = \cos(\theta_i(\mathbf{A},\mathbf{B}))$, where $\theta_i(\mathbf{A},\mathbf{B})$ is the ith principal angle separating $\mathbf{A}$ and $\mathbf{B}$, with associated left and right principal vectors $\mathbf{a}_i = A\mathbf{v}_i$ and $\mathbf{b}_i = B\mathbf{w}_i$ for $i = 1,\dots,k$.

Let $d : \mathrm{Gr}(k,n) \times \mathrm{Gr}(k,n) \to \mathbb{R}$ be a metric. If for all $\mathbf{A}, \mathbf{B} \in \mathrm{Gr}(k,n)$ and for all $Q_n \in \mathrm{O}(n)$ the left action of Q_n on A and B by multiplication does not change the value of the metric, that is, $d(\mathbf{A}, \mathbf{B}) = d(\mathbf{Q_n A}, \mathbf{Q_n B})$, then d is said to be orthogonally invariant. Orthogonally invariant metrics depend only on the relative position of $\mathbf{A}$ and $\mathbf{B}$, so as a result of [26, Thm. 3], d can be written as a function of the vector of principal angles separating $\mathbf{A}$ and $\mathbf{B}$, $\boldsymbol{\theta}(\mathbf{A}, \mathbf{B}) \in \mathbb{R}^k$. Additionally, for $Gr(k,n)$ with either $k \neq 2$ or $n \neq 2$ there is an essentially unique invariant Riemannian metric (up to scaling) which yields $d(\mathbf{A}, \mathbf{B}) = \|\boldsymbol{\theta}(\mathbf{A}, \mathbf{B})\|_2$, and is frequently referred to as the geodesic distance based on arc length [26].

Let $\mathcal{D} = \{\mathbf{X}_i\}_{i=1}^M$ be a finite collection of subspaces of $\mathbb{R}^n$ with possibly different dimensions, so that $\dim(\mathbf{X}_i) = p_i$. For the set of positive integers $\mathcal{P} = \{\dim(\mathbf{X}_i) : \mathbf{X}_i \in \mathcal{D}\}$ we can consider $\mathcal{D}$ as a collection of points lying on the disjoint union of Grassmann manifolds, $\mathbf{X}_i \in \coprod_{p \in \mathcal{P}} \mathrm{Gr}(p,n)$. To account for the difference in subspace dimensions, we adopt the convention of [24] by redefining $d(\mathbf{U}, \mathbf{X}_i)$ as the minimum distance between $\mathbf{U}$ and a subset of points on $\mathrm{Gr}(k,n)$, appropriately defined for each $\mathbf{X}_i \in \mathcal{D}$. Each subspace is associated with one of two types of subset, which are defined by

$$\begin{aligned} \Omega_+(\mathbf{X}_i) &\doteq \{\mathbf{Y} \in \mathrm{Gr}(k,n) : \mathbf{X}_i \subseteq \mathbf{Y}\} \text{ for } p_i < k, \text{ and} \\ \Omega_-(\mathbf{X}_i) &\doteq \{\mathbf{Y} \in \mathrm{Gr}(k,n) : \mathbf{Y} \subseteq \mathbf{X}_i\} \text{ for } p_i \geq k. \end{aligned} \tag{2}$$

We use $\Omega(\mathbf{X}_i)$ when referring to either type generically. For $\mathbf{X}_i$ such that $p_i < k$, $\Omega_+(\mathbf{X}_i)$ is the set of all points of $\mathrm{Gr}(k,n)$ containing $\mathbf{X}_i$. Alternatively when $\mathbf{X}_i$ is a p_i-plane with $p_i > k$, $\Omega_-(\mathbf{X}_i)$ is all k-dimensional subspaces contained in $\mathbf{X}_i$, and when $p_i = k$ the subset of points is just the singleton, $\mathbf{X}_i$.

Finally, we overload the notation for distance so that

$$d_{\mathrm{Gr}(k,n)}(\mathbf{U}, \mathbf{X}_i) \doteq d_{\mathrm{Gr}(k,n)}(\mathbf{U}, \Omega(\mathbf{X}_i)) = \min\{d(\mathbf{U}, \mathbf{Y}_i) : \mathbf{Y}_i \in \Omega(\mathbf{X}_i)\} \tag{3}$$

when the distance is being measured on $\mathrm{Gr}(k,n)$ and the data comes from Grassmann manifolds of possibly differing dimension. This is the proposed distance of [24], which is well-defined for a fixed value of k. Figure 1 shows an illustration of this distance as the length of the shortest path between a point, $\mathbf{U}$, and the sets of points, $\Omega(\mathbf{X}_i)$ for $i = 1, \ldots, 3$. In this particular case $\mathbf{Y}_3 \in \mathrm{Gr}(k,n)$ so $\mathbf{Y}_3 = \mathbf{X}_3 = \Omega(\mathbf{X}_3)$.

The minimum in Equation (3) always exists because $\Omega(\mathbf{X}_i)$ is a closed subset of the Grassmannian, and the points satisfying $\mathbf{Y}_i \in \arg\min_{\mathbf{Y} \in \Omega(\mathbf{X}_i)} d(\mathbf{U}, \mathbf{Y})$ are independent of the choice of orthogonally invariant distance measure. Let $U^T X_i = V \Sigma W^T$ be a thin SVD. One point that achieves the minimum distance is the column space of the matrix defined by

$$Y_i \doteq \begin{cases} [X_i \mathbf{w}_1, \ldots, X_i \mathbf{w}_k] & \text{for } p_i \geq k; \\ \left[X_i \mathbf{w}_1, \ldots, X_i \mathbf{w}_{p_i}, U \mathbf{v}_{p_i+1}, \ldots, U \mathbf{v}_k\right] & \text{otherwise.} \end{cases} \tag{4}$$

This derivation can be found in, e.g. [27].

This formalism implies that distances can be written as a function of exactly k principal angles regardless of the dimension of $\mathbf{X}_i$, and conveniently the definition agrees with many pseudo-metrics commonly used in the literature that measure similarity as a function of the (possibly less than k) principal angles between subspaces of different dimension. It should be clear, however, that this is not a metric because the distance between $\mathbf{A}$ and $\mathbf{B}$ will be zero if $\mathbf{A}$ is a proper subspace of $\mathbf{B}$, despite being non-identical.

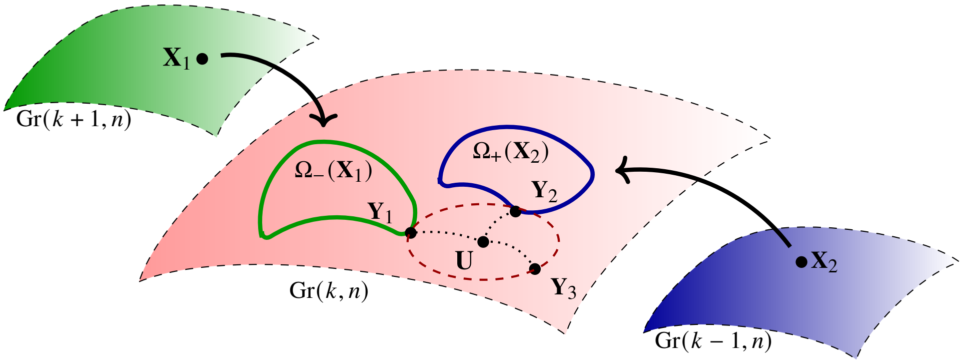


Figure 1: Illustration of the minimum point-to-set distance on $\text{Gr}(k,n)$ between $\mathbf{U}$ and the sets $\Omega_-(\mathbf{X}_1)$, $\Omega_+(\mathbf{X}_2)$, and $\mathbf{Y}_3$, associated with points on $\text{Gr}(k+1,n)$, $\text{Gr}(k-1,n)$, and $\text{Gr}(k,n)$, respectively. The points that realize the minimum distance are $\mathbf{Y}_1 \in \Omega_-(\mathbf{X}_1)$, $\mathbf{Y}_2 \in \Omega_+(\mathbf{X}_2)$, and $\mathbf{Y}_3$. The point $\mathbf{U}$ is the center of the minimum enclosing ball of $\mathbf{Y}_1$, $\mathbf{Y}_2$, and $\mathbf{Y}_3$.

This manuscript is concerned with computing the minimax center, i.e., the center of the GMEB, on $\text{Gr}(k,n)$ for the collection of subspaces, $\mathcal{D}$, using the point-to-set distance. However, rather than using a metric on $\text{Gr}(k,n)$ we measure dissimilarity by the squared chordal distance, $d(\mathbf{A},\mathbf{B}) = \|\sin(\boldsymbol{\theta}(\mathbf{A},\mathbf{B}))\|_2^2$. The minimum point-to-set distance using the squared chordal distance is

$$\begin{aligned} d_{\text{Gr}(k,n)}(\mathbf{U},\mathbf{X}_i) &= \|\sin(\boldsymbol{\theta}(\mathbf{U},\mathbf{Y}_i))\|_2^2 \\ &= \frac{1}{2}\|U_kU_k^T - Y_iY_i^T\|_F^2 \\ &= k - \text{Tr}(U^TY_iY_i^TU) \\ &= \min\{k,p_i\} - \text{Tr}(U^TX_iX_i^TU), \end{aligned} \tag{5}$$

where $\boldsymbol{\theta}(\mathbf{U},\mathbf{Y}_i) \in \mathbb{R}^k$ is the vector of principal angles between $\mathbf{U}$ and a point $\mathbf{Y}_i \in \Omega(\mathbf{X}_i)$ that attains the minimum. The final equality in Equation (5) can be seen from the definition of $\mathbf{Y}_i$ in Equation (4) and will be demonstrated in Equation (32). Note that it is not necessary to know $\mathbf{Y}_i$ in order to compute $d_{\text{Gr}(k,n)}(\mathbf{U},\mathbf{X}_i)$. With this definition and choice of distance measurement, the minimax problem we wish to solve is

$$\underset{\mathbf{U}\in\text{Gr}(k,n)}{\arg\min} \max_{i=1,\ldots,M} d_{\text{Gr}(k,n)}(\mathbf{U},\mathbf{X}_i). \tag{6}$$

Using the notion of distance from Equation (3), an algorithm was proposed by [3] to solve Problem (6) for a given value of k. Since the data is not of uniform dimension, it is one of our goals to find the solution across all possible values of k that best represents the common subspace in the data. In Section 5 we propose an order-selection rule for comparing solutions of different dimension, however we must first be able to find the solutions of different dimension efficiently. As we will see in Section 5.1, $\mathbf{U}^*(k) \in \text{Gr}(k,n)$ is not always contained in $\mathbf{U}^*(k+1) \in \text{Gr}(k+1,n)$, so it is not possible to construct the respective solutions iteratively via deflation. Instead the problem needs to be solved independently for each value of k.

3. Dual formulation

Problem (6) is nonconvex and challenging to optimize directly. Therefore, in this section we formulate the dual problem which can be solved efficiently. The dual variables also provide a primal-feasible solution, which can be tested for optimality.

Using Equation (5), Problem (6) can be written as one with matrix arguments that can be identified with the Grassmannian points they represent. That is,

$$\begin{aligned} &\operatorname*{arg\,min}_{U\in\mathbb{R}^{n\times k}} \max_{i=1,\dots,M} \left(\min\{k,p_i\} - \operatorname{Tr}(U^T X_i X_i^T U)\right) \\ &\text{s.t. } U^T U = I, \end{aligned} \tag{7}$$

where U is an orthonormal basis for $\mathbf{U}$, X_i is an orthonormal basis for $\mathbf{X}_i$, and $p_i = \dim(\mathbf{X}_i)$. A solution to (6) is then the column space of a solution to (7), $\mathbf{U}^* = \operatorname{col}(U^*)$. For ease of notation we will treat the dual problem as a minimization, so we reformulate the primal as,

$$\begin{aligned} &\operatorname*{arg\,max}_{U\in\mathbb{R}^{n\times k}} \min_{i=1,\dots,M} -\left(\min\{k,p_i\} - \operatorname{Tr}(U^T X_i X_i^T U)\right) \\ &\text{s.t. } U^T U = I. \end{aligned} \tag{8}$$

Adding an auxiliary variable τ, the quadratic cost function to be minimized is replaced by a smooth linear objective that is maximized with respect to quadratic inequality constraints,

$$\begin{aligned} &\operatorname*{arg\,max}_{U\in\mathbb{R}^{n\times k},\tau\in\mathbb{R}} \ \tau \\ &\text{s.t. } -\tau - \min\{k,p_i\} + \operatorname{Tr}(U^T X_i X_i^T U) \geq 0 \text{ for } i = 1,\dots,M, \\ &U^T U = I. \end{aligned} \tag{9}$$

This is essentially the same construction as in [3]. The authors of [3] go on to compute an intermediate solution to this problem via the Karush–Kuhn–Tucker conditions, and iterate to a stationary point by taking geodesic steps towards the subspace with the maximum distance to the current iterate of the primal variable. This contrasts with the proposed approach, where a solution to (6) is found by optimizing the dual problem.

Let $\boldsymbol{\lambda} = [\lambda_1,\dots,\lambda_M]^T$ be a vector of Lagrange multipliers associated with the inequality constraints in (9). Dualizing only the inequality constraints leads to the Lagrangian

$$\mathcal{L}(U,\tau,\boldsymbol{\lambda}) = \tau + \sum_{i=1}^{M} \lambda_i \left(-\tau - \min\{k,p_i\} + \operatorname{Tr}(U^T X_i X_i^T U)\right), \tag{10}$$

such that $U^T U = I$ and $\lambda_i \geq 0$ for $i = 1,\dots,M$. The dual cost function is then found by maximizing $\mathcal{L}$ over U and τ,

$$f(\boldsymbol{\lambda}) = \sup_{\tau} \left(\tau - \sum_{i=1}^{M} \lambda_i \tau\right) - \sum_{i=1}^{M} \lambda_i \min\{k,p_i\} + \sup_{U^T U = I} \operatorname{Tr}(U^T (\sum_{i=1}^{M} \lambda_i X_i X_i^T) U). \tag{11}$$

The maximum over τ yields $f(\boldsymbol{\lambda}) = \infty$ unless $\|\boldsymbol{\lambda}\|_1 = 1$, in which case the first term is zero. The final term in (11) is a well-known problem that is maximized by the sum of the k largest eigenvalues of $\sum_{i=1}^M \lambda_i X_i X_i^T$ [28]. Let $d_1(\boldsymbol{\lambda}) \geq d_2(\boldsymbol{\lambda}) \geq \cdots \geq d_n(\boldsymbol{\lambda})$ be the eigenvalues of $\sum_{i=1}^M \lambda_i X_i X_i^T$ and let $\mathbf{v}_1(\boldsymbol{\lambda}), \mathbf{v}_2(\boldsymbol{\lambda}), \ldots, \mathbf{v}_n(\boldsymbol{\lambda})$ be the associated orthonormal eigenvectors. The argument $\boldsymbol{\lambda}$ is included to emphasize that the eigendecomposition depends on $\boldsymbol{\lambda}$. The supremum is then $\sum_{j=1}^k d_j(\boldsymbol{\lambda})$, and is achieved by the matrix whose columns are the k dominant eigenvectors,

$$U_\lambda \doteq [\mathbf{v}_1(\boldsymbol{\lambda}), \ldots, \mathbf{v}_k(\boldsymbol{\lambda})]. \tag{12}$$

Thus the dual cost can be written as

$$f(\boldsymbol{\lambda}) = -\sum_{i=1}^M \lambda_i \min\{k, p_i\} + \sum_{j=1}^k d_j(\boldsymbol{\lambda}), \tag{13}$$

and finally, we wish to solve the problem,

$$\arg\min_{\boldsymbol{\lambda} \in \mathbb{R}^M} f(\boldsymbol{\lambda}) \text{ s.t. } \|\boldsymbol{\lambda}\|_1 = 1 \text{ and } \lambda_i \geq 0 \text{ for } i = 1, \ldots, M. \tag{14}$$

4. Solution via subgradient

The dual cost in (13) is a locally Lipschitz convex function. However, it is not differentiable at values of $\boldsymbol{\lambda}$ for which $d_k(\boldsymbol{\lambda}) = d_{k+1}(\boldsymbol{\lambda})$, that is, at values for which the kth and $(k+1)$st eigenvalues of $\sum_{i=1}^M \lambda_i X_i X_i^T$ are equal [28, Corr. 3.10]. There are many efficient ways to optimize such a function. In this section we recall how the subgradient method [29] can be applied to solve this dual problem. After a subgradient has been computed, the well-developed literature of subgradient algorithms provides a variety of techniques and step sizes to optimize Problem (14) with associated convergence guarantees.

Recall that a vector $\mathbf{g} \in \mathbb{R}^M$ is a subgradient of $f : \mathbb{R}^M \longrightarrow \mathbb{R}$ at $\mathbf{x}$ in the domain of f if for all $\mathbf{z}$ in the domain of f,

$$f(\mathbf{z}) \geq f(\mathbf{x}) + \mathbf{g}^T(\mathbf{z} - \mathbf{x}).$$

In this case we denote that $\mathbf{g}$ is in the subdifferential of f at $\mathbf{x}$ by writing $\mathbf{g} \in \partial f(\mathbf{x})$. If f is differentiable at $\mathbf{x}$ then the gradient is the only subgradient and $\mathbf{g} = \nabla f(\mathbf{x}) = \partial f(\mathbf{x})$.

To minimize f in Problem (14), the subgradient method uses the iteration

$$\boldsymbol{\lambda}^{(t+1)} = \Pi(\boldsymbol{\lambda}^{(t)} - \alpha^{(t)}\mathbf{g}^{(t)}), \tag{15}$$

where $\alpha^{(t)}$ is a step size selected to guarantee that the sequence $\{\boldsymbol{\lambda}^{(t)}\}_{t=1}^\infty$ converges (in distance) to the optimum, $\boldsymbol{\lambda}^*$, and $\Pi : \mathbb{R}^M \longrightarrow \{\mathbf{x} : \|\mathbf{x}\|_1 = 1, x_i \geq 0 \text{ for } i = 1, \ldots, M\} \subset \mathbb{R}^M$ projects the iterate into the unit simplex.

There is a standard trick for computing a subgradient of the dual function that can be adapted to this problem from nonlinear optimization texts such as [30]. Write the Lagrangian as $\mathcal{L}(U, \tau, \boldsymbol{\lambda}) = q(U, \tau) + \boldsymbol{\lambda}^T \mathbf{g}(U, \tau)$, where $q(U, \tau)$ is the primal objective function and $\mathbf{g}(U, \tau) \in \mathbb{R}^M$ is the vector of constraint values. Given the dual variable, $\boldsymbol{\lambda}^{(t)} \in \mathbb{R}^M$, at iteration t, let $(U_{\boldsymbol{\lambda}^{(t)}}, \tau_{\boldsymbol{\lambda}^{(t)}})$ be the primal variable that maximizes the Lagrangian. Then $\mathbf{g}^{(t)} = \mathbf{g}(U_{\boldsymbol{\lambda}^{(t)}}, \tau_{\boldsymbol{\lambda}^{(t)}})$ is a subgradient of f at $\boldsymbol{\lambda}^{(t)}$.

In our case $U_{\boldsymbol{\lambda}^{(t)}}$ is defined by Equation (12) and the ith element of the constraint vector is $g_i(U_{\boldsymbol{\lambda}^{(t)}}, \tau_{\boldsymbol{\lambda}^{(t)}}) = -\tau_{\boldsymbol{\lambda}^{(t)}} - \min\{k, p_i\} + \text{Tr}(U_{\boldsymbol{\lambda}^{(t)}}^T X_i X_i^T U_{\boldsymbol{\lambda}^{(t)}})$. However, the constant vector $[-\tau_{\boldsymbol{\lambda}^{(t)}}, \ldots, -\tau_{\boldsymbol{\lambda}^{(t)}}]^T \in \mathbb{R}^M$ does not affect the direction after projection onto the unit simplex, so a subgradient of $f(\lambda^{(t)})$ is

$$\mathbf{g}^{(t)} = \begin{pmatrix} -\min\{k, p_1\} + \mathrm{Tr}(U_{\lambda^{(t)}}^T X_1 X_1^T U_{\lambda^{(t)}}) \\ \vdots \\ -\min\{k, p_M\} + \mathrm{Tr}(U_{\lambda^{(t)}}^T X_M X_M^T U_{\lambda^{(t)}}) \end{pmatrix}. \tag{16}$$

We can check that $\mathbf{g}^{(t)}$ is a subgradient of f as follows. For any $\tilde{\lambda} \in \mathbb{R}^M$ such that $\|\tilde{\lambda}\|_1 = 1$ and $\tilde{\lambda}_i \geq 0$ for $i = 1, \ldots, M$ we have

$$\begin{aligned} f(\lambda^{(t)}) + \mathbf{g}^{(t)T}(\tilde{\lambda} - \lambda^{(t)}) &= f(\lambda^{(t)}) + \mathbf{g}^{(t)T}\tilde{\lambda} - \mathbf{g}^{(t)T}\lambda^{(t)} \\ &= f(\lambda^{(t)}) + \mathbf{g}^{(t)T}\tilde{\lambda} - f(\lambda^{(t)}) \\ &= -\sum_{i=1}^{M} \tilde{\lambda}_i \min\{k, p_i\} + \mathrm{Tr}(U_{\lambda^{(t)}}^T (\sum_{i=1}^{M} \tilde{\lambda}_i X_i X_i^T) U_{\lambda^{(t)}}) \\ &\leq -\sum_{i=1}^{M} \tilde{\lambda}_i \min\{k, p_i\} + \max_{U^T U = I} \mathrm{Tr}(U^T (\sum_{i=1}^{M} \tilde{\lambda}_i X_i X_i^T) U) \\ &= f(\tilde{\lambda}), \end{aligned} \tag{17}$$

and thus $\mathbf{g}^{(t)} \in \partial f(\lambda^{(t)})$. Additionally, it can be verified that this subgradient matches the general description provided by [28, Thm. 3.9] with the associated affine shift.

4.1. Convergence

The subgradient $\mathbf{g}^{(t)}$ can be used to update $\lambda^{(t)}$ via the iteration in (15). The subgradient method is not a descent method, so the value of the objective function at step $t + 1$ may be larger than it was at step t. Thus we keep track of the dual variable with the lowest cost at each iteration and denote it

$$\lambda_{\text{best}}^{(t+1)} = \begin{cases} \lambda_{\text{best}}^{(t)} & f(\lambda^{(t+1)}) > f(\lambda_{\text{best}}^{(t)}); \\ \lambda^{(t+1)} & \text{otherwise.} \end{cases} \tag{18}$$

Given an upper bound on the norm of the subgradients, $\|g^{(t)}\|_2 \leq G < \infty$ for all t, classical theory makes different guarantees on the convergence of the sequence of iterates, $\{\lambda^{(t)}\}_{t=1}^{\infty}$, and thus on the sequence of objective function values, $\{f(\lambda_{\text{best}}^{(t)})\}_{t=1}^{\infty}$, depending on the choice of step size, $\alpha^{(t)}$. For example, with step sizes independent of iteration like $\alpha^{(t)} = a$ or $\alpha^{(t)} = a/\|\mathbf{g}^{(t)}\|_2$ for some $a > 0$, the subgradient algorithm will converge respectively to within $G^2 a/2$ or $Ga/2$ of the optimal value [30]. Alternatively, if the step size converges to zero and the sequence is nonsummable or square-summable, that is, $\lim_{t\to\infty} \alpha^{(t)} = 0$ and

$$\sum_{t=1}^{\infty} \alpha^{(t)} = \infty \quad \text{or} \quad \sum_{t=1}^{\infty} (\alpha^{(t)})^2 < \infty, \tag{19}$$

the subgradient method converges to an optimal objective value, $\lim_{t\to\infty} f(\lambda_{\text{best}}^{(t)}) = f(\lambda^*)$. These conditions are satisfied by step sizes like, $\alpha^{(t)} = a/\sqrt{t}$ for $a > 0$, or $\alpha^{(t)} = a/(b+t)$ where $a > 0$ and $b \geq 0$. Proofs of these results can be found in standard literature on convex optimization for nonsmooth problems such as [30, 29, 31].

Although the theory requires $\alpha^{(t)}$ to satisfy the constraints in (19) for convergence, the small step size leads to very slow convergence. In practice we can find an approximate solution quickly by stepping in the direction of a subgradient but requiring the dual objective to decrease at each iteration. Algorithm 1 (in Appendix A) solves Problem (14) by performing a back-tracking line search in the direction of $\mathbf{g}^{(t)} \in \partial f(\boldsymbol{\lambda}^{(t)})$ to ensure that the dual objective decreases at each step, however, this method is not guaranteed to converge because $\mathbf{g}^{(t)}$ is not necessarily a descent direction. The practical implementation of Algorithm 1 is a hybrid of a back-tracking line search and a nonsummable diminishing step size and for a fixed dimension k it identifies a stationary point of the dual problem while providing a feasible solution to the primal problem. It is not intended to be a state-of-the-art subgradient algorithm, but rather just one example of an implementation that is faster than the standard $a/(b+t)$ square-summable step size. Alternatively, a well-established quasi-Newton method like the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [32] can be used to solve Equation (14), but empirically the convergence rates are comparable to those of the algorithm presented here for this problem.

4.2. Optimality

In addition to theoretical convergence guarantees, the optimality of a solution to the dual subgradient approach can be verified in some cases. Let $\boldsymbol{\lambda}^*$ be a solution to Problem (14). There exists a matrix $U_{\boldsymbol{\lambda}^*}$ whose columns are the k dominant eigenvectors of $\sum_{i=1}^{M} \lambda_i^* X_i X_i^T$, analogous to Equation (12). Then $U_{\boldsymbol{\lambda}^*}$ satisfies $U_{\boldsymbol{\lambda}^*}^T U_{\boldsymbol{\lambda}^*} = I$ and is thus a feasible solution to the primal problem in (7). If the primal and dual objective functions are equal, strong duality holds and implies that $\boldsymbol{\lambda}^*$ and $\mathbf{U}^* = \mathrm{col}(U_{\boldsymbol{\lambda}^*})$ are globally optimal dual and primal variables, respectively. Empirically the duality gap approaches zero for collections of data that satisfy an implicit assumption of minimax optimization; that the data collection is free of outliers. Even when strong duality does not hold, the duality gap gives a bound on the maximum possible improvement for a solution.

This verification of optimality is standard for problems where the primal and dual costs are both computable, but existing techniques for finding the GMEB do not offer this feature. For instance, using a primal method like [3] does not directly provide a solution to the dual problem, and thus the duality gap is unknown. Section 7.1 contains numerical experiments that demonstrate the accuracy of the proposed subgradient method.

5. Proposed order selection rule

Given a dimension, k, and a finite collection of subspaces, $\mathcal{D} = \{\mathbf{X}_i \in \mathrm{Gr}(p_i, n)\}_{i=1}^M$, there exist subspaces, $\mathbf{U}^*(k)$, that solve

$$\operatorname*{arg\,min}_{\mathbf{U} \in \mathrm{Gr}(k,n)} \max_{i=1,\dots,M} d_{\mathrm{Gr}(k,n)}(\mathbf{U}, \mathbf{X}_i), \tag{20}$$

for $k = 1, \dots \max_i\{\dim(\mathbf{X}_i)\}$. The argument k is now included in the notation for the GMEB center to emphasize that the subspace depends on the parameter k, and may differ significantly depending on the value of this parameter. Section 4 described a method to compute $\mathbf{U}^*(k)$ from the associated dual variable, $\boldsymbol{\lambda}^*(k) \in \mathbb{R}^M$. However, because $\mathcal{D}$ contains subspaces of differing dimension, it is unclear on which Grassmannian the minimum enclosing ball should be computed. Thus, given the set $\mathcal{D}$, in this section we would like to determine the optimal choice for k, in addition to the associated center $\mathbf{U}^*(k)$. Please note a change in notation; the costs associated with a particular order, k, are more intuitive when the primal is formulated as a minimization problem and the dual is a maximization. Therefore, as shown in Equation (20), the primal minimization formulation is used for the remainder of the manuscript. The prior formulation was only used for ease of notation in the subgradient method.

All orthogonally invariant distances on $\mathrm{Gr}(k, n)$ can be written as a function of the k principle angles between a pair of points. It should be clear from the definition in Equation (1) that each angle is bounded above by $\pi/2$, and thus that the squared chordal distance is bounded above by k. Scaling the primal objective function by $1/k$ normalizes the cost associated with $\mathbf{U}^*(k)$ so that the value of

$$c_{\text{obj}}(k) := \begin{cases} 0 & k = 0; \\ \max\limits_{i=1,\ldots,M} \dfrac{d_{\mathrm{Gr}(k,n)}(\mathbf{U}^*(k), \mathbf{X}_i)}{k} & k = 1, \ldots \max\limits_i \{\dim(\mathbf{X}_i)\}, \end{cases} \tag{21}$$

gives a fair comparison across different values of k. The normalized objective function achieves its maximum value, $c_{\text{obj}}(k) = 1$, when there exists an i such that $\mathbf{X}_i \perp \mathbf{U}^*(k)$. That is, $\mathbf{U}^*(k)$ contains no information about at least one of the points in $\mathcal{D}$. At the other extreme, the minimum occurs when $k = 0$, and when the point of each $\Omega(\mathbf{X}_i)$ closest to the center coincides with the center. That is, $c_{\text{obj}}(k) = 0$ when $\mathbf{Y}_i^*(k) = \mathbf{U}^*(k)$ for all i, where $\mathbf{Y}_i^*(k) = \operatorname*{arg\,min}\limits_{\mathbf{Y}_i \in \Omega(\mathbf{X}_i)} d_{\mathrm{Gr}(k,n)}(\mathbf{U}^*(k), \mathbf{Y}_i)$.

Simply minimizing $c_{\text{obj}}(k)$ with respect to k is not sufficient to identify the ideal dimension of $\mathbf{U}^*(k)$ because on average $c_{\text{obj}}(k) \leq c_{\text{obj}}(k+1)$ irrespective of the relationship between the data points, and of course $c_{\text{obj}}(0) = 0$ by definition. However, the dimension of the ideal center should represent all the common information without over-fitting, and should also indicate when no significant relationship exists between the data. Thus we propose a penalty term based on the dimensions of the data not represented by $\mathbf{U}^*(k)$ that balances the information lost by making k too small with the lack of specificity that comes from setting k too large.

Let $\mathbf{U}^{*\perp}(k)$ denote the orthogonal complement of $\mathbf{U}^*(k)$ and $\tilde{p}_j \doteq \min\{n - k, \dim(\mathbf{X}_j)\}$ for $j = 1, \ldots, M$. The expression

$$c_{\text{pen}}(k) := \begin{cases} 1 & k = 0; \\ \min\limits_{j=1,\ldots,M} 1 - \dfrac{d_{\mathrm{Gr}(\tilde{p}_j,n)}(\mathbf{U}^{*\perp}(k), \mathbf{X}_j)}{\tilde{p}_j} & k = 1, \ldots \max\limits_j \{\dim(\mathbf{X}_j)\}, \end{cases} \tag{22}$$

represents the minimum similarity between any point in $\mathcal{D}$ and the dimensions not contained in the center of the GMEB. A high minimum similarity between points in $\mathcal{D}$ and $\mathbf{U}^{*\perp}(k)$ implies that too much information is being left out of the central subspace, $\mathbf{U}^*(k)$. The penalty term takes a value of $c_{\text{pen}}(k) = 1$ when $\dim(\mathbf{U}^{*\perp}(k) \cap \mathbf{X}_j) = \tilde{p}_j$ for all j and $c_{\text{pen}}(k) = 0$ when there exists a j for which $\mathbf{X}_j \perp \mathbf{U}^{*\perp}(k)$. The sum of the terms in (21) and (22) leads to the proposed rule for selecting the optimal order k^*,

$$\operatorname*{arg\,min}_{k=0,\ldots,\max_i\{\dim(\mathbf{X}_i)\}} c_{\text{obj}}(k) + c_{\text{pen}}(k). \tag{23}$$

The two terms in (23) are computed independently so the GMEB center is not affected by the penalty term. The value of k^* that minimizes the sum of these two terms corresponds to the number of subspace dimensions needed to represent the common information present in $\mathcal{D}$ without over-fitting. Numerical experiments in Section 7.3 demonstrate the efficacy of the order selection rule on simulated data with ground truth.

5.1. Primal solutions are not nested in general for increasing values of k

Naively, the order selection rule in Equation (23) can be applied by computing the costs $c_{\text{obj}}(k)$ and $c_{\text{pen}}(k)$ independently for $k = 0, \ldots, \max_i\{\dim(\mathbf{X}_i)\}$ as follows,

1. Compute $\boldsymbol{\lambda}^*(k)$ using the subgradient method described in Section 4.
2. Find the associated primal variable, $\mathbf{U}^*(k)$, as the k-dimensional eigenspace of the weighted sum $\sum_{i=1}^{M} \lambda_i^*(k) X_i X_i^T$.
3. Compute the orthogonal complement, $\mathbf{U}^{*\perp}(k) = \text{col}\left(I - U^*(k)U^{*T}(k)\right)$.

Then k^* is selected as the value of k associated with the minimum cost, $c_{\text{obj}}(k) + c_{\text{pen}}(k)$. If $\boldsymbol{\lambda}^*(k) = \boldsymbol{\lambda}^*(k+1)$ for some $k < \max_i\{\dim(\mathbf{X}_i)\}$ then the solution on $\text{Gr}(k+1, n)$ can be constructed in a greedy fashion as the direct sum of the solution on $\text{Gr}(k, n)$ and the $(k+1)$st eigenvector of $\sum_{i=1}^{M} \lambda_i^*(k) X_i X_i^T$. Unfortunately, the dual variables are not generally equal for increasing values of k, so a greedy approach is not appropriate.

Observe that the central subspaces are not nested for increasing dimensions in the following illustrative example. Let

$$X_1 = \begin{bmatrix} \frac{\sqrt{2}}{\sqrt{3}} & 0 \\ \frac{1}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{6}} & 0 \\ 0 & \frac{\sqrt{7}}{\sqrt{8}} \\ 0 & \frac{1}{\sqrt{8}} \end{bmatrix}, \quad X_2 = \begin{bmatrix} \frac{1}{\sqrt{6}} & 0 \\ \frac{\sqrt{2}}{\sqrt{3}} & 0 \\ \frac{1}{\sqrt{6}} & 0 \\ 0 & \frac{1}{\sqrt{8}} \\ 0 & \frac{\sqrt{7}}{\sqrt{8}} \end{bmatrix}, \quad \text{and } X_3 = \begin{bmatrix} \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \\ \frac{\sqrt{2}}{\sqrt{3}} \\ 0 \\ 0 \end{bmatrix}, \tag{24}$$

be orthonormal bases for the three points $\mathbf{X}_1, \mathbf{X}_2 \in \text{Gr}(2,5)$ and $\mathbf{X}_3 \in \text{Gr}(1,5)$. One can check that the subspace that minimizes the maximum distance to these three points on $\text{Gr}(1,5)$ is the mean of their first columns. That is, the optimal primal and dual variables are

$$\mathbf{U}^*(1) = \text{col}\left(\begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & 0 & 0 \end{bmatrix}^T\right), \quad \text{and } \boldsymbol{\lambda}^*(1) = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix}^T, \tag{25}$$

with associated primal and dual costs of

$$\min_{\mathbf{U}\in\text{Gr}(1,5)} \max_{i=1,2,3} d_{\text{Gr}(1,5)}(\mathbf{U}, \mathbf{X}_i) = \max_{\lambda\in\mathbb{R}^3} \min_{U^TU=I} 1 - \sum_{i=1}^{3} \lambda_i \text{Tr}(U^T Y_i Y_i^T U) = \frac{1}{9}. \tag{26}$$

The duality gap in Equation (26) is zero, indicating that this is a global solution.

On $\text{Gr}(2,5)$, however, $\Omega_+(\mathbf{X}_3)$ consists of subspaces that span X_3 and any orthogonal direction. In particular there exists $\mathbf{Y}_3 \in \Omega_+(\mathbf{X}_3)$ such that the second column of Y_3 is $[0\ 0\ 0\ {}^1/\sqrt{2}\ {}^1/\sqrt{2}]^T$. This leads to a solution for the center of the minimum enclosing ball on $\text{Gr}(2,5)$ given by primal and dual variables

$$\mathbf{U}^*(2) = \text{col}\left(\begin{bmatrix} \frac{3}{\sqrt{22}} & \frac{3}{\sqrt{22}} & \frac{2}{\sqrt{22}} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}^T\right), \quad \text{and } \boldsymbol{\lambda}^*(2) = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}^T. \tag{27}$$

Notably, $\mathbf{X}_3$ is not in the support of the minimum enclosing ball on $\text{Gr}(2,5)$ and thus does not influence the central subspace. Strong duality also holds for this solution with

$$\min_{\mathbf{U}\in\mathrm{Gr}(2,5)} \max_{i=1,2,3} d_{\mathrm{Gr}(2,5)}(\mathbf{U},\mathbf{X}_i) = \max_{\lambda\in\mathbb{R}^3} \min_{U^TU=I} 2 - \sum_{i=1}^{3} \lambda_i \mathrm{Tr}(U^T Y_i Y_i^T U) = \frac{14-3\sqrt{7}}{24}. \tag{28}$$

Since $\mathbf{U}^*(1)$ is orthogonal to the second dimension of $\mathbf{U}^*(2)$ and noncollinear with the first, and the columns of $U^*(2)$ are orthogonal, we have $\mathbf{U}^*(1) \not\subset \mathbf{U}^*(2)$. Additionally we find that the optimal order selected by applying the rule in Equation (23) is $k^* = 1$, because

$$\begin{aligned}
c_{\mathrm{obj}}(0) + c_{\mathrm{pen}}(0) &= 0 + 1 = 1, \\
c_{\mathrm{obj}}(1) + c_{\mathrm{pen}}(1) &= \frac{1}{1}\left(\frac{1}{9}\right) + \frac{1}{1}\left(1 - \left(\frac{\sqrt{8}}{\sqrt{9}}\right)^2\right) \approx 0.22, \quad \text{and} \\
c_{\mathrm{obj}}(2) + c_{\mathrm{pen}}(2) &= \frac{1}{2}\left(\frac{14-3\sqrt{7}}{24}\right) + \frac{1}{2}\left(2 - \left(\frac{-1}{\sqrt{12}}\right)^2 + \left(\frac{1-\sqrt{7}}{\sqrt{16}}\right)^2\right) \approx 0.25.
\end{aligned} \tag{29}$$

This agrees with the intuition that the center of the minimum enclosing ball represents the common information in all points without over-fitting to any subset of points, but note that the optimal order is not always the dimension of the smallest subspace. The common subspace may have dimension smaller than any of the samples or there may be no common subspace.

Even though the primal solutions are not always nested, a good initial guess for the dual variable will reduce computational overhead. One benefit of the subgradient approach is that $\boldsymbol{\lambda}^*(k)$ is computed explicitly. Thus we can initialize the algorithm with $\boldsymbol{\lambda}^{(0)}(k+1) = \boldsymbol{\lambda}^*(k)$. The impact of this heuristic warm-start is discussed in the experiments in Section 7.2.

5.2. Related literature on order fitting for subspace averaging

A recent work from Santamaría *et al.* [22] also attempts to find a central subspace of ambiguous dimension. The authors minimize the mean-squared error (MSE) between a subspace and a collection of data in the space of $n \times n$ projection matrices using the squared Frobenius norm. That is,

$$E(k) = \min_{\mathbf{U}\in\mathrm{Gr}(k,n)} \frac{1}{M} \sum_{i=1}^{M} \|UU^T - X_i X_i^T\|_F^2. \tag{30}$$

Putting aside for a moment that the current work is interested in minimizing the maximum deviation rather than the mean-squared error, there remains a central difference between the technique in [22] and the proposed method. The optimization of Equation (30) is done in a vector space, after which the solution is mapped to the nearest point on the Grassmann manifold. This is subtly different than minimizing the MSE on the Grassmannian with respect to the squared chordal distance using the point-to-set interpretation of [24]. To see this, write half of the squared distance from [22] between the central subspace and the ith point as

$$\begin{aligned}\frac{1}{2}\|U^*(k)U^{*T}(k) - X_iX_i^T\|_F^2 &= \frac{k+p_i}{2} - \sum_{r=1}^{\min\{k,p_i\}} \cos^2(\theta_r(\mathbf{U}^*(k),\mathbf{X}_i)) \\ &= \frac{|k-p_i|}{2} + \sum_{r=1}^{\min\{k,p_i\}} \sin^2(\theta_r(\mathbf{U}^*(k),\mathbf{X}_i)).\end{aligned} \tag{31}$$

In contrast, the point-to-set squared chordal distance on $\mathrm{Gr}(k,n)$ is

$$\begin{aligned}d_{\mathrm{Gr}(k,n)}(\mathbf{U}^*(k),\mathbf{X}_i) &= \min\left\{d(\mathbf{U}^*(k),\mathbf{Y}_i) : \mathbf{Y}_i \in \Omega(\mathbf{X}_i)\right\} \\ &= \min\left\{\frac{1}{2}\|U^*(k)U^{*T}(k) - Y_iY_i^T\|_F^2 : \mathbf{Y}_i \in \Omega(\mathbf{X}_i)\right\} \\ &= k - \sum_{r=1}^{k} \cos^2(\theta_r(\mathbf{U}^*(k),\mathbf{Y}_i)) \\ &= \sum_{r=1}^{\min\{k,p_i\}} \sin^2(\theta_r(\mathbf{U}^*(k),\mathbf{X}_i))\end{aligned} \tag{32}$$

because $0 = \theta_{p_i}(\mathbf{U}^*(k),\mathbf{Y}_i) = \theta_{p_i+1}(\mathbf{U}^*(k),\mathbf{Y}_i) = \cdots = \theta_k(\mathbf{U}^*(k),\mathbf{Y}_i)$ if $p_i < k$ by the definition of $\mathbf{Y}_i$ in Equation (4). Thus the distances differ by $\frac{|k-p_i|}{2}$, which is the difference in dimensions between the central subspace and the ith data point.

The slight difference in distance measurements lends itself to an interesting interpretation when determining the appropriate rank of the central subspace. The solution, $\mathbf{U}^*(k)$, to

$$\underset{\mathbf{U}\in\mathrm{Gr}(k,n)}{\arg\min} \frac{1}{M}\sum_{i=1}^{M} \|UU^T - X_iX_i^T\|_F^2 \tag{33}$$

for a fixed k is the dominant k-dimensional eigenspace of the sum $\frac{1}{M}\sum_{i=1}^{M} X_iX_i^T$. That is, if

$$\frac{1}{M}\sum_{i=1}^{M} X_iX_i^T = FDF^T \tag{34}$$

is an eigendecomposition with eigenvectors $F = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_R]$ and associated eigenvalues $d_1 \geq d_2 \geq \cdots \geq d_R$, then the solution to Equation (33) is $\mathbf{U}^*(k) = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_k]$. Note that this $\mathbf{U}^*(k)$ is not the same subspace as the center of the minimum enclosing ball. The MSE in Equation (30) can be written as a function of all R eigenvalues,

$$E(k) = \sum_{r=1}^{k} 1 - d_r + \sum_{r=k+1}^{R} d_r, \tag{35}$$

and the minimum of Equation (35) is achieved when k^* is the smallest value for which $d_{k+1} < 0.5$. This eigenvalue threshold is then fixed regardless of the dimension of the ambient space, and as we will see in Section 7.3, the selected dimension could differ drastically for noisy data depending on the ambient dimension.

For a different interpretation of the k^* that minimizes Equation (30) we can rewrite Equation (35) as a function of the angles between each eigenvector and the subspaces,

$$
\begin{aligned}
E(k) &= \sum_{r=1}^{k} 1 - \mathbf{f}_r^T (\frac{1}{M}\sum_{i=1}^{M} X_i X_i^T)\mathbf{f}_r + \sum_{r=k+1}^{R} \mathbf{f}_r^T (\frac{1}{M}\sum_{i=1}^{M} X_i X_i^T)\mathbf{f}_r && (36)\\
&= \sum_{r=1}^{k} 1 - \frac{1}{M}\sum_{i=1}^{M} \cos^2(\theta(\mathbf{f}_r, \mathbf{X}_i)) + \sum_{r=k+1}^{R} \frac{1}{M}\sum_{i=1}^{M} \cos^2(\theta(\mathbf{f}_r, \mathbf{X}_i)) && (37)\\
&= \sum_{r=1}^{k} \frac{1}{M}\sum_{i=1}^{M} \sin^2(\theta(\mathbf{f}_r, \mathbf{X}_i)) + \sum_{r=k+1}^{R} \frac{1}{M}\sum_{i=1}^{M} \sin^2(\frac{\pi}{2} - \theta(\mathbf{f}_r, \mathbf{X}_i)) && (38)\\
&= \sum_{r=1}^{k} \frac{1}{M}\sum_{i=1}^{M} \sin^2(\theta(\mathbf{f}_r, \mathbf{X}_i)) + \sum_{r=k+1}^{R} \frac{1}{M}\sum_{i=1}^{M} \sin^2(\theta(\mathbf{f}_r, \mathbf{X}_i^{\perp})) && (39)\\
&= \sum_{r=1}^{k} \frac{1}{M}\sum_{i=1}^{M} d_{\mathrm{Gr}(1,n)}(\mathbf{f}_r, \mathbf{X}_i) + \sum_{r=k+1}^{R} \frac{1}{M}\sum_{i=1}^{M} d_{\mathrm{Gr}(1,n)}(\mathbf{f}_r, \mathbf{X}_i^{\perp}). && (40)
\end{aligned}
$$

The equality between (38) and (39) is due to [33, Thm. 2.7] which implies that $\frac{\pi}{2} - \theta(\mathbf{f}_r, \mathbf{X}_i) = \theta(\mathbf{f}_r, \mathbf{X}_i^{\perp})$. Note, however, that Equation (40) is *not* equivalent to

$$
\frac{1}{M}\sum_{i=1}^{M} d_{\mathrm{Gr}(k,n)}(\mathbf{U}^*(k), \mathbf{X}_i) + \frac{1}{M}\sum_{i=1}^{M} d_{\mathrm{Gr}(R-k,n)}(\mathbf{U}^{*\perp}(k), \mathbf{X}_i^{\perp}) \qquad (41)
$$

because linear combinations of the eigenvectors, $\mathbf{f}_r$, are not included in the expression. A new interpretation of the MSE-minimizing k becomes fairly apparent in light of Equation (40). The optimal k^* is the one that minimizes the mean-squared chordal distance between $\{\mathbf{f}_1, \ldots, \mathbf{f}_k\}$ and the data points, plus the mean-squared chordal distance between $\{\mathbf{f}_{k+1}, \ldots, \mathbf{f}_R\}$ and the orthogonal complements of the data points.

5.3. Hybrid rule

It is possible to create a hybrid of the order-selection rule of [22] and the proposed method with a slight modification. In [34], a robustification of the technique in [22] is proposed that leads to a weighted eigenvalue decomposition at optimality. The weights are determined using a variety of robust objective functions via a majorization-minimization scheme, which results in a down-weighting of outliers in the data. By minimizing the mean-squared error of the *weighted* average (similar to Equation (30)), this amounts to a hard eigenvalue threshold with the order chosen to be the number of dimensions with eigenvalues greater than 0.5.

For the hybrid method, weights will come from the values of the dual variable, $\boldsymbol{\lambda}^*(k)$, at optimality. Since these values depend on the parameter k, the hard eigenvalue threshold is not applicable. Let $d_1(k) \geq d_2(k) \geq \cdots \geq d_R(k)$ be the eigenvalues of $\sum_{i=1}^{M} \lambda_i^*(k) X_i X_i^T$ where $\boldsymbol{\lambda}^*(k)$ is the vector of optimal dual variables computed for the GMEB on $\mathrm{Gr}(k, n)$ using the proposed algorithm. For $k = 0$, let $\lambda_i^*(0) = \frac{1}{M}$ for $i = 1, \ldots, M$. We define a modified version of the MSE from Equation (35) as

$$\tilde{E}(k) = \sum_{r=1}^{k} 1 - d_r(k) + \sum_{r=k+1}^{R} d_r(k). \tag{42}$$

The order-selection rule of [22] applied to the GMEB center is then

$$\operatorname*{arg\,min}_{k=0,\ldots,\max_i\{\dim(\mathbf{X}_i)\}} \quad \tilde{E}(k). \tag{43}$$

It should be clear that the eigenvalues $\{d_r(k)\}_{r=1}^{R}$ will be different for different values of $\boldsymbol{\lambda}^*(k)$. In the experiments of Section 7.3, this combined method is referred to as “Hybrid” and performs favorably for all tests; out-performing the other techniques in 2 out of 3 scenarios.

6. Synthetic data generation

The numerical experiments in Section 7 require data for which the ground truth is known, and ideally data for which the center of the GMEB is distinct from the other generalized Grassmannian means. Thus, in this section we propose two different models for sampling points nonuniformly from a unit ball on the Grassmannian. The first is an asymmetrical nested ball structure, and the second samples more densely within a randomly selected arc of the boundary of a unit ball.

6.1. Asymmetrical nested ball model

A collection of subspaces, $\mathcal{D} = \{\mathbf{X}_i\}_{i=1}^{M}$, are uniformly sampled from two balls, $\mathcal{B}_{\epsilon_2}(\mathbf{Z}_2) \subset \mathcal{B}_{\epsilon_1}(\mathbf{Z}_1) \subset \mathrm{Gr}(k_0, n)$ with centers at $\mathbf{Z}_1$, $\mathbf{Z}_2$ and corresponding radii $\epsilon_1 > \epsilon_2$, respectively. The larger ball, $\mathcal{B}_{\epsilon_1}(\mathbf{Z}_1)$, is the minimum enclosing ball of the data so that $\mathbf{U}^*(k_0) = \mathbf{Z}_1$. The smaller ball is fully contained within the larger ball, i.e., $\mathcal{B}_{\epsilon_2}(\mathbf{Z}_2) \subset \mathcal{B}_{\epsilon_1}(\mathbf{Z}_1)$, but $\mathbf{Z}_1 \notin \mathcal{B}_{\epsilon_2}(\mathbf{Z}_2)$. Let M_1, M_2 be the number of points sampled from $\mathcal{B}_{\epsilon_1}(\mathbf{Z}_1), \mathcal{B}_{\epsilon_2}(\mathbf{Z}_2)$ respectively, with $M = M_1 + M_2$. When $M_2 = 0$, the generalized Grassmannian means are all equal to the point $\mathbf{Z}_1$. When more points are sampled from $\mathcal{B}_{\epsilon_2}(\mathbf{Z}_2)$ and the fraction M_2/M_1 grows, the generalized Grassmannian means for $p < \infty$ move away from $\mathbf{Z}_1$ in the direction of $\mathbf{Z}_2$, making the averages distinct without affecting the center of the GMEB. The radius of the large ball, ϵ_1, controls the similarity of the data points.

As described, the data points are all sampled from a single manifold, $\mathrm{Gr}(k_0, n)$. If ϵ_1 is small enough, then the optimal rank for the GMEB (or any of the generalized Grassmannian means) is $k^* = k_0$. This construction can be generalized in two ways.

1. For $i = 1, \ldots, M$, the basis for $\mathbf{X}_i$ can be completed to a p_i-dimensional subspace by taking the span of X_i and $p_i - k_0$ random dimensions. If the $p_i - k_0$ random dimensions are mutually orthogonal for $i = 1, \ldots, M$, then the optimal rank for the GMEB is still $k^* = k_0$.
2. Points from the large ball can be sampled from one manifold, $\mathcal{B}_{\epsilon_1}(\mathbf{Z}_1) \subset \mathrm{Gr}(k_1, n)$ while points from the small ball are sampled from another, $\mathcal{B}_{\epsilon_2}(\mathbf{Z}_2) \subset \mathrm{Gr}(k_2, n)$. If $k_1 \neq k_2$, the optimal rank of the central subspace is ambiguous. Experiments show that using the proposed order selection rule, $k^* = k_1$ independent of other parameters, but using the criteria of [22], k^* depends on ϵ_1 and M_2/M_1.

As an illustrative example, Figure 2 shows 2-dimensional embeddings via multidimensional scaling of data sets on $\mathrm{Gr}(1, 3)$ that have been generated according to the asymmetrical nested

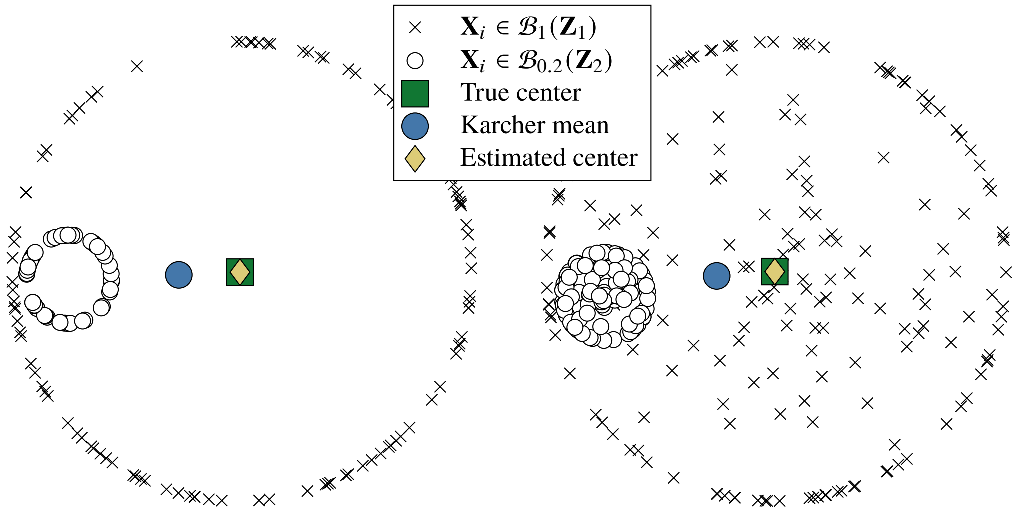


Figure 2: Two examples of point sets from $\mathrm{Gr}(1,3)$ generated using the nested ball model embedded into $\mathbb{R}^2$ by multidimensional scaling. The points from $\mathcal{B}_1(\mathbf{Z}_1)$ are indicated with x's, points from $\mathcal{B}_{0.2}(\mathbf{Z}_2)$ are marked with white circles, the true center is the green square, the Karcher mean is the blue circle, and the estimated GMEB center is the yellow diamond.

ball model. The yellow diamond indicates the center of the GMEB (computed via the proposed method) and the blue circle marks the Karcher mean of each data collection.

6.2. Unit ball with higher sampling density from a random arc

Another practical scenario where the GMEB center may differ from other generalized Grassmannian means is when data has been sampled unevenly. This setting is simulated by selecting a random arc from the boundary of a unit ball and sampling additional points from that region. A collection of subspaces, $\mathcal{D} = \{\mathbf{X}_i\}_{i=1}^{M}$, are uniformly sampled from the ball $\mathcal{B}_{\epsilon_1}(\mathbf{Z}_1) \subset \mathrm{Gr}(k_0, n)$ with center at $\mathbf{Z}_1$ and radius ϵ_1. M_1 points are sampled from $\mathcal{B}_{\epsilon_1}(\mathbf{Z}_1)$ so that $\mathbf{U}^*(k_0) = \mathbf{Z}_1$. Two points are randomly selected from the boundary of $\mathcal{B}_{\epsilon_1}(\mathbf{Z}_1)$, and M_2 additional points are uniformly sampled from the arc connecting them on the boundary to create $M = M_1 + M_2$ samples. The data points are all sampled from a single manifold, $\mathrm{Gr}(k_0, n)$, and for sufficiently small ϵ_1, the optimal rank for the GMEB (or any of the generalized Grassmannian means) is $k^* = k_0$. To generalize this construction, additional dimensions can be included to create points from a disjoint union of Grassmannians.

For $i = 1, \ldots, M$, the basis for $\mathbf{X}_i$ can be completed to a p_i dimensional subspace by taking the span of X_i and $p_i - k_0$ random dimensions. If the $p_i - k_0$ random dimensions are mutually orthogonal for $i = 1, \ldots, M$, then the optimal rank for the GMEB is still $k^* = k_0$. Figure 3 shows 2-dimensional embeddings via multidimensional scaling of data sets on $\mathrm{Gr}(1,3)$ that have been generated as a unit ball with higher sampling density along a random arc. The yellow diamond indicates the center of the GMEB (computed via the proposed method) and the blue circle marks the Karcher mean of each data collection.

It should be noted that using either data model the point at the center of $\mathcal{B}_{\epsilon_1}(\mathbf{Z}_1)$ is only the ground-truth center of the minimum enclosing ball of the data collection, $\mathbf{U}(k^*)$, if the points have been sampled with a high enough density from the surface of the ball. The minimum

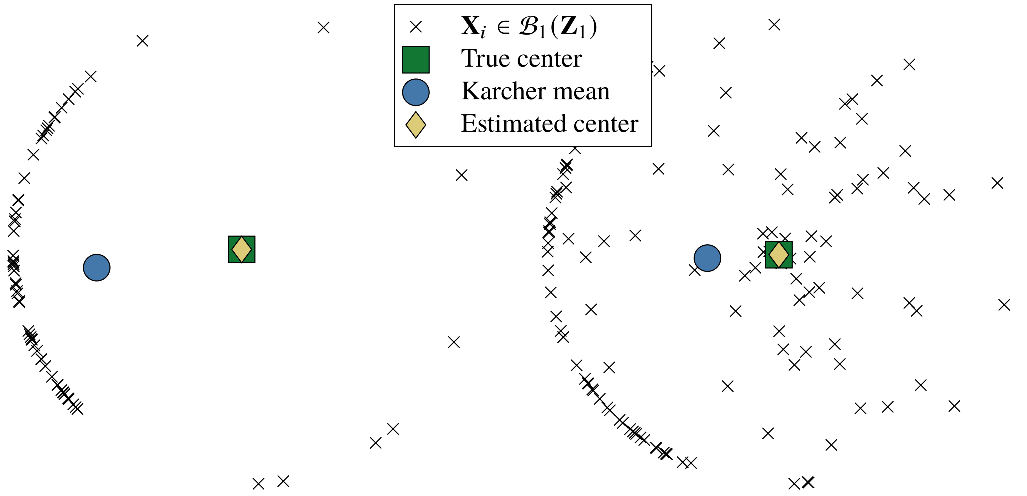


Figure 3: Two examples of point sets from $\mathrm{Gr}(1,3)$ on the unit ball, $\mathcal{B}_1(\mathbf{Z}_1)$, sampled with nonuniform density on the boundary, embedded into $\mathbb{R}^2$ by multidimensional scaling. Points from $\mathcal{B}_1(\mathbf{Z}_1)$ are indicated with x's, the true center is the green square, the Karcher mean is the blue circle, and the estimated GMEB center is the yellow diamond.

number of uniformly distributed points needed grows with the ambient dimension, n, so in high dimensional spaces the number of points, M, needed to create a ground-truth center may become prohibitively large. The experimental data can be generated exclusively from the boundary of the balls or interior points can be added.[1]

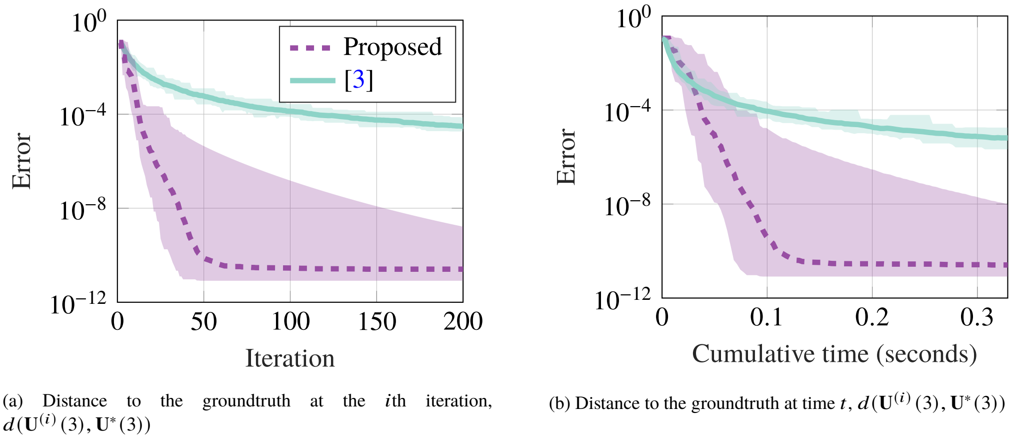
7. Numerical experiments

The experiments in this section are meant to illustrate three properties of the proposed GMEB algorithm and associated order-selection rule. First, we demonstrate the speed and accuracy of the proposed method for estimating the center of the GMEB. Second, we demonstrate that a warm-start on $\mathrm{Gr}(k+1,n)$ using the optimal solution from $\mathrm{Gr}(k,n)$ can reduce the number of iterations required for the algorithm to converge. And finally, we compare results of the proposed order-selection rule and the rule of [22] in a variety of scenarios to gain intuition about when and how they differ.

7.1. Experiment 1: Accuracy of the GMEB

To test the accuracy and efficiency of the proposed dual subgradient approach, data sets are generated according to the each of two data models from Section 6. For each data collection, the GMEB center is approximated using the proposed method and the algorithm of Renard *et al.* [3], and the residual error is measured as the distance between the approximate centers and the true centers. For the first data set, $M = 100$ points are sampled from $\mathrm{Gr}(3,10)$ using the asymmetrical nested ball model in Section 6.1 with neither of the proposed generalizations. That is, $k_0 = k_1 = k_2 = 3$ so that all points are sampled from the same Grassmann manifold. $M_1 = 70$

[1] Matlab code for the algorithms, data generation procedures, and numerical experiments in this manuscript is available at https://sites.google.com/site/nicolasgillis/code.



(a) Distance to the groundtruth at the ith iteration, $d(\mathbf{U}^{(i)}(3), \mathbf{U}^*(3))$

(b) Distance to the groundtruth at time t, $d(\mathbf{U}^{(i)}(3), \mathbf{U}^*(3))$

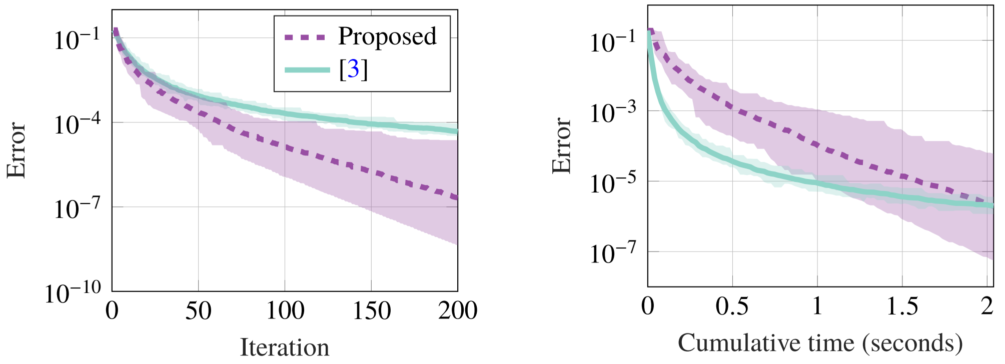
Figure 4: Median distance to the groundtruth and cumulative time for the GMEB on $\mathrm{Gr}(3, 10)$ of data generated with the asymmetrical nested ball model from Section 6.1 over 100 Monte Carlo trials. The data consists of 100 points in $\mathrm{Gr}(3, 10)$. The proposed method is indicated by the dashed purple line and the method of Renard *et al*. [3] is represented by the solid turquoise line. The shaded regions span the extreme values.

of the points come from the boundary of $\mathcal{B}_1(\mathbf{Z}_1)$ and $M_2 = 30$ from the boundary of $\mathcal{B}_{0.125}(\mathbf{Z}_2)$. No points are sampled from the interior of either ball. Both algorithms are initialized using the extrinsic mean of the data [35, 9], that is, $\lambda^{(0)} = [1/100, 1/100, \dots, 1/100]^T$, and $\mathbf{U}^{(0)}(3)$ is the dominant 3-dimensional eigenspace of $\sum_{i=1}^{100} \lambda_i^{(0)} X_i X_i^T$. The groundtruth center is $\mathbf{U}^*(3) = \mathbf{Z}_1$.

Figure 4a shows the median distance to the groundtruth over 100 Monte Carlo trials between the iterate with the lowest primal cost and the ground-truth center. Figure 4b shows the same median distance to the groundtruth relative to cumulative computation time for each algorithm. In both plots the proposed method is indicated by the dashed purple line and the method of [3] is represented by the solid turquoise line. The shaded regions denote the complete range of values across all trials. This is a setting in which all data points live on a single Grassmann manifold. Therefore the point-to-set distances reduce to the traditional Grassmannian distances and the technique of [3] is equivalent to that of [1].

The proposed method clearly outperforms the existing technique in terms of accuracy relative to both iterations and computation time for this collection of data. However, the cumulative computation time is affected by many of the parameters in the experimental setup. Let $P = \max_i\{\dim(\mathbf{X}_i)\}$. For the technique of [3, 1], the per iteration complexity is $O\left(MP(nk + k^2)\right)$ due to the M matrix products and subsequent thin SVDs. The proposed method computes these same M products and SVDs, but must additionally compute the compact SVD of a matrix of size $n \times MP$ in order to get the updated center. Assuming that $n \le MP$ (as it is in all the experiments), the complexity of the proposed algorithm is then $O\left(MP(nk + k^2 + n^2)\right)$. There are an additional M SVDs for each back-tracking step taken, but those steps are infrequent and thus dominated by the other terms. From these complexities we can see that an increase in the ambient dimension, n, number of subspaces, M, or subspace dimension, P, would all lead to a relative decrease in the efficiency of the proposed method.

In the second example we employ the data model from Section 6.2, with the inclusion of interior points and the generalization that the data points come from a disjoint union of Grassmannians, that is, they are subspaces of differing dimensions. Initially, $M_1 = 100$ points are



(a) Distance to the groundtruth at the ith iteration, $d(\mathbf{U}^{(i)}(3), \mathbf{U}^*(3))$

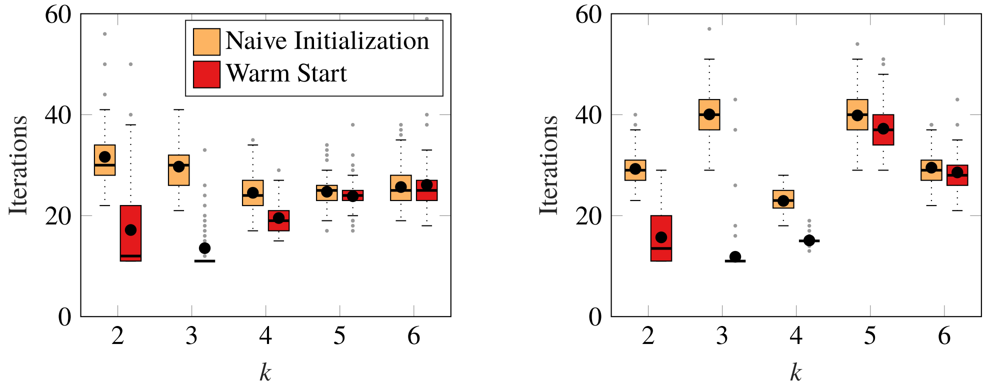
(b) Distance to the groundtruth at time t, $d(\mathbf{U}^{(i)}(3), \mathbf{U}^*(3))$

Figure 5: Median distance to the groundtruth and cumulative computation time for the GMEB on $\mathrm{Gr}(3, 15)$ of data generated with the nonuniform sampling model from Section 6.2 over 100 Monte Carlo trials. The data consists of 300 points in $\coprod_{p \in \mathcal{P}} \mathrm{Gr}(p, 15)$ for $\mathcal{P} = \{3, 4, 5, 6\}$. The proposed method is indicated by the dashed purple line and the method of Renard *et al.* [3] is represented by the solid turquoise line. The shaded regions span the extreme values.

sampled from the boundary of $\mathcal{B}_1(\mathbf{Z}_1)$ on $\mathrm{Gr}(3, 15)$. An additional $M_2 = 100$ points are selected from an arc on the boundary of the ball between two randomly selected points. Finally $M_3 = 100$ points are selected uniformly at random from the interior of the ball. Each of the $M = 300$ points is then completed to a basis for a p_i-dimensional subspace where p_i is randomly selected from the set $\mathcal{P} = \{3, 4, 5, 6\}$. Both algorithms are again initialized using the extrinsic mean of the data on $\mathrm{Gr}(3, 15)$ where $\boldsymbol{\lambda}^{(0)} = [1/300, 1/300, \dots, 1/300]^T$, and $\mathbf{U}^{(0)}(3)$ is the dominant 3-dimensional eigenspace of $\sum_{i=1}^{300} \lambda_i^{(0)} X_i X_i^T$. Figure 5a shows the median distance to the groundtruth over 100 Monte Carlo trials between the iterate with the lowest primal cost and the ground-truth center, while Figure 5b shows the median error relative to cumulative computation time. The proposed method is indicated by the dashed purple line and the method of Renard *et al.* [3] is represented by the solid turquoise line. The shaded regions span the extreme values. The groundtruth center is $\mathbf{U}^*(3) = \mathbf{Z}_1$.

As shown in Figure 5a, the proposed method achieves a higher accuracy in fewer iterations than [3]. However, the greater complexity of the proposed method means that the primal algorithm initially achieves a lower error, as shown in Figure 5b. The increased number of points in the data set and specifically in the support of the GMEB lead to a slower overall convergence for the proposed algorithm. This reduced efficiency would grow with the size of the data, however the subgradient technique consistently achieves lower overall error given enough time. Moreover, the proposed method provides duality-gap optimality guarantees.

One direction for future work is to combine the two methods to get the best of both worlds; fast initial estimates of the center and high accuracy solutions over time. Using $\mathbf{U}^{(t)}(k)$ computed via t iterations of [3] as an estimate of the center, we can find dual-feasible variables that are non-zero only for points in the support set of the enclosing ball centered at $\mathbf{U}^{(t)}(k)$. For example, let $\mathcal{I} = \{i : d_{\mathrm{Gr}(k,n)}(\mathbf{U}^{(t)}(k), \mathbf{X}_i) = \max_i d_{\mathrm{Gr}(k,n)}(\mathbf{U}^{(t)}(k), \mathbf{X}_i)\}$. Then let $\lambda_i^{(0)} = 1/|\mathcal{I}|$ for $i \in \mathcal{I}$ and $\lambda_i^{(0)} = 0$ otherwise, and proceed with the subgradient algorithm from this warm-start. An alternative initialization strategy is proposed in Section 7.2.



(a) Results from 100 trials with the asymmetrical nested ball model where $k^* = 4$ and $M = 50$ points sampled from $\mathrm{Gr}(p_i, 10)$ with $p_i \in \{4, 5, 6\}$.

(b) Results from 100 trials with the nonuniform sampling model where $k^* = 4$ and $M = 300$ points sampled from $\mathrm{Gr}(p_i, 10)$ with $p_i \in \{4, 5, 6\}$.

Figure 6: Number of iterations needed for the proposed algorithm to reach a stationary point using a naive initialization, $\boldsymbol{\lambda}^{(0)}(k+1) = [1/M, 1/M, \ldots, 1/M]^T$ (orange), and a warm start, $\boldsymbol{\lambda}^{(0)}(k+1) = \boldsymbol{\lambda}^*(k)$ (red) for two data sets.

7.2. Experiment 2: Faster convergence by initializing with previous solutions

To apply the order selection criteria in Section 5, the GMEB center must be computed for $k = 1, \ldots, \max_i\{\dim(\mathbf{X}_i)\}$. The example in Section 5.1 demonstrates that the subspace at the center of the minimum enclosing ball cannot be built in a greedy fashion, because the center $\mathbf{U}^*(k-1) \in \mathrm{Gr}(k-1, n)$ is not in general a subspace of the center $\mathbf{U}^*(k) \in \mathrm{Gr}(k, n)$. However, the solutions are often *nearly* nested. As a result, the vector, $\boldsymbol{\lambda}^*(k-1)$, that provides the optimal value of the dual objective function for the problem on $\mathrm{Gr}(k-1, n)$ can offer a good initialization for the dual subgradient algorithm used to find the GMEB center on $\mathrm{Gr}(k, n)$, significantly reducing the total computation time needed to identify the optimal dimension, k^*. In [36] the authors also used a warm-starting strategy on a similar problem to improve the efficiency of a rank-adaptive matrix optimization scheme. Their proposed method alternates between greedy rank increase and smooth Riemannian optimization on fixed-rank manifolds, and they show that the strategy significantly improves the number of iterations and computational time to convergence.

This experiment demonstrates the advantages of warm-starting the dual problem, which has the benefit of efficiently generating primal feasible solutions with lower error. By way of a baseline comparison, simple initializations of $\boldsymbol{\lambda}^{(0)}(k)$ would be to randomly select the dual variables or to set all of the dual variables equal so that $\boldsymbol{\lambda}^{(0)}(k) = [1/M, \ldots, 1/M]^T$. For these experiments the latter strategy is chosen. The initial iterate for the primal variable when the dual variables are all equal is then the uniformly weighted extrinsic mean of the data, that is, $\mathbf{U}^{(0)}(k)$ is the dominant k-dimensional eigenspace of $\sum_{i=1}^{M} \lambda_i^{(0)} X_i X_i^T$. On $\mathrm{Gr}(1, n)$, no warm-start initialization is possible because $\boldsymbol{\lambda}^*(0)$ is undefined, so the algorithm is run using only the naive initialization. For $k = 2, \ldots, \max_i\{\dim(\mathbf{X}_i)\}$ Figure 6 illustrates the relative speed-up due to smart initialization by comparing the number of iterations needed to find a stationary point for different choices of the initial dual variable using each of the data models. Both data models are intentionally structured so that the extrinsic mean is not the center of the GMEB on $\mathrm{Gr}(k^*, n)$. The naive initialization is indicated by the orange box-and-whisker plots, while the warm-start is denoted with red. The black dots mark the mean number of iterations and the solid line is the median.

In Figure 6a the data has been generated using the asymmetrical nested ball model with $M = 50$ points sampled from $\mathrm{Gr}(p_i, 10)$ for $p_i \in \{4, 5, 6\}$ and an optimal dimension of $k^* = 4$. The warm start converged in less iterations than the naive initialization in 359 out of 500 possible trials. An experiment using data generated by sampling more densely from a randomly selected arc of a unit ball is displayed in Figure 6b. Here, $M = 300$ points were generated on $\mathrm{Gr}(p_i, 10)$ with $p_i \in \{4, 5, 6\}$ where $k^* = 4$. In 415 out of 500 possible trials, the warm start converged in less iterations than the naive initialization.

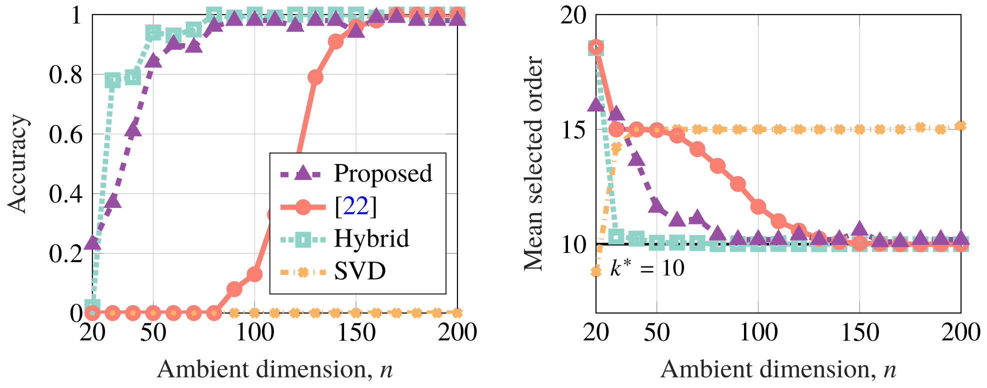
7.3. Experiment 3: Order-selection comparison

The previous experiments demonstrated the effectiveness of the proposed approach for computing the subspace at the center of the GMEB in a noise-free scenario. However the end-goal is to find a central subspace *and* the optimal size to best represent the common dimensions in a collection of data. Adding noise to the subspaces makes it difficult to identify how many common dimensions exist, thus the third experiment compares the ability of the proposed order-selection rule to identify the optimal dimension of the common subspace with that of the technique from Santamaria *et al.* [22] as the difficulty of the task varies.

In many machine learning applications, a common low-rank subspace is extracted from data as a pre-processing task, but the rank of this subspace is selected with little care. The most commonly used solutions are heuristics for locating the elbow of the scree plot, that is, computing the SVD of the concatenated data sets, finding the the singular values that represent the significant information, and keeping the dimensions corresponding to these singular values. This can be done with a variety of techniques such as the L-method [37], which estimates the elbow as the intersection of the two lines that minimize the root mean-squared error of the projection of the points in the of the scree plot onto the lines, the method of [38], which maximizes the profile log-likelihood under an independence assumption, and even just visually inspecting the scree plot to identify the first significant change in the first derivative [39]. To justify the need for a more principled way of selecting a subspace dimension, we additionally compare to the elbow of the scree plot using the L-method, and expect it to provide bad results. In the experiments this technique is denoted “SVD.”

Figure 7 shows a comparison of order-selection rules for $M = 20$ points generated using the asymmetrical nested ball model from Section 6.1 with both generalizations. The data has $M_1 = 10$ points that are sampled uniformly from the boundary of $\mathcal{B}_1(\mathbf{Z}_1) \subset \mathrm{Gr}(10, n)$ and $M_2 = 10$ points that are sampled from the boundary of $\mathcal{B}_{.5}(\mathbf{Z}_2) \subset \mathrm{Gr}(15, n)$. Each of the points is completed to a basis for a point on $\mathrm{Gr}(p_i, n)$ for $p_i \in \{10, 11, \ldots, 20\}$ and $n = 20, 30, \ldots, 200$. Zero-mean Gaussian noise is added to create noisy data sets. The signal-to-noise ratio (SNR) of the data is the total power of the signal divided by the total power of the noise. In order to have the same SNR for each subspace despite differing dimensions, the noise variance per component is scaled by the number of subspace dimensions. Since X_i is an orthonormal basis for $\mathbf{X}_i$, the magnitude of each basis vector is 1. Thus the total power of signal subspace is k^*, and the SNR is computed as $\mathrm{SNR} = 10 \log_{10}({k^*}/{\sigma_N^2})$, where σ_N^2 is the total variance of the noise. In this example the order of the common subspace is $k^* = 10$ and $\sigma_N^2 = 1.259$ meaning that the data has an SNR of 9dB.

Figure 7a shows the percentage of 100 Monte Carlo trials for which the proposed order-selection rule (purple dashed line with triangle markers), the method of Santamaría *et al.* [22] (pink solid line with circle markers), the hybrid method (turquoise dotted line with square markers), and the elbow point of the SVD (orange dash-dotted line with circle markers) were able to correctly identify the optimal order of the common subspace relative to the ambient dimension.



(a) Accuracy, $\frac{\text{Number of times } k^*=10}{\text{Number of trials}}$

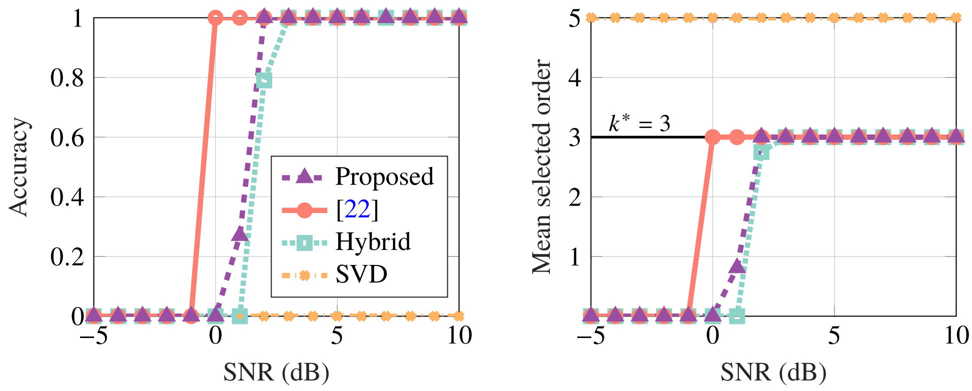
(b) Mean selected order

Figure 7: Order-selection accuracy and mean selected order relative to the ambient dimension of the data from 100 Monte Carlo trials using the proposed order-selection rule (purple dashed line with triangle markers), the method of Santamaría *et al.* [22] (pink solid line with circle markers), the hybrid method (turquoise dotted line with square markers), and the elbow point of the SVD (orange dash-dotted line with circle markers). The data consists 20 points from $\coprod_{p \in \mathcal{P}} \mathrm{Gr}(p, n)$ for $\mathcal{P} = \{10, 11, \ldots, 20\}$ and $n = 20, 30, \ldots, 200$ with an SNR of 9dB generated according to the model in Section 6.1.

Figure 7b shows the mean selected order, averaged across all trials. We can see that when the ambient dimension is small, all methods other than the SVD tend to overestimate the order of the common subspace. This is a result of the noise dimensions being relatively close in the low-dimensional spaces. The dimension of $\mathrm{Gr}(k, n)$ is $k(n - k)$, so for $k \approx \max_i\{p_i\} \approx n$ all samples are very similar regardless of the data model. As the ambient dimension grows and the randomly selected dimensions become further apart on average, the proposed method and the hybrid method correctly select the order with a high degree of accuracy. The proposed method achieves slightly lower accuracy and has less stable performance than the hybrid method because $c_{\text{pen}}(k)$ can be significantly affected by even one subspace that is similar to $\mathbf{U}^{*\perp}(k)$. However, this behavior is consistent with the assumption that every sample is valid and there are no outliers in the collection of data. As expected, [22] initially estimates the order as the dimension of the common subspace for the smaller ball and over-estimates the order as 15, while the methods that rely on the minimum enclosing ball estimate the dimension of the common subspace for that support set. Predictably, the elbow point of the SVD has low accuracy regardless of the ambient dimension. In essence, this method attempts to preserve all dimensions that are not pure noise.

Figure 8 shows a comparison using data from the second model, a ball that is sampled more densely from a random arc. For some $\mathbf{Z}_1 \in \mathrm{Gr}(3, 100)$, $M_1 = 200$ points are sampled uniformly from $\mathcal{B}_{0.5}(\mathbf{Z}_1) \subset \mathrm{Gr}(3, 100)$ and $M_2 = 25$ additional points are then sampled from a random arc on the same ball. No points were sampled from the interior of the ball. Each of these $M = 225$ subspaces is completed to a basis for a point on $\mathrm{Gr}(p_i, 100)$ for $p_i \in \{3, 4, 5\}$, and zero-mean Gaussian noise is added to each basis to create noisy data sets. In this experiment, the ambient dimension is fixed and we allow the SNR to vary from −5dB to 10dB.

With this data the optimal order of the common subspace is $k^* = 3$ and center of the ball is $\mathbf{U}^*(3) = \mathbf{Z}_1$. Figure 8a shows the percentage of 100 Monte Carlo trials for which the proposed order-selection rule (purple dashed line with triangle markers), the method of Santamaría *et al.* [22] (pink solid line with circle markers), the hybrid method (turquoise dotted line with square



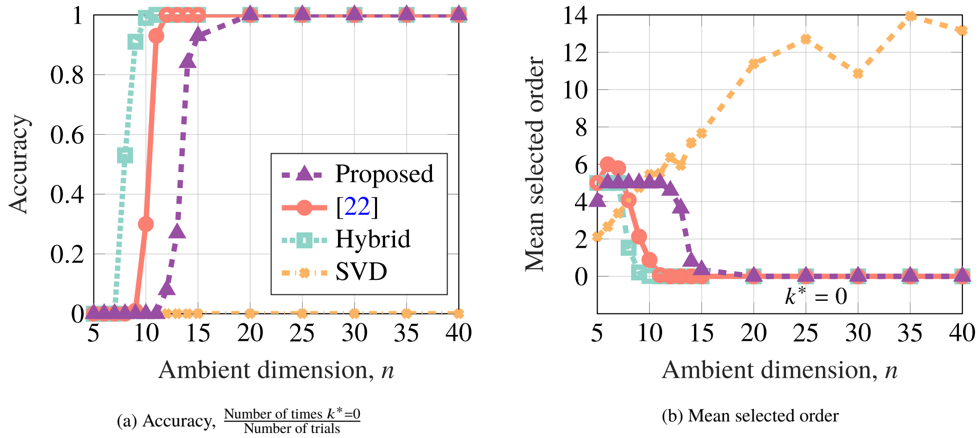
(a) Accuracy, $\frac{\text{Number of times } k^*=3}{\text{Number of trials}}$

(b) Mean selected order

Figure 8: Order-selection accuracy and mean selected order relative to the signal-to-noise ratio of the data (in dB) from 100 Monte Carlo trials using the proposed order-selection rule (purple dashed line with triangle markers), the method of Santamaría *et al*. [22] (pink solid line with circle markers), the hybrid method (turquoise dotted line with square markers), and the elbow point of the SVD (orange dash-dotted line with circle markers). The data consists 225 points from $\coprod_{p\in\mathcal{P}} \mathrm{Gr}(p, 100)$ for $\mathcal{P} = \{3, 4, 5\}$ generated according to the model in Section 6.2.

markers), and the elbow point of the SVD (orange dash-dotted line with circle markers) were able to correctly identify the optimal order of the common subspace relative to the signal-to-noise ratio. Figure 8b shows the mean selected order in the same trials. This experiment demonstrates the behavior of the different rules when all of the points are in the support of the minimum enclosing ball on $\mathrm{Gr}(k^*, n)$. Each of the subspace averaging methods should theoretically select the same order in this experiment, because all of the points share the same number of dimensions and there is no ambiguity about the optimal solution. Thus even though the mean computed by [22] is not the same point as the center of the GMEB, they lead to the same estimated rank. We see that in this scenario, the behavior of the rules using ℓ_∞-norm and the ℓ_2-norm are similar with a sharp phase transition when the power of the signal and the power of the noise are almost equal, although the ℓ_2-norm transitions to the correct order at a slightly higher noise power. This suggests that for situations where the data is free from outliers and the ℓ_∞-mean is close to the ℓ_2-mean, either technique will accurately estimate the number of common dimensions. The elbow point of the singular value decomposition fails to identify the common dimension in all trials.

Finally, in Figure 9 we see the ability of each method to identify when there is no subspace common to a collection of points. This is a valuable test because estimating $k^* = 0$ suggests that there is no information shared across all the data and that averaging the points is not an appropriate way to aggregate the information in the data. The data in this experiment consists of 50 subspaces chosen uniformly at random from $\mathrm{Gr}(p_i, n)$ for $p_i \in \{3, 4, 5\}$ for $i = 1, \ldots, 10$ with ambient dimensions $n = 5, 6, \ldots, 15, 20, 25, \ldots, 40$. The noise variance does not affect performance in this task because there is no signal so the SNR is undefined. In Figure 9a we see a similar phase transition to that of Figure 8. The hybrid method is able to achieve perfect accuracy for ambient dimensions greater than 10, while [22] and the proposed method transition shortly thereafter. The SVD fails every time, but that is to be expected in this scenario. The elbow point method computes two lines that minimize the residual for the scree plot, and chooses dimension as the index of the singular value just larger than the intersection of those lines. A



(a) Accuracy, $\frac{\text{Number of times } k^*=0}{\text{Number of trials}}$

(b) Mean selected order

Figure 9: Order-selection accuracy and mean selected order relative to the ambient dimension of the data when there is no common subspace. Results are from 100 Monte Carlo trials using the proposed order-selection rule (purple dashed line with triangle markers), the method of Santamaría *et al.* [22] (pink solid line with circle markers), the hybrid method (turquoise dotted line with square markers), and the elbow point of the SVD (orange dash-dotted line with circle markers). The data consists 50 points from $\coprod_{p \in \mathcal{P}} \mathrm{Gr}(p, n)$ for $\mathcal{P} = \{3, 4, 5\}$ and $n = 5, 6, \ldots, 15, 20, 25, \ldots, 40$.

line cannot be fit to zero points, so the method will not select $k^* = 0$ or $k^* = n$ as a solution. However, in Figure 9b we see that the SVD is significantly overestimating the dimension of the (non-existent) common subspace, so the poor performance is not an issue of the method being unable to select 0 as the optimal dimension. When n is small the proposed algorithm incorrectly identifies a relationship between the subspaces, but as the ambient dimension grows the optimal order, $k^* = 0$, is selected with increasing accuracy. As noted in discussion of Figure 7, the misidentifications in low dimensions are due to the minimum similarity between the points and $\mathbf{U}^{*\perp}(k)$ being higher when $k \approx \max_i\{p_i\} \approx n$.

8. Conclusions

The recent trend of performing machine learning tasks on linear subspace data has created a need for flexible subspace averages, ones that can be computed accurately and in a principled manner for subspaces of differing dimension. In response to this need, we have proposed an algorithm to find the ℓ_∞-center of mass using a subgradient algorithm to solve the dual problem with respect to a point-to-set distance. We additionally proposed a flexible data generation model to create subspaces of differing dimensions with ground-truth for the GMEB that emulates realistic settings where an ℓ_∞-average would be appropriate. On this synthetic data, the proposed algorithm provides estimates of the GMEB center with high accuracy. However, the high computational complexity means that an existing primal method can provide low-accuracy solutions more quickly for large data sets. One direction for future expansion is to develop a core-set theory akin to that of [2] in order to estimate the GMEB on a subset of the data with theoretical accuracy guarantees. A related area for further study is to develop an active-set approach for ℓ_∞-averaging of mixed-dimensional subspaces, à la John [40]. Active-set methods also attempt to minimize the cost function over a subset of the data. However, the active-set approach looks for a subset of the data that solves the original problem exactly, whereas the core-set technique computes error bounds on the solution provided by *any* subset of a given size. One theoretical hurdle to achieving an active-set method is a theorem on the minimum number of points required to define a Grassmannian ball given a fixed Grassmann manifold and subspaces of differing dimensions.

Finally, we proposed a geometric order-fitting rule that estimates the best dimension for the common subspace. This rule fits the common dimensions of the subspaces in the support set of the minimum enclosing ball, which is appropriate for data where all subspace samples are assumed to be valid examples of the model of interest. We additionally implement a hybrid technique for estimating the dimension of the common subspace that modifies the order-selection rule of [22] for use with the ℓ_∞-average. This hybrid method would not be possible for existing techniques that estimate the GMEB, because it uses the values of the dual variables as weights for an eigenvalue decomposition at each potential order. The hybrid approach outperforms the proposed technique and that of [22] when the ambient dimension is close to the subspace dimension of the data points.

A high-accuracy estimate of the GMEB center combined with an order-selection rule for the number of common dimensions results in a powerful technique for detecting and estimating similarity in a collection of subspaces. We anticipate that many practical applications will arise in the form of distributed large-scale problems, where the subspace averaging can be used for aggregation, for example the sparse subspace clustering of [13].

Acknowledgments

The authors would like to thank Emilie Renard for the stimulating discussions that improved the ideas presented here.

References

[1] M. Arnaudon, F. Nielsen, On approximating the Riemannian 1-center, Computational Geometry 46 (1) (2013) 93–104. 1, 2, 18

[2] M. Bădoiu, K. L. Clarkson, Smaller core-sets for balls, in: Proc. ACM-SIAM Symposium on Discrete Algorithms, SIAM, 2003, pp. 801–802. 1, 2, 24

[3] E. Renard, K. A. Gallivan, P.-A. Absil, A Grassmannian minimum enclosing ball approach for common subspace extraction, in: Proc. International Conference on Latent Variable Analysis and Signal Separation, Springer, 2018, pp. 69–78. 1, 2, 5, 6, 9, 17, 18, 19

[4] P. Kumar, J. S. Mitchell, E. A. Yildirim, Approximate minimum enclosing balls in high dimensions using core-sets, Journal of Experimental Algorithmics 8 (2003) 1–1. 1

[5] K. Fischer, B. Gärtner, The smallest enclosing ball of balls: combinatorial structure and algorithms, International Journal of Computational Geometry & Applications 14 (04n05) (2004) 341–378. 1

[6] E. A. Yildirim, Two algorithms for the minimum enclosing ball problem, SIAM Journal on Optimization 19 (3) (2008) 1368–1391. 1

[7] F. Nielsen, R. Nock, Approximating smallest enclosing balls with applications to machine learning, International Journal of Computational Geometry & Applications 19 (05) (2009) 389–414. 1

[8] P. Turaga, A. Veeraraghavan, A. Srivastava, R. Chellappa, Statistical computations on Grassmann and Stiefel manifolds for image and video-based recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (11) (2011) 2273–2286. 2

[9] Q. Rentmeesters, P. Absil, P. Van Dooren, K. Gallivan, A. Srivastava, An efficient particle filtering technique on the Grassmann manifold, in: Proc. International Conference on Acoustics, Speech and Signal Processing, IEEE, 2010, pp. 3838–3841. 2, 18

[10] T. Marrinan, J. Beveridge, B. Draper, M. Kirby, C. Peterson, Flag-based detection of weak gas signatures in long-wave infrared hyperspectral image sequences., in: Proc. SPIE Defense, Security, and Sensing, International Society for Optics and Photonics, 2016. 2

[11] B. Afsari, Riemannian l^p center of mass: existence, uniqueness, and convexity, Proceedings of the American Mathematical Society 139 (2) (2011) 655–673. 2

[12] E. Renard, P.-A. Absil, K. A. Gallivan, Minimax center to extract a common subspace from multiple datasets, in: Proc. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2019. 2

[13] M. Abdolali, N. Gillis, M. Rahmati, Scalable and robust sparse subspace clustering using randomized clustering and multilayer graphs, Signal Processing 163 (2019) 166–180. 2, 25

[14] A. Srivastava, E. Klassen, Bayesian and geometric subspace tracking, Advances in Applied Probability 36 (1) (2004) 43–56. 2

[15] J.-M. Chang, C. Peterson, M. Kirby, et al., Feature patch illumination spaces and Karcher compression for face recognition via Grassmannians, Advances in Pure Mathematics 2 (4) (2012) 226–242. 2

[16] R. Chakraborty, B. C. Vemuri, Recursive Frechet mean computation on the Grassmannian and its applications to computer vision, in: Proc. International Conference on Computer Vision and Pattern Recognition, IEEE, 2015, pp. 4229–4237. 2

[17] S. O'Hara, B. A. Draper, Scalable action recognition with a subspace forest, in: Proc. Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 1210–1217. 2

[18] X. Ma, M. Kirby, C. Peterson, L. Scharf, Self-organizing mappings on the Grassmannian with applications to data analysis in high dimensions, Neural Computing and Applications (2018) 1–12. 2

[19] E. Jurrus, N. Hodas, N. Baker, T. Marrinan, M. D. Hoover, Adaptive visual sort and summary of micrographic images of nanoparticles for forensic analysis, in: Proc. Symposium on Technologies for Homeland Security, IEEE, 2016, pp. 1–6. 2

[20] T. Franz, R. Zimmermann, S. Görtz, N. Karcher, Interpolation-based reduced-order modelling for steady transonic flows via manifold learning, International Journal of Computational Fluid Dynamics 28 (3-4) (2014) 106–121. 2

[21] K. Sim, R. Hartley, Removing outliers using the l_∞ norm, in: Proc. Conference on Computer Vision and Pattern Recognition, IEEE, 2006. 2

[22] I. Santamaría, L. L. Scharf, C. Peterson, M. Kirby, J. Francos, An order fitting rule for optimal subspace averaging, in: Proc. Statistical Signal Processing Workshop, IEEE, 2016, pp. 1–4. 3, 12, 13, 14, 15, 17, 21, 22, 23, 24, 25

[23] G. W. Stewart, J.-G. Sun, Matrix Perturbation Theory, Elsevier, 1990. 3

[24] K. Ye, L.-H. Lim, Schubert varieties and distances between subspaces of different dimensions, SIAM Journal on Matrix Analysis and Applications 37 (3) (2016) 1176–1197. 3, 4, 12

[25] A. Björck, G. Golub, Numerical methods for computing angles between linear subspaces, Mathematics of Computation 27 (123) (1973) 579–594. 3

[26] Y.-C. Wong, Differential geometry of Grassmann manifolds, Proceedings of the National Academy of Sciences of the United States of America 57 (3) (1967) 589. 4

[27] A. N. Schwickerath, Linear models, signal detection, and the Grassmann manifold, Ph.D. thesis, Colorado State University. Libraries (2014). 4

[28] M. L. Overton, R. S. Womersley, Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices, Mathematical Programming 62 (1-3) (1993) 321–357. 7, 8

[29] N. Z. Shor, Minimization Methods for Non-Differentiable Functions, Vol. 3, Springer Science & Business Media, 2012. 7, 8

[30] D. P. Bertsekas, Nonlinear programming, Journal of the Operational Research Society 48 (3) (1997) 334–334. 7, 8

[31] J.-B. Hiriart-Urruty, C. Lemaréchal, Convex Analysis and Minimization Algorithms I: Fundamentals, Vol. 305, Springer Science & Business Media, 2013. 8

[32] F. E. Curtis, T. Mitchell, M. L. Overton, A BFGS-SQP method for nonsmooth, nonconvex, constrained optimization and its evaluation using relative minimization profiles, Optimization Methods and Software 32 (2017) 148–181. 9

[33] A. V. Knyazev, M. E. Argentati, Majorization for changes in angles between subspaces, Ritz values, and graph Laplacian spectra, SIAM Journal on Matrix Analysis and Applications 29 (1) (2006) 15–32. 14

[34] V. Garg, I. Santamaría, D. Ramírez, L. L. Scharf, Subspace averaging and order determination for source enumeration, IEEE Transactions on Signal Processing 67 (11) (2019) 3028–3041. 14

[35] T. Marrinan, B. Draper, J. R. Beveridge, M. Kirby, C. Peterson, Finding the subspace mean or median to fit your need, in: Proc. Conference on Computer Vision and Pattern Recognition, IEEE, 2014, pp. 1082–1089. 18

[36] A. Uschmajew, B. Vandereycken, Greedy rank updates combined with Riemannian descent methods for low-rank optimization, in: Proc. International Conference on Sampling Theory and Applications, IEEE, 2015, pp. 420–424. 20

[37] S. Salvador, P. Chan, Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms, in: Proc. International Conference on Tools with Artificial Intelligence, IEEE, 2004, pp. 576–584. 21

[38] M. Zhu, A. Ghodsi, Automatic dimensionality selection from the scree plot via the use of profile likelihood, Computational Statistics & Data Analysis 51 (2) (2006) 918–930. 21

[39] M. Steyvers, Multidimensional scaling, Encyclopedia of Cognitive Science (2006). 21

[40] F. John, Extremum problems with inequalities as subsidiary conditions, in: Traces and Emergence of Nonlinear Programming, Springer, 2014, pp. 197–215. 24

Appendix A. GMEB dual subgradient algorithm

Algorithm 1 Algorithm to minimize Equation (14) with back-tracking line search

```
function GMEB($\{\mathbf{X}_i\}_{i=1}^M, k, a, \eta, \zeta, \beta$)
    input: Data: $\{\mathbf{X}_i\}_{i=1}^M$, Rank: $k$, Step size parameter: $a$, Stopping criteria: $\eta$, Step size threshold: $\zeta$, Growth parameter: $\beta$
    output: Weights: $\lambda^*$, Minimax center: $\mathbf{U}^*$
    $t \leftarrow 0$
    $\lambda^{(t)} \leftarrow [1/M, \dots, 1/M]^T \in \mathbb{R}^M$        ▷ $\lambda^{(t)} \leftarrow \lambda^*(k-1)$ for warm-start
    $\mathbf{U}^{(t)} \leftarrow$ dominant $k$ eigenvectors$\big( \sum_{i=1}^M \lambda_i^{(t)} X_i X_i^T \big)$
    $\mathbf{g}^{(t)} \leftarrow -\big[d_{\mathrm{Gr}(k,n)}(\mathbf{U}^{(t)}, \mathbf{X}_1), d_{\mathrm{Gr}(k,n)}(\mathbf{U}^{(t)}, \mathbf{X}_2), \dots, d_{\mathrm{Gr}(k,n)}(\mathbf{U}^{(t)}, \mathbf{X}_M)\big]^T$
    $f_{\mathrm{primal}}(\mathbf{U}^{(t)}) \leftarrow \min_{i=1,\dots,M}\{-d_{\mathrm{Gr}(k,n)}(\mathbf{U}^{(t)}, \mathbf{X}_i)\}$        ▷ Primal cost at iteration $t$
    $f_{\mathrm{dual}}(\lambda^{(t)}) \leftarrow \lambda^{(t)T} \mathbf{g}^{(t)}$        ▷ Dual cost at iteration $t$
    while $f_{\mathrm{dual}}(\lambda^{(t)}) - f_{\mathrm{primal}}(\mathbf{U}^{(t)}) > \eta$ and $\max\limits_{i=1,\dots,10}\{f_{\mathrm{dual}}(\lambda^{(t-i)}) - f_{\mathrm{dual}}(\lambda^{(t)})\} > \eta$ do
        $t \leftarrow t + 1$
        $\alpha^{(t)} \leftarrow a/\sqrt{t}$
        $\lambda^{(t)} \leftarrow \lambda^{(t-1)} - \alpha^{(t)} \mathbf{g}^{(t-1)}$, $\lambda^{(t)} \leftarrow \lambda^{(t)}/\|\lambda^{(t)}\|_1$
        $\mathbf{U}^{(t)} \leftarrow$ dominant $k$ eigenvectors$\big( \sum_{i=1}^M \lambda_i^{(t)} \mathbf{X}_i \mathbf{X}_i^T \big)$
        $\mathbf{g}^{(t)} \leftarrow -\big[d_{\mathrm{Gr}(k,n)}(\mathbf{U}^{(t)}, \mathbf{X}_1), d_{\mathrm{Gr}(k,n)}(\mathbf{U}^{(t)}, \mathbf{X}_2), \dots, d_{\mathrm{Gr}(k,n)}(\mathbf{U}^{(t)}, \mathbf{X}_M)\big]^T$
        $\tilde{\alpha}^{(t)} \leftarrow \alpha^{(t)}$
        $\tilde{\lambda}^{(t)} \leftarrow \lambda^{(t)}$
        $f_{\mathrm{dual}}(\tilde{\lambda}^{(t)}) \leftarrow \tilde{\lambda}^{(t)T} \mathbf{g}^{(t)}$
        while $f_{\mathrm{dual}}(\tilde{\lambda}^{(t)}) > f_{\mathrm{dual}}(\lambda^{(t-1)})$ and $\tilde{\alpha}^{(t)} > \zeta \alpha^{(t)}$ do        ▷ Back-tracking line search
            $a \leftarrow a/2$
            $\tilde{\alpha}^{(t)} \leftarrow a/\sqrt{t}$
            $\tilde{\lambda}^{(t)} \leftarrow \lambda^{(t-1)} - \tilde{\alpha}^{(t)} \mathbf{g}^{(t-1)}$, $\tilde{\lambda}^{(t)} \leftarrow \tilde{\lambda}^{(t)}/\|\tilde{\lambda}^{(t)}\|_1$
            $\tilde{\mathbf{U}}^{(t)} \leftarrow$ dominant $k$ eigenvectors$\big( \sum_{i=1}^M \tilde{\lambda}_i^{(t)} \mathbf{X}_i \mathbf{X}_i^T \big)$
            $\tilde{\mathbf{g}}^{(t)} \leftarrow -\big[d_{\mathrm{Gr}(k,n)}(\tilde{\mathbf{U}}^{(t)}, \mathbf{X}_1), d_{\mathrm{Gr}(k,n)}(\tilde{\mathbf{U}}^{(t)}, \mathbf{X}_2), \dots, d_{\mathrm{Gr}(k,n)}(\tilde{\mathbf{U}}^{(t)}, \mathbf{X}_M)\big]^T$
            $f_{\mathrm{dual}}(\tilde{\lambda}^{(t)}) \leftarrow \tilde{\lambda}^{(t)T} \tilde{\mathbf{g}}^{(t)}$
            if $f_{\mathrm{dual}}(\tilde{\lambda}^{(t)}) \le f_{\mathrm{dual}}(\lambda^{(t-1)})$ then        ▷ Update variables if $f_{\mathrm{dual}}$ decreases
                $a \leftarrow \beta a$
                $\lambda^{(t)} \leftarrow \tilde{\lambda}^{(t)}$
                $\mathbf{U}^{(t)} \leftarrow \tilde{\mathbf{U}}^{(t)}$
                $\mathbf{g}^{(t)} \leftarrow \tilde{\mathbf{g}}^{(t)}$
        $f_{\mathrm{primal}}(\mathbf{U}^{(t)}) \leftarrow \min_{i=1,\dots,M}\{-d_{\mathrm{Gr}(k,n)}(\mathbf{U}^{(t)}, \mathbf{X}_i)\}$
        $f_{\mathrm{dual}}(\lambda^{(t)}) \leftarrow \lambda^{(t)T} \mathbf{g}^{(t)}$
    return $\lambda^{(t)}$, $\mathbf{U}^{(t)}$
```