arXiv:2304.09154v1 [stat.ME] 18 Apr 2023

Sharp-SSL：用于半监督学习的选择性高维轴对齐随机投影

TengyaoWang*、Edgar Dobriban†、Milana Gataric‡和 Richard J. Samworth‡

*伦敦政治经济学院统计学系
†宾夕法尼亚大学统计学与数据科学系
‡剑桥大学统计实验室

2023年4月19日

摘要

We propose a new method for high-dimensional semi-supervised learning problems based on the careful aggregation of the results of a low-dimensional procedure applied to many axis-aligned random projections of the data. Our primary goal is to identify important variables for distinguishing between the classes; existing low-dimensional methods can then be applied for final class assignment. Motivated by a generalized Rayleigh quotient, we score projections according to the traces of the estimated whitened between-class covariance matrices on the projected data. This enables us to assign an importance weight to each variable for a given projection, and to select our signal variables by aggregating these weights over high-scoring projections. Our theory shows that the resulting `Sharp-SSL` algorithm is able to recover the signal coordinates with high probability when we aggregate over sufficiently many random projections and when the base procedure estimates the whitened between-class covariance matrix sufficiently well. The Gaussian EM algorithm is a natural choice as a base procedure, and we provide a new analysis of its performance in semi-supervised settings that controls the parameter estimation error in terms of the proportion of labeled data in the sample. Numerical results on both simulated data and a real colon tumor dataset support the excellent empirical performance of the method.

1 引言

Semi-supervised learning, where we attempt to assign observations to one of finitely many groups based on partially-labeled training data, represents a core modern statistical challenge. It is sufficiently general to incorporate, at either extreme, the unsupervised case of no labeled training data (clustering) and the supervised setting of fully-labeled training data (classification). Such tasks abound in many application areas, including genomics (e.g., Eisen et al., 1998), image processing (Jain and Flynn, 1996; Cheplygina, de Bruijne and Pluim, 2019), natural language processing (Liang, 2005; Turian, Ratinov and Bengio, 2010) and anomaly detection (Akcay, Atapour-Abarghouei and Breckon, 2019; Wang et al., 2019). Entry points to the literature on semi-supervised learning

include Zhu (2005), Zhu and Goldberg (2009), Chapelle, Schölkopf and Zien (2006) and Van Engelen and Hoos (2020). For introductions to clustering, see Xu and Wunsch (2005), Kaufman and Rousseeuw (2009) and Xu and Tian (2015), and for classification, see Devroye, Györfi and Lugosi (2013) and Hastie, Tibshirani and Friedman (2009).

A common feature of contemporary semi-supervised learning problems is high-dimensionality, since we may record many covariates having a possible association with the labels corresponding to different observations. This represents a significant challenge, as can be seen by considering a simple two-class problem with more covariates than observations. For any given assignment of class labels, if no subset of n_0 observations lies in an $(n_0 - 2)$-dimensional affine space, then we can find hyperplanes with orthogonal normal vectors, each of which achieves zero training error (in other words, they perfectly separate the classes). Nevertheless, even in the simple setting where the true Bayes decision boundary is linear, many such hyperplanes may be little better than a random guess on test data.

An appealing approach to tackling high-dimensionality is via random projections into lower-dimensional spaces. Such projections may almost preserve the pairwise distances between observations, as seen from the Johnson–Lindenstrauss lemma (Johnson and Lindenstrauss, 1984; Dasgupta and Gupta, 2003). Moreover, in cases where we have reason to believe that only a relatively small proportion of the variables recorded are relevant for the learning task, we can choose our random projections to be axis-aligned in order to preserve this structure. A third benefit is the possibility of aggregating results over multiple random projections, though this must be done with care so as to avoid noise accumulation. These attractions have meant that random projections have now been employed in many high-dimensional statistical problems, including precision matrix estimation (Marzetta, Tucci and Simon, 2011), two-sample mean testing (Lopes, Jacob and Wainwright, 2011), classification (Durrant and Kabán, 2015; Cannings and Samworth, 2017), (sparse) principal component analysis (Yang et al., 2021; Gataric, Wang and Samworth, 2020), linear regression (Thanei, Heinze and Meinshausen, 2017; Slawski, 2018; Dobriban and Liu, 2019; Ahfock, Astle and Richardson, 2021), clustering (Dasgupta, 1999; Fern and Brodley, 2003; Han and Boutin, 2015; Yellamraju and Boutin, 2018; Anderlucci, Fortunato and Montanari, 2022) and dimensionality reduction (Bingham and Mannila, 2001; Reeve, Kabán and Bootkrajang, 2022). See Cannings (2021) for a review of recent developments in the area.

In this paper, we propose a new method, called `Sharp-SSL` (short for **S**elective **h**igh-dimensional, **a**xis-aligned **r**andom **p**rojections for **S**emi-**S**upervised **L**earning). Our primary goal is to identify a small subset of variables that are particularly helpful for label assignment; existing low-dimensional methods can then be used to complete the learning task. To this end, we generate a large number of axis-aligned random projections, and apply a base learning procedure such as a semi-supervised version of the Gaussian Expectation–Maximization (EM) algorithm to our projected data. Motivated by the notion of a generalized Rayleigh quotient (see (2) below for a formal definition), and to avoid the noise accumulation issue mentioned above, we score the projections by computing the trace of the corresponding estimated whitened between-class covariance matrices. This enables us to assign an importance weight to each variable for a given projection, and we select our signal variables by aggregating these importance weights over the high-scoring projections. See Section 2 for a more detailed description of our methodology.

第3 节致力于对我们的 `Sharp-SSL` 算法进行理论分析。我们首先在定理2 中证明，只要低维基学习过程满足

a guarantee on the proximity of the estimated whitened between-class covariance matrix to its population analogue, the corresponding high-dimensional semi-supervised learning algorithm can recover the signal coordinates with high probability when we aggregate over sufficiently many random projections. It turns out that both Linear Discriminant Analysis and an EM algorithm are examples of low-dimensional learning procedures that satisfy this proximity guarantee, as we prove in Theorems 3 and 6 respectively. The latter is particularly challenging, and one of the main novel contributions of our analysis is to provide a guarantee on the performance of a d-dimensional Gaussian EM algorithm in a semi-supervised setting. In particular, we control the parameter estimation error in terms of the proportion of labeled data in the sample, showing that with a sample size of n it smoothly interpolates between the $(d/n)^{1/4}$ rate for unsupervised learning and the $(d/n)^{1/2}$ rate for fully-labeled data, up to logarithmic factors. An advantage of the modular approach to our analysis is that it illustrates the way in which the `Sharp-SSL` algorithm can be combined with different base learning algorithms to adapt to different problem settings and reflect the preferences of the practitioner.

In Section 4, we study the numerical performance of the `Sharp-SSL` algorithm. Our first goal, in Section 4.1, is to study the effect of the choices of input parameters to our method, which allows us to recommend sensible default choices for application in our subsequent comparisons. Section 4.2 presents the results of a simulation study involving the `Sharp-SSL` method, as well as five alternative approaches, on high-dimensional clustering tasks (since not all of the competing methods are able to leverage partial label information). We find that the `Sharp-SSL` algorithm is able to attain a misclustering rate very close to that of the optimal Bayes classifier, even with only around 50 observations per cluster, in settings where these alternative techniques may perform poorly. In Section 4.3, we investigate the extent to which the different versions of the `Sharp-SSL` method are able to leverage partial label information. The results here are consistent with the phase transition phenomenon articulated by our theory. Finally, in Section 4.4, we apply the `Sharp-SSL` algorithm, as well as the other methods from our simulation study, on a colon tumor dataset, where we withhold the true labels from the algorithms in order to assess performance. Our analysis supports the ability of the `Sharp-SSL` algorithm to identify signal coordinates (genes) that are useful for identifying patients with and without tumors.

In the broader literature on high-dimensional learning problems, a large number of methods have been developed to leverage sparse low-dimensional structures for both clustering (Witten and Tibshirani, 2010; Azizyan, Singh and Wasserman, 2013; Wasserman, Azizyan and Singh, 2014; Azizyan, Singh and Wasserman, 2015; Jin and Wang, 2016; Verzelen and Arias-Castro, 2017; Löffler, Wein and Bandeira, 2022; Löffler, Zhang and Zhou, 2021) and classification (Cai and Liu, 2011; Witten and Tibshirani, 2011; Mai, Zou and Yuan, 2012; Cai and Zhang, 2019). These methods are not designed for partially-labeled (semi-supervised) settings. Another common approach is to project the data into the span of the top few principal components, and run a standard low-dimensional method such as k-means clustering or the EM algorithm (Butler et al., 2018). This approach can fail if the directions of largest variation in the data are not aligned with the directions separating the clusters. Finally, recent developments in other aspects of semi-supervised learning include self-training (Oymak and Gulcu, 2020), mean estimation (Zhang, Brown and Cai, 2019), choice of k in k-nearest neighbour classification (Cannings, Berrett and Samworth, 2020) and linear regression (Chakrabortty and Cai, 2018).

证明 我们所有结果的详细内容见第5节，我们在此结束本引言 n

with some notation used throughout the paper. We write $\mathbb{S}^{d\times d}$ for the set of d-dimensional symmetric matrices, write $\mathbb{S}^{d\times d}_{+}$ for the subset that are invertible, and write $\mathbb{S}^{d\times d}_{K-1}$ the subset of matrices in $\mathbb{S}^{d\times d}$ of rank at most $K-1$. We write $\mathbb{R}^{d\times d}$ for the set of d-dimensional matrices. For $p \geq d$, let $\mathbb{O}^{p\times d}$ denote the set of $p \times d$ matrices with orthonormal columns. The Euclidean norm is denoted by $\|\cdot\|$, and the operator norm of a matrix is denoted by $\|\cdot\|_{\mathrm{op}}$, so that $\|A\|_{\mathrm{op}} := \sup_{\{x:\|x\|=1\}} \|Ax\|$. Given two sequences (a_n) and (b_n), we write $a_n \lesssim b_n$ when there exists a universal constant $C > 0$ such that $a_n \leq Cb_n$, and, given an additional problem parameter R, we write $a_n \lesssim_R b_n$ when there exists $C > 0$, depending only on R, such that $a_n \leq Cb_n$.

For any set $S \subseteq \mathbb{R}^d$ and $d \leq |S|$, we write $\binom{S}{d} := \{A \subseteq S : |A| = d\}$. If $S \subseteq \mathbb{R}^d$, we define $\operatorname{sargmax} S$ to be the smallest element in the argmax in the lexicographic order. For a positive integer k, we define $[k] := \{1, \ldots, k\}$. For a vector $v = (v_1, \ldots, v_k)^\top \in \mathbb{R}^k$, and $j \in [k]$, we define $v_{-j} = (v_1, \ldots, v_{j-1}, v_{j+1}, \ldots, v_k)^\top \in \mathbb{R}^{k-1}$.

2 Sharp-SSL算法

In this section, we describe in detail the `Sharp-SSL` algorithm for K-class semi-supervised learning, with $K \geq 2$. We aim to provide a unified treatment of clustering, semi-supervised learning and classification. To this end, we assume that for $i \in [n]$, the observation $x_i \in \mathbb{R}^p$ has a true label $y_i^* \in [K]$, but it may be the case that we do not observe y_i^*. Instead, we assume that our observed label y_i takes values in $[K] \cup \{0\}$, where $y_i := y_i^*$ when the true class label is observed, and $y_i := 0$ otherwise. Thus, our data can be regarded as $(x_1, y_1)\ \ldots, (x_n, y_n) \in \mathbb{R}^p \times ([K] \cup \{0\})$, and our goal is to construct a *data-dependent classifier*[1], i.e. a Borel measurable function $C : \mathbb{R}^p \times \big(\mathbb{R}^p \times ([K] \cup \{0\})\big)^n \to [K]$, with the interpretation that $C\big(x; (x_1, y_1)\ \ldots, (x_n, y_n)\big)$ is the predicted class of $x \in \mathbb{R}^p$.

To motivate our `Sharp-SSL` algorithm, it is instructive first to consider a canonical Gaussian classification problem, where our data can be regarded as n independent realizations of a pair (X, Y) taking values in $\mathbb{R}^p \times [K]$, with prior probability $\pi_k := \mathbb{P}(Y = k)$ for the kth class and $X \mid Y = k \sim \mathcal{N}_p(\nu_k, \Sigma_{\mathrm{w}})$, for class means $\nu_1, \ldots, \nu_K \in \mathbb{R}^p$ and *within-class covariance matrix* $\Sigma_{\mathrm{w}} \in \mathbb{S}^{p\times p}_{+}$. Let $\nu := \sum_{k=1}^K \pi_k \nu_k \in \mathbb{R}^p$ denote the grand population mean, let

$$\Sigma_{\mathrm{b}} := \sum_{k=1}^{K} \pi_k (\nu_k - \nu)(\nu_k - \nu)^\top \in \mathbb{S}^{p\times p}_{K-1} \tag{1}$$

表示类间协方差矩阵，并考虑列空间由 $\Sigma_{\mathrm{w}}^{-1}(\nu_1 - \nu), \ldots, \Sigma_{\mathrm{w}}^{-1}(\nu_K - \nu)$ 张成的 $D \in \mathbb{O}^{p\times(K-1)}$ 。注意，对于 $k \neq \ell$,

$$\log\left\{\frac{\mathbb{P}(Y = k \mid X = x)}{\mathbb{P}(Y = \ell \mid X = x)}\right\} = \log\Big(\frac{\pi_k}{\pi_\ell}\Big) - \frac{1}{2}(\nu_k + \nu_\ell)^\top \Sigma_{\mathrm{w}}^{-1}(\nu_k - \nu_\ell) + x^\top \Sigma_{\mathrm{w}}^{-1}(\nu_k - \nu_\ell),$$

from which we deduce that this likelihood ratio, and hence the Bayes classifier $x \mapsto \operatorname{argmax}_{k\in[K]} \mathbb{P}(Y = k \mid X = x)$, only depends on x through $D^\top x$. Thus, for the purposes of classification, no signal would be lost (and the noise would be reduced) if X were replaced with $D^\top X$.

[1]这里使用“分类器”一词很方便，尽管部分或全部标签可能未被观测到。

In high-dimensional settings with $p \gg n$, the matrix Σ_{w}^{-1} is not consistently estimable in general, but we can nevertheless make progress if the vectors $\Sigma_{\mathrm{w}}^{-1}(\nu_1 - \nu)$ $\ldots, \Sigma_{\mathrm{w}}^{-1}(\nu_K - \nu)$ are sparse. In other words, writing S_0 for the union of the set of coordinates for which these vectors are non-zero, we suppose that $|S_0| \ll p$; this is a very common assumption in high-dimensional LDA (e.g. Cai and Liu, 2011; Witten and Tibshirani, 2011; Mai, Zou and Yuan, 2012; Cai and Zhang, 2019).

In such a setting, the column space of D has a sparse basis, so it is natural to consider projecting the data onto a small subset of its coordinates. For $d \in [p]$, define the set of axis-aligned projection matrices $\mathcal{P}_d := \{P \in \{0,1\}^{d \times p} : PP^\top = I_d\}$, i.e. the set of binary $d \times p$ matrices with orthonormal rows. By the argument above, if $d \geq |S_0|$ then there exists $P^* \in \mathcal{P}_d$ such that the error of the Bayes classifier is unchanged by projecting the data along P^*. In practice, it would typically be computationally too expensive to enumerate through all $p(p-1)\cdots(p-d+1)$ axis-aligned projections. Instead, we consider a randomly chosen subset of projections within $\mathcal{P}_d$. An axis-aligned projection chosen uniformly at random is unlikely to capture all the signal coordinates S_0, but by aggregating over a carefully-chosen subset of these random projections, we can nevertheless recover the set of signal coordinates under suitable conditions; see Theorem 2 below. To describe our method for choosing good projections, for $V \in \mathbb{O}^{p \times d}$, we define the *generalized Rayleigh quotient* along V by

$$J(V; \Sigma_{\mathrm{b}}, \Sigma_{\mathrm{w}}) := \mathrm{tr}\{(V^\top \Sigma_{\mathrm{w}} V)^{-1}(V^\top \Sigma_{\mathrm{b}} V)\}. \tag{2}$$

命题 1 下面的结果说明了为何要通过选择投影来最大化广义瑞利商：任意最大化者 $J(V; \Sigma_{\mathrm{b}}, \Sigma_{\mathrm{w}})$ 在$V \in \mathbb{O}^{p \times d}$ 上的列空间必须包含 D的列空间。

Proposition 1. *Let $K \geq 2$ and $d \geq K-1$. Assume that the convex hull of $\nu_1, \ldots, \nu_K$ is $(K-1)$-dimensional, and let $V^* \in \mathrm{argmax}_{V \in \mathbb{O}^{p \times d}} J(V; \Sigma_{\mathrm{b}}, \Sigma_{\mathrm{w}})$. Then the column space of V^* contains the eigenspace corresponding to the $K-1$ non-zero eigenvalues[2] of $\Sigma_{\mathrm{w}}^{-1}\Sigma_{\mathrm{b}}$, which is equal to the space spanned by $\big(\Sigma_{\mathrm{w}}^{-1}(\nu_k - \nu) : k \in [K]\big)$.*

Based on Proposition 1, a natural conceptual approach to maximizing the generalized Rayleigh quotient is to compute the leading $(K-1)$-dimensional eigenspace of $\Sigma_{\mathrm{w}}^{-1}\Sigma_{\mathrm{b}}$. This strategy, however, runs into difficulties when we replace these population quantities with their sample versions in the setting of the opening paragraph of this section. More precisely, writing $n_k := \sum_{i=1}^n \mathbb{1}_{\{y_i = k\}}$ for $k \in [K]$, as well as

$$\tilde{\Sigma}_{\mathrm{w}} := \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}(x_i - \hat{\nu}_k)(x_i - \hat{\nu}_k)^\top \mathbb{1}_{\{y_i = k\}} \in \mathbb{R}^{p \times p}$$

$$\tilde{\Sigma}_{\mathrm{b}} := \sum_{k=1}^{K}\frac{n_k}{n}(\hat{\nu}_k - \hat{\nu})(\hat{\nu}_k - \hat{\nu})^\top \in \mathbb{R}^{p \times p},$$

for the sample versions of the within-class and between-class covariance matrices respectively, the matrix $\tilde{\Sigma}_{\mathrm{w}}$ is not invertible whenever $p > n$. Fortunately, though, this issue can be resolved by working with the projected data, as long as we choose $d \leq n - K$: the projected data $\{PX_i : i \in [n]\}$ has within-class covariance matrix $P\Sigma_{\mathrm{w}}P^\top \in \mathbb{R}^{d \times d}$

[2]尽管 $\Sigma_{\mathrm{w}}^{-1}\Sigma_{\mathrm{b}}$ 不保证是对称的，但它与对称矩阵 $\Sigma_{\mathrm{w}}^{-1/2}\Sigma_{\mathrm{b}}\Sigma_{\mathrm{w}}^{-1/2}$ 相似（即共轭），因此具有实特征值和特征向量。

以及类间协方差矩阵 $P\Sigma_{\mathrm{b}}P^\top \in \mathbb{R}^{d\times d}$，因此以概率 1，样本版本 $P\tilde{\Sigma}_{\mathrm{w}}P^\top$ 可逆。

Returning to the general setting of the opening paragraph of this section, then, we seek projections P with large $J(P^\top; \tilde{\Sigma}_{\mathrm{b}}, \tilde{\Sigma}_{\mathrm{w}}) = \mathrm{tr}\big\{(P\tilde{\Sigma}_{\mathrm{w}}P^\top)^{-1}(P\tilde{\Sigma}_{\mathrm{b}}P^\top)\big\}$. To this end, for fixed $A, B \in \mathbb{N}$, we sample a set of projections $\{P^{a,b} : a \in [A], b \in [B]\}$ uniformly at random from $\mathcal{P}_d$. For each a and b, we apply a low-dimensional base algorithm $\psi : \big(\mathbb{R}^d \times ([K] \cup \{0\})\big)^n \to \mathbb{S}_+^{d\times d}$ to the projected data $(P^{a,b}x_1, y_1) \ldots, (P^{a,b}x_n, y_n)$ to obtain an estimator $\hat{Q}^{a,b}$ of $(P^{a,b}\Sigma_{\mathrm{w}}P^{a,b,\top})^{-1}(P^{a,b}\Sigma_{\mathrm{b}}P^{a,b,\top})$, the *whitened between-class covariance matrix* of the projected data. We assume throughout for convenience that ψ is *permutation equivariant* in the sense that $\psi\big((\Pi z_1, y_1) \ldots, (\Pi z_n, y_n)\big) = \Pi\psi\big((z_1, y_1) \ldots, (z_n, y_n)\big)\Pi^\top$ for every permutation matrix $\Pi \in \mathbb{R}^{d\times d}$. One choice for the base algorithm is to set $\hat{Q}^{a,b} = (\hat{\Sigma}_{\mathrm{w}}^{a,b})^{-1}\hat{\Sigma}_{\mathrm{b}}^{a,b}$, where $\hat{\Sigma}_{\mathrm{w}}^{a,b}$ and $\hat{\Sigma}_{\mathrm{b}}^{a,b}$ are estimated projected within- and between-class (or cluster) covariance matrices.

为了选择投影，对于每个 $a \in [A]$，我们定义

$$b^*(a) := \operatorname*{sargmax}_{b\in[B]} \mathrm{tr}(\hat{Q}^{a,b}),$$

and select $P^{a,b^*(a)}$. The main rationale for dividing the projections into A groups and selecting one within each group—as opposed to selecting the A projections with the largest values of $\mathrm{tr}(\hat{Q}^{a,b})$—is that, conditional on the original data, the selected projections are independent and identically distributed. This facilitates our theoretical analysis by enabling the application of concentration inequalities in the proof of Theorem 2.

$\{$的对角元素 $\hat{Q}^{a,b^*(a)} : a \in [A]\}$ measure the importance of the projected variables for the semi-supervised learning task. These can then be converted into importance scores for the original variables by 'back-projecting' into the higher-dimensional space, i.e. by forming the vector $\hat{w} = (\hat{w}_1, \ldots, \hat{w}_p)^\top \in \mathbb{R}^p$ given by

$$\hat{w}_j := \frac{1}{A}\sum_{a=1}^{A}\Big[P^{a,b^*(a),\top}\hat{Q}^{a,b^*(a)}P^{a,b^*(a)}\Big]_{j,j}, \quad j \in [p].$$

Finally, we rank the variables by their importance scores, and our estimate $\hat{S}$ of the set of signal coordinates is given by the largest ℓ entries in $\hat{w}$, breaking ties arbitrarily if necessary, where $\ell \in [p]$ is specified by the practitioner. Pseudocode for the `Sharp-SSL` procedure is given in Algorithm 1.

$\hat{}$ 在应用算法 1 获得信号变量的估计集合 S 后，我们便可以应用任何现有的低维数据半监督学习方法，其输入为$\hat{}$ $(P_{\hat{}S}x_i,\ y_i)_{i\in[n]}$，其中 $P_{\hat{S}}$表示投影到 S中的坐标。

2.1 基础学习方法

Algorithm 1 relies on a base learning method for low-dimensional data to estimate the projected whitened between-class covariance matrix from the projected data. When all or almost all of the input data are labeled, we can use the procedure outlined in Algorithm 2, which ignores any unlabeled data, for this purpose. On the other hand, when we have a substantial amount of unlabeled data, Algorithm 2 may be inaccurate. In such circumstances, it may be preferable to use Algorithm 3, which runs an *Expectation–Maximization* (EM) procedure to predict the unobserved labels and subsequently estimate the whitened between-class covariance matrix. More precisely, from M random initializations of the

算法 1： Sharp-SSL：通过集成进行聚类和半监督学习-
轴对齐随机投影的表格。

输入： 数据 $(x_1, y_1) \ldots, (x_n, y_n) \in \mathbb{R}^p \times ([K] \cup \{0\})$（其中 $y_i = 0$ 表示一个缺失标签）；
投影维度 $d \in [\min(p, n - K)]$, 所选信号的数量
坐标 $\ell \in [p]$;
投影组的数量 $A \in \mathbb{N}$，每组投影的数量 $B \in \mathbb{N}$
每个组;
置换等变基础算法 $\psi : \left(\mathbb{R}^d \times ([K] \cup \{0\})\right)^n \to \mathbb{S}_+^{d \times d}$。

生成轴对齐随机投影 $\{P^{a,b} : a \in [A], b \in [B]\}$ 独立地
并从 $\mathcal{P}_d$ 中均匀地抽取。

for $a \in [A]$ **do**

  对于 $b \in [B]$ **do**

    Let $\hat{Q}^{a,b} := \psi((P^{a,b} x_1, y_1) \ldots (P^{a,b} x_n, y_n))$。

  end

  集合 $b^*(a) := \operatorname{sargmax}_{b \in [B]} \operatorname{tr}(\hat{Q}^{a,b})$.

end

设 $\hat{w} = (\hat{w}_1, \ldots, \hat{w}_p)^\top$，其中 $\hat{w}_j := \frac{1}{A} \sum_{a=1}^{A} [P^{a,b^*(a)\top} \hat{Q}^{a,b^*(a)} P^{a,b^*(a)}]_{j,j}$.

输出： $\hat{S} \subseteq [p]$，定义为$\hat{w}$中最大的 ℓ 个分量的索引集，
随机打破平局。

cluster means and the within-class covariance matrix, Algorithm 3 uses the EM algorithm to update these quantities, and thereby compute the whitened between-cluster sample covariance matrix estimators $\left\{\hat{Q}^{[m]} = (\hat{\Sigma}_{\mathrm{w}}^{[m]})^{-1}(\hat{\Sigma}_{\mathrm{b}}^{[m]}) : m \in [M]\right\}$. We select $\hat{m} \in [M]$ such that $\hat{Q}^{[\hat{m}]}$ is in best agreement with results from the other EM runs; this is made precise in (6).

The algorithm also allows the practitioner to incorporate prior knowledge about the true cluster means and within-cluster covariance matrices, both through optimizing over a restricted constraint set $\mathcal{C}$ in the M step of the EM algorithm, and through the choice of a distribution supported on $\mathcal{C}$ for the initialization of these quantities. An alternative to the EM algorithm for unsupervised learning would be to apply k-means clustering as a base procedure. Previous studies have suggested that these approaches have comparable empirical performance (e.g., de Souto et al., 2008; Rodriguez et al., 2019, and references therein), but the EM algorithm is more amenable to theoretical analysis in our setting.

3 理论保证

3.1 高层算法的结果

In this subsection, we consider independent triples $(X_1, Y_1, Y_1^*) \ldots, (X_n, Y_n, Y_n^*)$ taking values in $\mathbb{R}^p \times ([K] \cup \{0\}) \times [K]$. We recall that Y_i^* denotes the true label of the ith observation, and that $Y_i := Y_i^*$ if the ith label is observed, and $Y_i := 0$ otherwise. For $k \in [K]$, let $\pi_k := \mathbb{P}(Y_1^* = k)$ and $\nu_k^* := \mathbb{E}(X_1 \mid Y_1^* = k)$ denote the prior probability and the cluster mean of the kth cluster respectively, let $\nu^* := \sum_{k=1}^{K} \pi_k \nu_k^*$ denote the weighted cluster mean and let $\Sigma_{\mathrm{w}} := \operatorname{Cov}(X_1 \mid Y_1^* = k)$ denote the common within-cluster covariance matrix. With the between-cluster covariance matrix Σ_{b} from (1), our

算法 2： 仅使用有标签数据进行基学习

Input: $(z_1, y_1)\ \ldots, (z_n, y_n) \in \mathbb{R}^d \times ([K] \cup \{0\})$, where the convex hull of $(z_i)_{i:y_i=k}$ is d-dimensional for at least one $k \in [K]$.

Compute $\hat{\mu} := n^{-1} \sum_{i=1}^n z_i$

for $k \in [K]$ **do**

  Set $n_k := |\{i : y_i = k\}|$ and $n' := \sum_{k=1}^K n_k$

  Compute $\hat{\mu}_k := n_k^{-1} \sum_{i:y_i=k} z_i$ (with the convention that $\hat{\mu}_k := 0$ if $n_k = 0$).

end

计算类内和类间协方差矩阵为

$$\hat{\Sigma}_{\mathrm{w}} := \frac{1}{n'} \sum_{i=1}^n (z_i - \hat{\mu}_{y_i})(z_i - \hat{\mu}_{y_i})^\top \quad \text{and} \quad \hat{\Sigma}_{\mathrm{b}} := \sum_{k=1}^K \frac{n_k}{n'} (\hat{\mu}_k - \hat{\mu})(\hat{\mu}_k - \hat{\mu})^\top. \tag{3}$$

输出： $\hat{}\ Q$: $= {}^{\hat{}}\Sigma_{\mathrm{w}}^{-1}{}^{\hat{}}\Sigma_{\mathrm{b}}$。

目标是估计信号坐标集合，

$$S_0 := \{j \in [p] : (\Sigma_{\mathrm{w}}^{-1}\Sigma_{\mathrm{b}})_{j,j} \neq 0\},$$

并记作 $s_0 := |S_0|$。

我们的首个主要理论结果表明，如果基础算法在每个低维投影上都准确无误，并且 A 足够大，那么以高概率，所有信号坐标都会被选中。

Theorem 2. *Define* $\gamma_{\min} := \min_{j \in S_0} (\Sigma_{\mathrm{w}}^{-1}\Sigma_{\mathrm{b}})_{j,j}$ *and* $\gamma_{\max} := \max_{j \in S_0} (\Sigma_{\mathrm{w}}^{-1}\Sigma_{\mathrm{b}})_{j,j}$. *Let* $\hat{S}$ *be the output of Algorithm 1 with input* K, p, $(X_1, Y_1)\ \ldots, (X_n, Y_n)$, A, B, $d \geq s_0$, $\ell \geq s_0$ *and permutation equivariant base procedure* ψ. *Write*

$$\varepsilon := \mathbb{P}\left(\max_{P \in \mathcal{P}_d} \left\| \psi\big((PX_i, Y_i)_{i \in [n]}\big) - P\Sigma_{\mathrm{w}}^{-1}\Sigma_{\mathrm{b}}P^\top \right\|_{\mathrm{op}} \geq \frac{\gamma_{\min}}{4(K-1)} \right). \tag{7}$$

Then

$$\mathbb{P}(S_0 \subseteq \hat{S}) \geq 1 - \varepsilon - pe^{-A\gamma_{\min}^2/(50p^2\gamma_{\max}^2)}.$$

事实上，我们可以从定理的证明中看出2 以下更强的结论成立：对于满足以下条件的数据的任意一次实现（x_i, y_i）$_{i \in [n]}$

$$\max_{P \in \mathcal{P}_d} \left\| \psi\big((Px_i, y_i)_{i \in [n]}\big) - P\Sigma_{\mathrm{w}}^{-1}\Sigma_{\mathrm{b}}P^\top \right\|_{\mathrm{op}} < \frac{\gamma_{\min}}{4(K-1)}, \tag{8}$$

we have $\mathbb{P}\big(S_0 \subseteq \hat{S} \mid (X_i, Y_i)_{i \in [n]} = (x_i, y_i)_{i \in [n]}\big) \geq 1 - pe^{-A\gamma_{\min}^2/(50p^2\gamma_{\max}^2)}$. Note here that, after conditioning on the data, the probability is taken over the randomness in the projections. An attraction of Theorem 2 is its generality, and in particular the fact that we do not impose strong distributional assumptions — we simply require control of ε in (7). The price we pay for this generality is that the probability bound may be loose in particular cases; for example, the bound holds even with $B = 1$, though in practice we would expect it to improve as B increases.

算法3： 通过EM算法使用部分有标签数据进行基学习

输入： 数据 $(z_1,\ y_1),\ldots,(z_n,\ y_n) \in \mathbb{R}^d \times ([K] \cup \{0\})$。 一个约束集 $\mathcal{C} \subseteq (\mathbb{R}^d)^K \times \mathbb{S}_+^{d\times d}$和支撑于 $\mathcal{C}$ 上的概率分布 $\pi_{\mathcal{C}}$。
随机初始化次数 M。迭代次数 T。

对于 $m \in [M]$ **执行**

随机采样 $(\hat{\mu}_1, \ldots, \hat{\mu}_K, \hat{\Sigma}_{\mathrm{w}}) \sim \pi_{\mathcal{C}}$.

对于 $t \in [T]$ **执行此操作**

（E 步）计算软标签矩阵 $(L_{i,\ k})_{i\in[n],\ k\in[K]}$

$$L_{i,k} := \left(\frac{e^{-\frac{1}{2}(z_i-\hat{\mu}_k)^\top \hat{\Sigma}_{\mathrm{w}}^{-1}(z_i-\hat{\mu}_k)}}{\sum_{\ell=1}^K e^{-\frac{1}{2}(z_i-\hat{\mu}_\ell)^\top \hat{\Sigma}_{\mathrm{w}}^{-1}(z_i-\hat{\mu}_\ell)}} \right) \mathbb{1}_{\{y_i=0\}} + \mathbb{1}_{\{y_i=k\}}. \tag{4}$$

（M 步）更新参数估计值通过

$$(\hat{\mu}_1, \ldots, \hat{\mu}_K, \hat{\Sigma}_{\mathrm{w}}) := \operatorname*{argmin}_{(\mu_1,\ldots,\mu_K,\Sigma)\in\mathcal{C}} \left\{ \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K L_{i,k}(z_i-\mu_k)^\top \Sigma^{-1}(z_i-\mu_k) \ + \text{对数行列式} \Sigma \right\}. \tag{5}$$

end

使用 $(L_{i,k})_{i\in[n],k\in[K]}$ 计算 $(\hat{\mu}_1, \ldots, \hat{\mu}_K,$的最终值。 $\hat{\Sigma}_{\mathrm{w}})$ 如（4）。

计算 $\hat{\mu}_{\mathrm{tot}} := n^{-1} \sum_{i=1}^n \sum_{k=}{}_1^K L_{i,k} \hat{\mu}_k$ 以及类间协方差矩阵

$$\hat{\Sigma}_{\mathrm{b}} := \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K L_{i,k}(\hat{\mu}_k - \hat{\mu}_{\mathrm{tot}})(\hat{\mu}_k - \hat{\mu}_{\mathrm{tot}})^\top.$$

Set $\hat{Q}^{[m]} := \hat{\Sigma}_{\mathrm{w}}^{-1} \hat{\Sigma}_{\mathrm{b}}$.

end

Set

$$\hat{m} \in \operatorname*{argmin}_{m\in[M]} \text{中位数}\big(\| \hat{Q}^{[m]} - \hat{Q}^{[m']} \|_{\mathrm{op}} :\ m' \in [M] \setminus \{m\} \big). \tag{6}$$

输出： $\hat{Q} := \hat{Q}^{[\hat{m}]}$.

3.2 Theory for base learning using labeled data

In this subsection, we demonstrate how the high-level result in Theorem 2 can be used to derive performance guarantees for a high-dimensional classification algorithm that uses the `Sharp-SSL` procedure in Algorithm 1 in conjunction with the low-dimensional base method described in Algorithm 2 for estimating the projected whitened between-class covariance matrix. The following theorem provides uniform control of the output of Algorithm 2 for all axis-aligned d-dimensional projected datasets.

Theorem 3. *Fix $\varepsilon \in (0,1]$, $K \in \{2,3,\dots,\}$, $p, d \in \mathbb{N}$ with $p \geq d$ and $n \in \mathbb{N}$ with $n \geq Kd+1$. Suppose that $(X_1,Y_1)\dots,(X_n,Y_n)$ are independent and identically distributed pairs, with $\mathbb{P}(Y_1=k)=\pi_k$ and $X_1 \mid Y_1=k \sim \mathcal{N}_p(\nu_k^*,\Sigma_{\mathrm{w}})$ for $k\in[K]$, and let $\psi\big((PX_i,Y_i)_{i\in[n]}\big)$ be the output of Algorithm 2 with input $(X_i,Y_i)_{i\in[n]}$, for $P\in\mathcal{P}_d$. Suppose that $\|\nu_k^*-\nu_\ell^*\|\leq R_1$ for all $k,\ell\in[K]$ and some $R_1>0$, and that Σ_{w} is diagonal and well-conditioned in the sense that $\max\{\|\Sigma_{\mathrm{w}}\|_{\mathrm{op}},\|\Sigma_{\mathrm{w}}^{-1}\|_{\mathrm{op}}\}\leq R_2$ for some $R_2\geq 1$. If*

$$\frac{16R_2^2K}{n}\leq 1 \quad \text{and} \quad \frac{32R_2^2}{n^{1/2}}\log^{1/2}\biggl(\frac{8\cdot 9^d\binom{p}{d}}{\varepsilon}\biggr)\leq 1, \tag{9}$$

然后，以至少 $1-\varepsilon$ 的概率，我们有

$$\max_{P\in\mathcal{P}_d}\bigl\|\psi\big((PX_i,Y_i)_{i\in[n]}\big)-(P\Sigma_{\mathrm{w}}P^\top)^{-1}P\Sigma_{\mathrm{b}}P^\top\bigr\|_{\mathrm{op}} \lesssim_{R_1,R_2} \frac{K}{n}+\sqrt{\frac{d\log(ep/d)+\log(1/\varepsilon)}{n}}.$$

The sample size condition (9) can be restated as $n \gtrsim_{R_1,R_2} d\log p+\log(1/\varepsilon)+K$, so may be regarded as mild. Regarding K as a constant, Theorem 3 confirms that the uniform control of Algorithm 2 is at the parametric rate, up to a logarithmic factor. The following corollary then follows immediately by combining Theorems 2 and 3.

Corollary 4. *Fix $\varepsilon\in(0,1]$. Suppose that the conditions of Theorem 3 hold, and moreover that $\lambda_{\min}(\Sigma_{\mathrm{b}})\geq 1/R_3$ for some $R_3>0$. Then there exist $C_1,C_2>0$, depending only on R_1,R_2 and R_3, such that if*

$$C_1\biggl(\frac{K}{n}+\sqrt{\frac{d\log(ep/d)+\log(1/\varepsilon)}{n}}\biggr)\leq\frac{1}{K},$$

ˆ那么，算法 S 的输出1以 K, p, $d\geq s_0$, $\ell\geq s_0$, $(X_1,Y_1)\dots,(X_n,Y_n)$,A, B，以及算法 ψ 2中的基础程序 满足

$$\mathbb{P}(S_0\subseteq\hat{S})\geq 1-\varepsilon-p\exp\biggl(-\frac{A}{C_2p^2}\biggr).$$

Thus, under the conditions of Corollary 4, the `Sharp-SSL` algorithm can, with high probability, select the signal variables in the top s_0 output variables, provided that the number A of groups of random projections is large by comparison with p^2. In other words, the algorithm reduces the problem to a low-dimensional one, for which standard learning techniques can be applied. The guarantees for these methods (e.g. Anderson, 2003, Theorem 6.6.1) can then be combined on the high-probability event of Corollary 4 to establish theoretical results for the full procedure.

3.3 半监督学习理论

When the proportion of labeled data is low, Algorithm 2 may be inaccurate when used as the base procedure in Algorithm 1. The aim of this subsection, therefore, is to study the base procedure of Algorithm 3, which is able to leverage both the labeled and unlabeled data via an EM algorithm to estimate the whitened between-class covariance matrix for each projected data set. Our analysis builds on several recent breakthroughs in our understanding of the EM algorithm. This line of work includes Balakrishnan, Wainwright and Yu (2017), Daskalakis, Tzamos and Zampetakis (2017), Yan, Yin and Sarkar (2017), Dwivedi et al. (2020a), Dwivedi et al. (2020b), Davis, Diaz and Wang (2021), Ho et al. (2020), Ndaoud (2022), Wu and Zhou (2022) and Doss et al. (2023), all of which focus on the unsupervised case.

为简便起见，我们将重点讨论以下情形：独立同分布的 $(X_1, Y_1^*), \ldots, (X_n, Y_n^*)$ 生成自一个均值相反且具有单位协方差矩阵的双高斯混合：

$$Y_i^* \sim \mathrm{Unif}(\{1,2\}),\ X_i \mid Y_i^* \sim \mathcal{N}_p\big((-1)^{Y_i^*}\nu^*, I_p\big),\ \text{and } Y_i = Y_i^* \mathbb{1}_{\{i \le n_{\mathrm{L}}\}} \text{ for all } i \in [n]. \tag{10}$$

We assume that we observe $(X_1, Y_1) \ldots, (X_{n_{\mathrm{L}}}, Y_{n_{\mathrm{L}}})\ X_{n_{\mathrm{L}}+1}, \ldots, X_n$ for some $n_{\mathrm{L}} \in \{0, \ldots, n\}$. In other words, we are given n_{L} labeled observations and $n_{\mathrm{U}} := n - n_{\mathrm{L}}$ unlabeled ones. Thus, $n_{\mathrm{L}} = 0$ corresponds to the fully unsupervised case, i.e., clustering, while $n_{\mathrm{L}} = n$ corresponds to the supervised case, i.e., classification. We define $Y_i = Y_i^*$ for $i \in [n_{\mathrm{L}}]$, and $Y_i = 0$ for $i \in \{n_{\mathrm{L}} + 1, \ldots, n\}$.

We first study the performance of the EM procedure after the covariates have been projected into a lower-dimensional space. In other words, for some fixed $P \in \mathcal{P}_d$, define $Z_i := PX_i$ for $i \in [n]$ and $\mu^* := P\nu^* \in \mathbb{R}^d$, so that $Z_i \mid Y_i^* \sim \mathcal{N}_d((-1)^{Y_i^*}\mu^*, I_d)$. In this setting, we have a single unknown parameter μ^* to estimate, and this can be achieved by applying Algorithm 3 to $(Z_i, Y_i)_{i\in[n]}$ with $K = 2$ and the constraint set

$$\mathcal{C} := \big\{(-\mu, \mu, I_d) : \mu \in \mathbb{R}^d\big\}. \tag{11}$$

在某个固定值处初始化 EM算法之后 $(-\hat{}\mu^{(0)}, \hat{}\mu^{(0)}, I_d) \in \mathcal{C}$，对于 $t \in \mathbb{N}$，EM迭代在 (4) 和 (5) 中描述的第 t 次迭代结果为 $(-\hat{}\mu^{(t)}, \hat{}\mu^{(t)}, I_d)$，其中

$$\hat{\mu}^{(t)} := \frac{1}{n}\bigg\{\sum_{i:Y_i\neq 0}(-1)^{Y_i} Z_i + \sum_{i:Y_i=0} Z_i \tanh\big\langle Z_i, \hat{\mu}^{(t-1)}\big\rangle\bigg\}; \tag{12}$$

see Lemma 15. Since we allow $n_{\mathrm{L}} = 0$, where μ is only identifiable up to sign, and since the between-class sample covariance matrix $\hat{\Sigma}_{\mathrm{b}}$ computed in Algorithm 3 is equal to $\hat{\Sigma}_{\mathrm{b}} = \hat{\mu}_1\hat{\mu}_1^\top - \hat{\mu}_{\mathrm{tot}}\hat{\mu}_{\mathrm{tot}}^\top$, which is invariant to flipping the signs of $\hat{\mu}_1$ and $\hat{\mu}_2$ simultaneously, it is natural to consider the loss function $L : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$ given by

$$L(\mu, \mu') := \|\mu - \mu'\| \wedge \|\mu + \mu'\|.$$

Proposition 5 below provides a theoretical guarantee for this semi-supervised EM algorithm. For notational simplicity, we define $\gamma := n_{\mathrm{L}}/n$, $\omega_0 := \sqrt{\{d \log n + \log(1/\delta)\}/n_{\mathrm{U}}}$ and $\zeta_0 := \min\{\omega_0\gamma^{-1/2}, \omega_0^{1/2}\}$ throughout this section. Thus, treating d as a constant and ignoring polylogarithmic terms, ω_0 is of order $n_{\mathrm{U}}^{-1/2}$ and ζ_0 is of order $\min\{n_{\mathrm{L}}^{-1/2}\ n_{\mathrm{U}}^{-1/4}\}$ when $\gamma < 1/2$. We remark that $n_{\mathrm{L}}^{-1/2}$ is the critical ℓ_2-testing radius for distinguishing the means of two labeled Gaussian distributions with identity

covariance using n_{L} observations. On the other hand, as we show in Lemma 16, no test of the null hypothesis $H_0 : \mathcal{N}_d(0, I_d)$ against the two-component mixture alternative $H_1 : \frac{1}{2}\mathcal{N}_d(\mu^*, I_d) + \frac{1}{2}\mathcal{N}_d(-\mu^*, I_d)$ based on n_{U} observations can have large power unless the signal strength $\|\mu^*\|$ is at least of order $n_{\mathrm{U}}^{-1/4}$.

Proposition 5. *Fix $\delta \in (2e^{-n}, 1]$ and $r \geq 1$, and suppose that $\|\mu^*\| \leq r$ and $\gamma < 1/2$. There exists $c > 0$, depending only on r, such that if $\omega_0 \leq c$ and $n \geq 3$, then the following statements hold:*

(i) 对于任意 $\hat{}\mu^{(0)} \in \mathbb{R}^d$，满足 $\|\hat{}\mu^{(0)}\| \leq r + 3$，我们以至少 $1 - 2\delta$ 的概率有时时彩官网

$$\limsup_{t\to\infty} L(\hat{\mu}^{(t)}, \mu^*) \lesssim_r \zeta_0 \vee \|\mu^*\|.$$

(ii) There exists $C > 0$, depending only on r, such that if $\|\mu^*\| \geq C\zeta_0\sqrt{d \log n}$ and $\hat{\mu}^{(0)} = (\zeta_0 \vee r\omega_0)\eta_0$ with $\eta_0 \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, then with probability at least $1 - 2\delta - \sqrt{2/(\pi \log n_{\mathrm{U}})}$, we have

$$\limsup_{t\to\infty} L(\hat{\mu}^{(t)}, \mu^*) \lesssim_r \frac{\omega_0}{\|\mu^*\|} \wedge \frac{\omega_0}{\gamma^{1/2}}.$$

In order to interpret Proposition 5*(i)*, consider the regime where $\|\mu^*\| \leq \zeta_0$. In this case, as discussed above, the two mixture components are essentially indistinguishable, and the bound reveals that the EM algorithm performs no worse than the trivial zero estimator, up to constant factors. On the other hand, part *(ii)* studies the more interesting regime where the two mixture components are distinguishable, and we establish a faster convergence rate for the EM algorithm in this strong signal regime.

The following theorem combines the two convergence regimes in Proposition 5 to derive a convergence guarantee for the estimated whitened between-class covariance matrix output by Algorithm 3. To state the result, recall the definition of $\mathcal{C}$ from (11). For any $\zeta > 0$, we write $U(\zeta)$ for the pushforward measure on $\mathcal{C}$ induced by $\mathrm{Unif}(\zeta\mathbb{S}^{d-1})$ under the map $\mu \mapsto (-\mu, \mu, I_d)$.

Theorem 6. *Fix $\delta \in (2e^{-n}, 1]$, and $r \geq 1$ and suppose that $\|\mu^*\| \leq r$ and $\gamma < 1/2$. There exists $c > 0$, depending only on r, such that if $\omega_0 \leq \min\{c, (d \log n)^{-3}\}$ and $n \geq 108$, then the sequence of outputs $(\hat{Q}^{(T)})_{T\in\mathbb{N}}$ of Algorithm 3 with inputs $(Z_1, Y_1) \ldots, (Z_n, Y_n)$, $\mathcal{C}$, $\pi_{\mathcal{C}} = U(\zeta_0 \vee r\omega_0)$, $M \in \mathbb{N}$ and $T \in \mathbb{N}$ satisfies with probability at least $1 - 3\delta - e^{-M/50}$ that*

$$\limsup_{T\to\infty} \|\hat{Q}^{(T)} - \mu^*\mu^{*\top}\|_{\mathrm{op}} \lesssim_r \frac{\omega_0}{\|\mu^*\|} \wedge \zeta_0.$$

Finally in this section, we study the implications of Theorem 6 for the recovery of the signal coordinates in the semi-supervised learning setting. We write $\psi^{(M,T)}$ for the base procedure that takes $(z_i, y_i)_{i\in[n]} \in \mathbb{R}^d \times ([K] \cup \{0\})$ as input and returns the output of Algorithm 3 when run with these inputs together with $\mathcal{C}$, $\pi_{\mathcal{C}}$, M and T. Let S_0 denote the set of coordinates where $\nu^* \in \mathbb{R}^p$ is non-zero, and let $s_0 := |S_0|$.

Corollary 7. *Fix $\varepsilon \in (8e^{-n/2}, 1]$, $r \geq 1$, and suppose that $\|\mu^*\| \leq r$, $M \geq 50\log(4/\varepsilon) + 50d\log p$ and $\gamma < 1/2$. Let $\nu^*_{\max} := \|\nu^*\|_\infty$ and let $\nu^*_{\min}$ denote the minimum absolute value of a non-zero component of ν^*. There exist $C_1, C_2 > 0$, depending only on r, such that if $n \geq C_1(d\log p)^6\{d\log p + \log(1/\varepsilon)\}$, and*

$$C_2 \min\left[\left\{\frac{d\log(p\vee n) + \log(1/\varepsilon)}{n}\right\}^{1/4}, \sqrt{\frac{d\log(p\vee n) + \log(1/\varepsilon)}{n_{\mathrm{L}}}}\right] \leq \frac{(\nu^*_{\min})^2}{4},$$

然后是输出序列 $(\hat{S}^{(T)})_{T\geq 1}$ 算法 1 的输入为 $K = 2$, p, $d \geq s_0$, $\ell \geq s_0$, $(X_i, Y_i)^{i\in[n]}$, A, B 和基础程序 $\psi^{(M,T)}$ 满足

$$\liminf_{T\to\infty} \mathbb{P}(S_0 \subseteq \hat{S}^{(T)}) \geq 1 - \varepsilon - pe^{-A(\nu^*_{\min})^4/(50p^2(\nu^*_{\max})^4)}.$$

Corollary 7 reveals in particular that, treating $\nu^*_{\max}$ and $\nu^*_{\min}$ as constants and under the stated sample size conditions, we again recover all of the signal coordinates in the top s_0 output entries, provided that A is large by comparison with p^2. Thus, in this sense, we can achieve a similar guarantee to that provided by Corollary 4, though the number of groups of projections required for a high probability guarantee in Corollary 7 may be significantly larger in settings where the ratio $\nu^*_{\max}/\nu^*_{\min}$ is large.

4 数值研究

Throughout this section, unless otherwise stated, data $(X_i, Y_i, Y_i^*)_{i\in[n]}$ are sampled from an equal-probability normal mixture as follows: $\mathbb{P}(Y_i^* = k) = 1/K$ for $k \in [K]$, $\mathbb{P}(Y_i = Y_i^*) = 1 - \mathbb{P}(Y_i = 0) = \gamma$ and $X_i \mid Y_i^* \sim N_p(\mu_{Y_i^*}, \Sigma_{\mathrm{w}})$. The cluster means $(\mu_k)_{k\in[K]}$ are chosen to be s_0-sparse and we define the signal-to-noise ratio of the problem to be[3]

$$\mathrm{SNR} := \frac{\min_{k,k'\in[K],k\neq k'} \|\mu_k - \mu_{k'}\|}{\sqrt{\mathrm{tr}(\Sigma_{\mathrm{w}})/p}}. \tag{13}$$

In our numerical studies, we slightly modify Algorithm 3 so that instead of randomly initializing the cluster means and the covariance matrix, we use the output of hierarchical clustering to initialize the EM algorithm as implemented in the `mclust` `R` package (Fraley and Raftery, 1998). This allow us to run Algorithm 3 with $M = 1$.

4.1 调优参数的选择

The purpose of this subsection is to investigate the effect of the various input parameters A, B, d and ℓ in Algorithm 1, and to recommend sensible default choices. In Figure 1, we plot the misclustering rate with Algorithm 3 as a base procedure in our Gaussian semi-supervised learning setting as each of these parameters varies, for four different SNR levels. After applying Algorithm 1, we obtain our final estimated cluster labels by using Algorithm 3 again on the data projected onto the selected coordinates with a single hierarchical clustering initialization. We then output the predicted labels, computed as $\hat{y}_i := \mathrm{sargmax}_{k\in[K]} L_{i,k}$ for $i \in [n]$, instead of $\hat{Q}$.

The panels in Figure 1 reveal that the misclustering rate is quite robust to the choices of A, B and d, and that it is less serious (and may even help) to choose ℓ larger—rather than smaller—than s_0. In particular, it seems that $A = 150$ suffices for almost optimal performance (though there appears to be some penalty for choosing it to be as small as 50), and $B = 75$ appears adequate. There is no clear trend on performance with the choice of d, so for simplicity we took $d = s_0$ in our remaining simulations below. Finally, the misclustering rate appears to decrease as ℓ increases, with an elbow in the curve visible at the highest value of the SNR when ℓ is set to the true sparsity level s_0. Of course, if ℓ is chosen to be very large, then we will include many noise variables, and

[3]在我们的一些模拟实验中，Σ_{w} 为了方便，随机生成了。在此类设置下，我们将 $(\Sigma_{\mathrm{w}})/p$ 在 (13) 的分母中的 tr($\mathbb{E}\{\mathrm{tr}(\Sigma_{\mathrm{w}})\}/p$替换为。

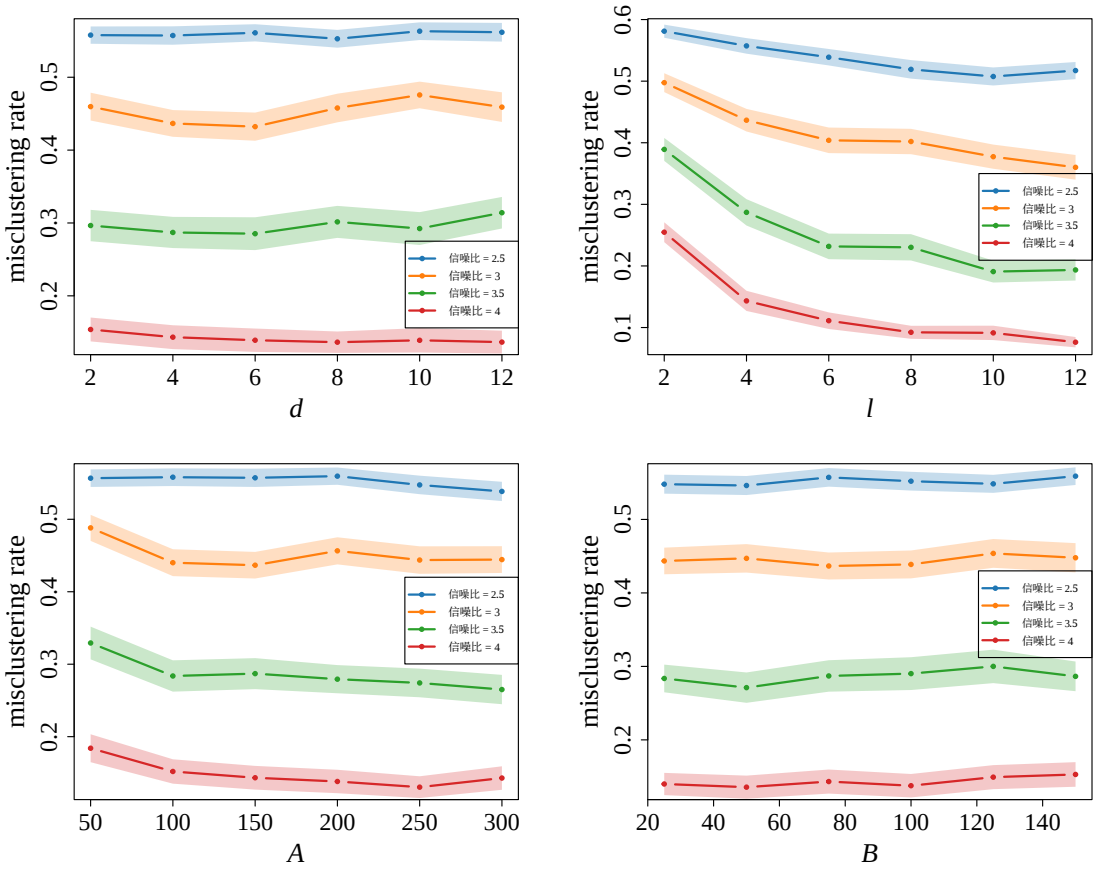

Figure 1: Average misclustering rate over 200 repetitions in our anisotropic Gaussian semi-supervised learning problem with $n = 250$, $p = 600$, $s = 4$, $K = 3$, $\gamma = 0.05$, SNR $\in \{2.5, 3, 3.5, 4\}$ and $\Sigma_{\mathrm{w}} = V\Lambda V^\top$, where $\Lambda \in \mathbb{R}^{p\times p}$ is diagonal with independent Unif$[0, 2]$ diagonal entries and V is independent of Λ, and generated according to the Haar measure on $\mathbb{O}^{p\times p}$. For each of the four panels, we fix three of $d = 4$, $\ell = 4$, $A = 150$, $B = 75$, and vary the remaining one. The shaded regions represent interpolated 95% confidence intervals at each of the points.

the misclustering rate will eventually deteriorate. Nevertheless, the bottom-right panel of Figure 1 indicates that the gain in increasing the probability of including all signal variables may outweigh the penalty of also including more noise variables—as expected, this effect is larger when the SNR is larger. For simplicity we choose $\ell = s_0$ in our remaining simulations, though we recommend practitioners err on the side of choosing larger ℓ.

4.2 与现有方法的比较

In this subsection, we compare the empirical performance of the `Sharp-SSL` algorithm in high-dimensional clustering tasks with several existing approaches. We apply the `Sharp-SSL` algorithm using the EM algorithm of Algorithm 3 as a base procedure, with input parameters $A = 150$, $B = 75$, $d = \ell = s_0$ as discussed in Section 4.1, and our final estimated cluster labels are then obtained as described there.

我们将该 `Sharp-SSL` 算法与五种备择高维聚类方法进行比较：谱聚类（例如von Luxburg，2007年），ℓ_1惩罚方法（由

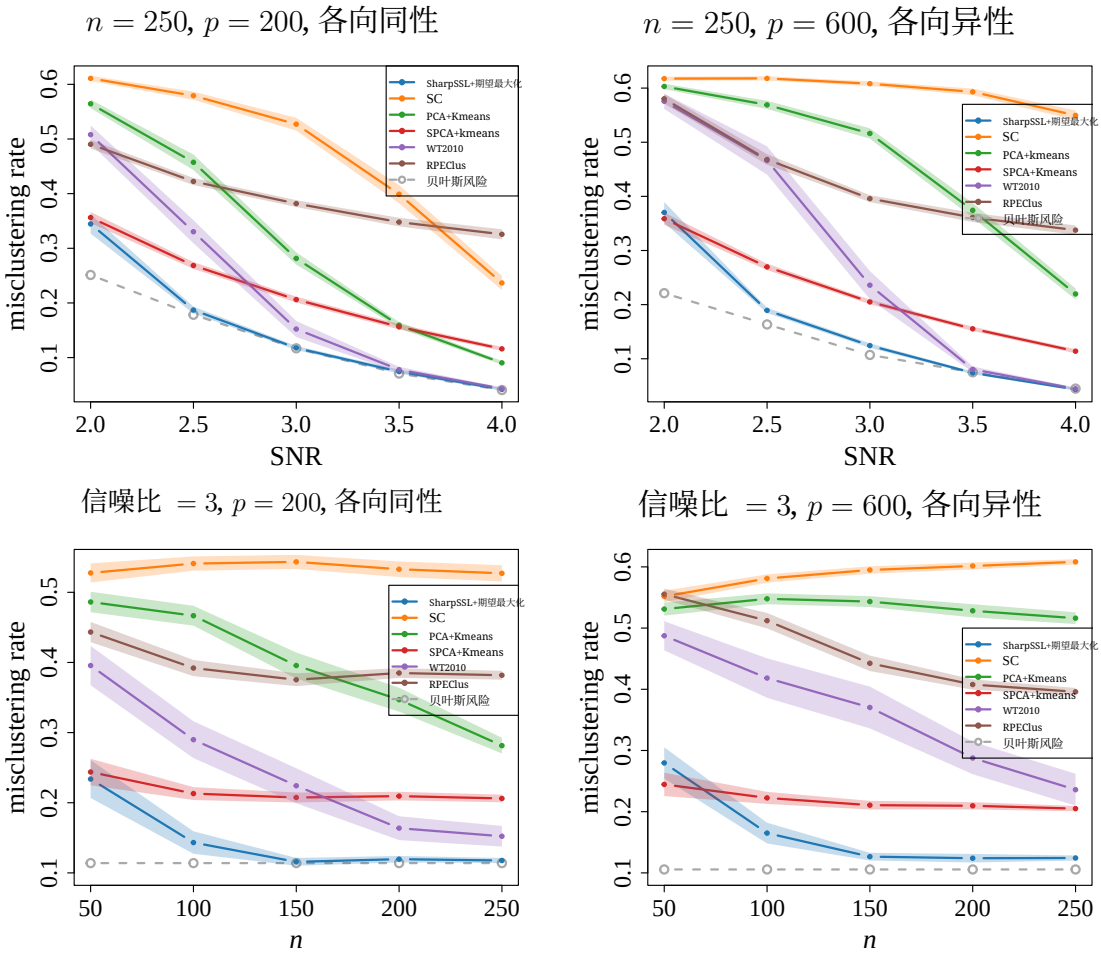

Figure 2: Average misclustering rate over 100 repetitions using `Sharp-SSL` followed by the EM algorithm, as well as using the other methods from Section 4.2. Data are generated from the normal mixture distribution described at the beginning of Section 4 with $K = 3$ and $p = 200$ (left) as well as $p = 600$ (right). The three cluster means are given by $\mu_1 = a(1, 1, 0, \mathbf{0}_{p-3})$, $\mu_2 = a(-1, 0, 1, \mathbf{0}_{p-3})$ and $\mu_3 = a(0, -1, -1, \mathbf{0}_{p-3})$, where the scale a is chosen such that their pairwise distances are all equal to SNR. For isotropic settings (left), $\Sigma_{\mathrm{w}} = I_p$; for anisotropic settings (right), $\Sigma_{\mathrm{w}} = V\Lambda V^\top$, where $\Lambda \in \mathbb{R}^{p\times p}$ is diagonal with independent Unif$[0, 2]$ diagonal entries and V is independent of Λ, and sampled from the Haar measure on $\mathbb{O}^{p\times p}$. The Bayes risk is shown as the gray dashed line. In the top panels, $n = 250$ and the SNR varies; in the bottom panels, SNR $= 3$ and n varies. The shaded regions represent interpolated 95% confidence intervals at each of the points.

Witten and Tibshirani (2010年）以及 RPEClus 算法Anderlucci、Fortunato 和 Montanari（2022年），以及另外两种方法。这些方法与 Sharp-SSL一样，在低维聚类算法之前应用降维。

In more detail, the spectral clustering approach first constructs a J-nearest neighbour graph adjacency matrix $A = (A_{i,i'})_{i,i' \in [n]} \in \{0,1\}^{n \times n}$, where $A_{i,i'} := 1$ if either X_i is one of the $J = 10$ nearest neighbours of $X_{i'}$ in Euclidean distance or vice versa, and $A_{i,i'} := 0$ otherwise. It then computes an $n \times K$ matrix of eigenvectors associated with the K smallest nonzero eigenvalues of the Laplacian matrix $L := D - A$, where $D \in \mathbb{R}^{n \times n}$ is a diagonal matrix with diagonal entries $D_{i,i} := \sum_{i' \in [n]} A_{i,i'}$. The final step is to apply the K-means clustering algorithm (Lloyd, 1982), as implemented in the `kmeans` base `R` function with 100 random initializations, to the rows of L with the oracle choice of K.

The Witten and Tibshirani (2010) method, which is implemented in the `sparcl R` package, determines the estimated cluster memberships by maximizing a coordinatewise-weighted between-cluster sum of squares criterion, subject to an ℓ_1 constraint on the weights. A permutation approach is used to select the ℓ_1 tuning parameter.

In the `RPEClus` algorithm of Anderlucci, Fortunato and Montanari (2022), we generate B random orthogonal projections and incorporate the d-dimensional projected data as covariates for a linear regression with the orthogonal complement of the projected data as the response. We then use the Bayesian Information Criteria (BIC) from both an application of the EM algorithm to the projected data and the aforementioned regression to identify good projections, and aggregate using the consensus clustering technique of Dimitriadou, Weingessel and Hornik (2002) over the best B^* projections chosen according to the sum of the BIC scores. Following the recommendation of Anderlucci, Fortunato and Montanari (2022), we took $B = 1000$ and $B^* = 100$ as well as $d = s_0$. It turned out that this approach had a misclustering rate almost identical to that of a random guess, likely because it did not leverage the sparsity of the signal. We therefore modified this method by generating random axis-aligned projections instead of orthogonal ones, and report this version in our comparison.

The first of the two-stage approaches applies principal component analysis (PCA) to project the data into the oracle choice of $K - 1$ dimensions (the dimension of the space spanned by the K cluster means); the second uses sparse principal component analysis (SPCA), as implemented in the `SPCAvRP` algorithm (Gataric, Wang and Samworth, 2020) with inputs $A = 600$, $B = 200$, and the oracle choices $d = \ell = s_0$, to project into s_0 dimensions. Thereafter, both algorithms apply K-means to the projected data as above. We also explored the option of replacing the K-means steps in these latter algorithms with the EM algorithm, but observed very little difference, so do not report these results here.

给定真实标签 $y_1, \ldots\ldots, y_n \in [K]$ 和聚类算法得到的估计标签 $\hat{y}_1, \ldots\ldots, \hat{y}_n \in [K]$，我们通过其错误聚类率来衡量该算法的性能，定义为[4]

$$L(\{y_1, \ldots, y_n\}, \{\hat{y}_1, \ldots, \hat{y}_n\}) := \min_{\sigma \in \mathcal{S}_K} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{\sigma(\hat{y}_i) \neq y_i\}},$$

其中 S_K 是 $[K]$的全体排列组成的群。特别是，图2展示了上述不同高维聚类算法在 100 次蒙特卡洛重复试验中得到的平均错误聚类率。在两个不同的维度 $p \in \{200, 600\}$,

[4]这里取所有排列中的最小值，因为重要的只是聚类分组，而不是标签本身。

isotropic and anisotropic settings, and for different values of n and SNR, we see a consistent picture of the `Sharp-SSL` algorithm combined with EM producing the lowest misclustering rates, often by a large margin. Indeed, for all but the smallest sample sizes or values of SNR, the `Sharp-SSL+EM` algorithm nearly attains the Bayes risk in all of the problems considered here.

4.3 观测比例对错误聚类率的影响

我们方法的主要优势之一是，它提供了一个统一框架，可以利用任意比例的有标签观测进行分类或聚类。在本小节中，我们将通过改变观测标签的比例来探究算法的性能。

Recall that we have two different options for the way in which we implement the `Sharp-SSL` algorithm to estimate the set of signal coordinates: we can either use only the labeled data, as in the supervised learning approach of Algorithm 2, or we can try to leverage in addition the unlabeled data via the semi-supervised EM approach of Algorithm 3. In the extreme case of this latter version, we have no labeled data, so the algorithm is unsupervised. In Figure 3 we compare the performance of these three methods in both high- and low-dimensional versions of the normal mixture distribution data generation mechanism described at the beginning of Section 4 as the proportion γ of observed labels varies.

More precisely, for the semi-supervised and unsupervised algorithms, we adopt the same implementation of `Sharp-SSL` as described at the beginning of Section 4.2. The supervised algorithm is very similar, but applies Algorithm 2 in place of Algorithm 3 to select coordinates, and obtains final predicted labels by applying LDA again on the projected labeled data. In cases where the proportion of labeled data was so small that the convex hull of the projected labeled data was less than full-dimensional for every class, we forced Algorithm 2 to return a zero matrix (this only happened when γ was very small).

The top panels of Figure 3 present the results in high-dimensional settings with $p \in \{200, 600\}$. Since the unsupervised approach has no access to the labels, it has constant misclustering rate. The performance of the semi-supervised approach is always at least as good as that of the unsupervised algorithm, and improves as γ increases. In other words, it effectively leverages the additional information provided by the class labels. When γ is very small, the supervised algorithm—which ignores the unlabeled data—is inaccurate, as it has very little data to work with. On the other hand, its performance also improves as γ increases, and once around 5% of our data are labeled, it outperforms the unsupervised algorithm. Further, it essentially matches the semi-supervised approach when about a third of the data are labeled. We truncate the plot at $\gamma = 1/2$ to ensure that we have enough test data on which to compute the misclustering rate.

在图3的下方图中，我们考察上述三种算法在两种不同 `SNR`值的低维设置中的表现，以进一步理解上一段所述的现象。这里，我们取$K = 2$ 并报告平均弗罗贝尼乌斯范数损失

$$\mathcal{L}\big((\hat{\mu}_1, \hat{\mu}_2), (\mu_1, \mu_2)\big) := \min\big\{\|(\hat{\mu}_1, \hat{\mu}_2) - (\mu_1, \mu_2)\|_{\mathrm{F}}, \|(\hat{\mu}_2, \hat{\mu}_1) - (\mu_1, \mu_2)\|_{\mathrm{F}}\big\}$$

的估计均值，重复 100 次。如果有标签数据不足以运行算法 2，则我们输出 $\hat{}\mu_1 = \hat{}\mu_2 = \mathbf{0}_p$。我们看到，即使在这些低-

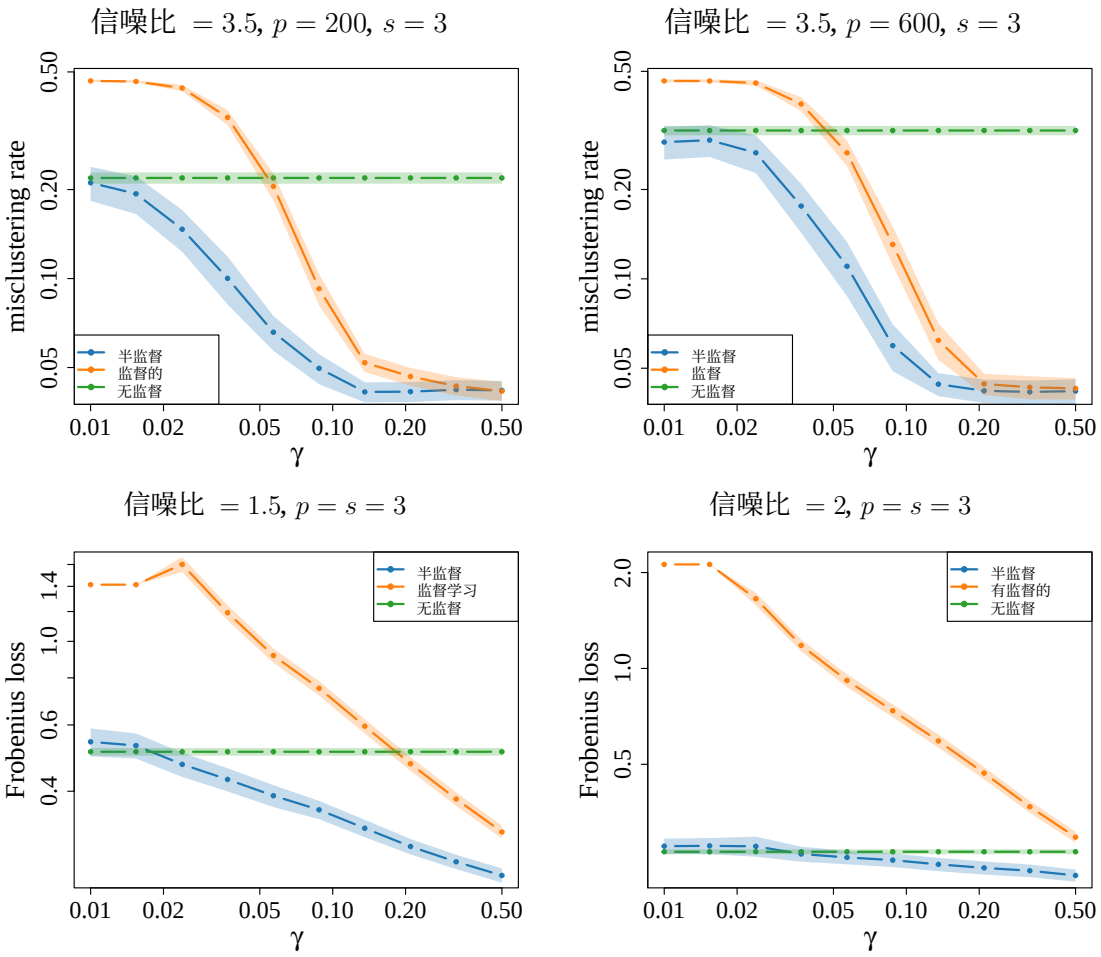

Figure 3: Effect of label fraction on performance of supervised, semi-supervised and unsupervised `Sharp-SSL` learning methods. Data are generated from the normal mixture distribution described at the beginning of Section 4 with $K = 2$ and $\Sigma_w = I_p$, $\mu_1 = -\mu_2 = a(\mathbf{1}_s, \mathbf{0}_{p-s})^\top \in \mathbb{R}^p$, where a is chosen such that $\|\mu_1 - \mu_2\| = \text{SNR}$. Bottom: average Frobenius loss of estimating the $(\mu_1, \mu_2) \in \mathbb{R}^{p\times 2}$ over 100 repetitions via the semi-supervised approach (Algorithm 3), supervised approach (Algorithm 2) and unsupervised approach (Algorithm 3 without using the labels). Top: average misclustering rate over 100 repetitions from applying the above three methods as base algorithms in Algorithm 1. The shaded regions represent interpolated 95% confidence intervals at each of the points.

dimensional problems, a similar picture emerges: if the proportion of labeled data is small, then the unsupervised algorithm outperforms the supervised one, but this situation may be reversed when γ is larger. The semi-supervised algorithm is able to leverage both the unlabeled and labeled data to obtain the best of both worlds. These empirical observations agree with our theory from Section 3, in particular in the way in which Theorem 6 bounds the accuracy of mean estimation for the semi-supervised algorithm by a minimum of a term that does not depend on γ and one that decreases as γ increases. It appears that the switch in the minimum occurs around $\gamma = 0.02$ in these examples.

4.4 实证数据分析

In this subsection we apply the `Sharp-SSL` algorithm, as well as several competing methods, to the gene expression data set from Alon et al. (1999), which contains observations on 62 patients. A preprocessed version of the data can be downloaded from the R package 'datamicroarray' (Ramey, 2016), with a total of 2000 features (genes) measured on 40 patients with colon tumors and 22 without tumors. We first exclude 9 genes to remove perfect collinearity and then standardize each of the remaining $p = 1991$ columns of the dataset to have unit variance.

We apply the `Sharp-SSL` algorithm using EM (Algorithm 3) as the base procedure, with input parameters $A = 150$, $B = 75$, $d = \ell = 5$. In addition to our approach (`Sharp-SSL+EM`), we also compare the performance of the spectral clustering (`SC`) method, the Witten and Tibshirani (2010) method (`WT2010`, as well as four two-stage methods (`PCA+Kmeans`, `PCA+EM`, `SPCA+Kmeans`, `SPCA+EM`), where we first reduce dimension of the data to a 5-dimensional subspace using either PCA or SPCA and then apply either the EM algorithm or K-means clustering on the low-dimensional data. For SPCA, we use the `SPCAvRP` algorithm (Gataric, Wang and Samworth, 2020) with inputs $A = 600$, $B = 200$ and $d = \ell = 5$. The true labels are hidden to all algorithms and are only used to evaluate the final misclustering rate.

Over 100 Monte Carlo repetitions of the randomized algorithms, the `Sharp-SSL+EM` method had an average misclustering rate of 28.8%, whereas all other competitors had a misclustering rate above 40%, as can be seen from the right-hand data points in Figure 4. To investigate this performance further, we applied each method to a subset of the features. These were constructed from the top $\ell = 5$ genes identified through `Sharp-SSL`, together with $m = 0, 10, 50, 200$ and 600 randomly chosen genes from the remaining 1986. The results are presented as the other data points in Figure 4. We see that the improved performance of the `Sharp-SSL+EM` relative to the other methods persists, even when only a small number of potentially non-discriminative covariates are present. When $m = 0$, `Sharp-SSL+EM` has a slight disadvantage as other algorithms benefit from the ensemble effect of combining two different learning methods; nevertheless it remains competitive. This reinforces the point that the primary contribution of the `Sharp-SSL` algorithm is to identify signal coordinates that are helpful for semi-supervised learning, and once this task has been accomplished, a variety of low-dimensional procedures are available to the practitioner.

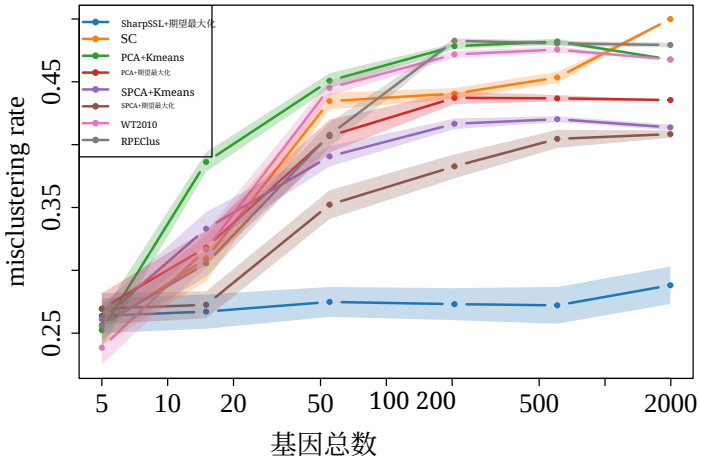

Figure 4: Average misclustering rate (over 100 repetitions for randomized algorithms) for the colon tumor data, using `Sharp-SSL` followed by the EM algorithm, as well as the other methods described in Section 4.4. The right-hand data points plot the average misclustering rate on the full data set. The other points were obtained by applying each method to a subset of genes formed from the top five genes identified by `Sharp-SSL` together with randomly sampled genes. The shaded regions represent interpolated 95% confidence intervals at each of the points.

5 主要结果的证明

5.1 命题 证明 1

The assumption that the convex hull of $\nu_1, \ldots, \nu_K$ is $(K-1)$-dimensional implies that Σ_{b} is of rank $K-1$. Define $A := \Sigma_{\mathrm{w}}^{-1/2}\Sigma_{\mathrm{b}}\Sigma_{\mathrm{w}}^{-1/2} \in \mathbb{R}^{p\times p}$, which has rank $K-1$. Given $V \in \mathbb{O}^{p\times d}$, we can find $Q \in \mathbb{O}^{p\times d}$ with the same column span as that of $\Sigma_{\mathrm{w}}^{1/2}V$ and let $R := Q^\top \Sigma_{\mathrm{w}}^{1/2} V \in \mathbb{R}^{d\times d}$, so that R is invertible, and $\Sigma_{\mathrm{w}}^{1/2}V = QR$. We observe that

$$\mathrm{tr}\{(V^\top \Sigma_{\mathrm{w}} V)^{-1}(V^\top \Sigma_{\mathrm{b}} V)\} = \mathrm{tr}\{(R^\top R)^{-1}(R^\top Q^\top A Q R)\} = \mathrm{tr}(Q^\top A Q).$$

Thus, $J(V; \Sigma_{\mathrm{b}}, \Sigma_{\mathrm{w}})$ depends on V only through the column space of $\Sigma_{\mathrm{w}}^{1/2}V$. Moreover, $\mathrm{tr}(Q^\top AQ)$ is maximized when Q, or equivalently $\Sigma_{\mathrm{w}}^{1/2}V$, spans a d-dimensional space that contains the $(K-1)$-dimensional eigenspace corresponding to the non-zero eigenvalues of A. Note that if for some $v \in \mathbb{R}^p \setminus \{0\}$ and $\lambda \geq 0$, we have $Av = \lambda v$, then $\Sigma_{\mathrm{w}}^{-1}\Sigma_{\mathrm{b}}\Sigma_{\mathrm{w}}^{-1/2}v = \Sigma_{\mathrm{w}}^{-1/2}Av = \lambda \Sigma_{\mathrm{w}}^{-1/2}v$, so $\Sigma_{\mathrm{w}}^{-1/2}v$ is an eigenvector of $\Sigma_{\mathrm{w}}^{-1}\Sigma_{\mathrm{b}}$ with eigenvalue λ. Hence V maximizes $J(V; \Sigma_{\mathrm{b}}, \Sigma_{\mathrm{w}})$ over $\mathbb{O}^{p\times d}$ if and only if V spans a d-dimensional space that contains the $(K-1)$-dimensional eigenspace corresponding to the $K-1$ non-zero eigenvalues of $\Sigma_{\mathrm{w}}^{-1}\Sigma_{\mathrm{b}}$. Finally, for any $v \in \mathbb{R}^p \setminus \{0\}$,

$$v^\top \Sigma_{\mathrm{w}}^{-1/2}\Sigma_{\mathrm{b}}\Sigma_{\mathrm{w}}^{-1/2}v = \sum_{k=1}^{K} \pi_k v^\top \Sigma_{\mathrm{w}}^{-1/2}(\nu_k - \nu)(\nu_k - \nu)^\top \Sigma_{\mathrm{w}}^{-1/2} v \neq 0$$

if and only if $v^\top \Sigma_{\mathrm{w}}^{-1/2}(\nu_k - \nu) \neq 0$ for some $k \in [K]$. Thus, the eigenspace corresponding to the non-zero eigenvalues of A is spanned by $\big(\Sigma_{\mathrm{w}}^{-1/2}(\nu_k - \nu) : k \in [K]\big)$, and so the eigenspace corresponding to non-zero eigenvalues of $\Sigma_{\mathrm{w}}^{-1}\Sigma_{\mathrm{b}}$ is spanned by $\big(\Sigma_{\mathrm{w}}^{-1}(\nu_k - \nu) : k \in [K]\big)$.

5.2 定理证明2

We write $S^{a,b} := \{j \in [p] : (P^{a,b,\top}P^{a,b})_{j,j} = 1\}$. For any $S = \{j_1, \dots, j_d\}^\top \in \binom{[p]}{d}$, we identify the set S with the sequence $j_{i_1} < \dots < j_{i_d}$ sorted in increasing order; and with a slight abuse of notation, we will use S to refer to either object, which will always be clear depending on the context. We define $P^S \in \mathcal{P}_d$ by $(P^S)_{\ell,j} := \mathbb{1}_{\{j=j_\ell\}}$, so that $P^{S,\top}P^S = \mathrm{diag}\big((\mathbb{1}_{\{j\in S\}})_{j\in[p]}\big)$. Define $Q^S := (P^S\Sigma_{\mathrm{w}}P^{S,\top})^{-1}P^S\Sigma_{\mathrm{b}}P^{S,\top} \in \mathbb{R}^{d\times d}$ and $\hat{Q}^S := \psi\big((P^S X_i, Y_i)_{i\in[n]}\big) \in \mathbb{R}^{d\times d}$. Note that $\hat{Q}^{a,b} = \hat{Q}^{S^{a,b}}$ in this notation, and we will similarly denote $Q^{a,b} := Q^{S^{a,b}}$ for simplicity. Recalling the definition of Ω from (8), in our new notation, and recalling that ψ is permutation-equivariant, we can write

$$\Omega = \left\{ \max_{S\in\binom{[p]}{d}} \big\|\hat{Q}^S - Q^S\big\|_{\mathrm{op}} < \frac{\gamma_{\min}}{4(K-1)} \right\},$$

并且有 $\mathbb{P}(\Omega) \geq 1-\varepsilon$ 由（7）。在证明余下部分中，我们将始终在事件 Ω 上进行讨论，并且还假设 $\gamma_{\min} > 0$，否则结论显然成立。

根据Weyl不等式（Weyl，1912年；Stewart和Sun，1990，推论IV.4.9），我们在 Ω 上有

$$|\mathrm{tr}(\hat{Q}^S) - \mathrm{tr}(Q^S)| \leq (K-1)\|\hat{Q}^S - Q^S\|_{\mathrm{op}} \leq \frac{\gamma_{\min}}{4}.$$

On the other hand, we note that $\mathrm{tr}(Q^S) = \sum_{j\in S\cap S_0}(\Sigma_{\mathrm{w}}^{-1}\Sigma_{\mathrm{b}})_{j,j}$. Therefore, by the triangle inequality, for any $S, S' \in \binom{[p]}{d}$ such that $S\cap S_0$ is a proper subset of $S'\cap S_0$, we have on Ω that

$$\mathrm{tr}(\hat{Q}^S) - \mathrm{tr}(\hat{Q}^{S'}) \leq \frac{\gamma_{\min}}{2} - \sum_{j\in(S'\setminus S)\cap S_0} (\Sigma_{\mathrm{w}}^{-1}\Sigma_{\mathrm{b}})_{j,j} < 0. \tag{14}$$

Fix $a \in [A]$, and for any $\tilde{j} \in [p]$, define $q_{\tilde{j}} := \mathbb{P}\big(\tilde{j} \in S^{a,b^*(a)} \mid (X_i, Y_i)_{i\in[n]}\big)$. Now fix some $j \in S_0$ and $j' \in [p]\setminus S_0$. We claim that

$$(q_j - q_{j'})\mathbb{1}_\Omega \geq 0. \tag{15}$$

˜为验证这一结论，对 $j \in \{j, j'\}$ 和 $b \in [B]$ 定义集合

$\mathcal{S}_{b,\tilde{j}} := \{(S^{a,1}, \dots, S^{a,B}) : b^*(a) = b, \tilde{j} \in S^{a,b}\}$ and $\mathcal{S}_b := \big\{(S^{a,1}, \dots, S^{a,B}) : b^*(a) = b\big\}$.
设 $f\colon \binom{[p]}{d} \to \binom{[p]}{d}$ 是一个映射，其定义为

$$f(S) := \begin{cases} (S\setminus\{j'\})\cup\{j\} & \text{if } j\notin S \text{ and } j'\in S \\ S & \text{otherwise.} \end{cases}$$

If $j \notin S^{a,b}$ and $j' \in S^{a,b}$, then $f(S^{a,b})\cap S_0 = (S^{a,b}\cup\{j\})\cap S_0 = (S^{a,b}\cap S_0)\cup\{j\}$, so $S^{a,b}\cap S_0$ is a proper subset of $f(S^{a,b})\cap S_0$; on the other hand, if either $j \in S^{a,b}$ or $j, j' \notin S^{a,b}$, then $f(S^{a,b}) = S^{a,b}$. It follows by (14) that on Ω we have

$$\mathrm{tr}(\hat{Q}^{S^{a,b}}) \leq \mathrm{tr}(\hat{Q}^{f(S^{a,b})}). \tag{16}$$

现在令 $F\colon \mathcal{S}_{b,j'} \to \mathcal{S}_{b,j}$ 定义为

$$F(S^{a,1}, \dots, S^{a,B}) := \big(S^{a,1}, \dots, S^{a,b-1}, f(S^{a,b}), S^{a,b+1}, \dots, S^{a,B}\big).$$

We claim that F is both well-defined and injective on Ω. For the first of these claims, we note that since $j' \in S^{a,b}$, we must have $j \in f(S^{a,b})$. Moreover, if $(S^{a,1}, \dots, S^{a,B}) \in \mathcal{S}_{b,j'}$, then $b^*(a) = b$. But (16) holds on Ω, so $\big(S^{a,1}, \dots, S^{a,b-1}, f(S^{a,b})\ S^{a,b+1}, \dots, S^{a,B}\big) \in \mathcal{S}_{b,j}$. Hence F is well-defined. For the second claim, suppose that $S_1, S_2 \in \binom{[p]}{d}$ are such that $j' \in S_1 \cap S_2$ and $f(S_1) = f(S_2)$. If $j \in S_1 \cap S_2$, then $S_1 = f(S_1) = f(S_2) = S_2$; if $j \in S_1$ but $j \notin S_2$, then $j \in f(S_2) \setminus f(S_1)$, a contradiction. Similarly, we cannot have $j \notin S_1$ but $j \in S_2$. Finally, if $j \in S_1^{\mathrm{c}} \cap S_2^{\mathrm{c}}$, then $S_1 = \big(f(S_1) \setminus \{j\}\big) \cup \{j'\} = \big(f(S_2) \setminus \{j\}\big) \cup \{j'\} = S_2$. We deduce that f is injective on $\{S : j' \in S\}$. Since $j' \in S^{a,b}$ for $(S^{a,1}, \dots, S^{a,B}) \in \mathcal{S}_{b,j'}$, this establishes the injectivity of F. In particular, $|\mathcal{S}_{b,j'}| \le |\mathcal{S}_{b,j}|$. Consequently, on Ω, we have for all $b \in [B]$ that

$$\begin{aligned}
\mathbb{P}\big(j \in S^{a,b^*(a)} \mid (X_i, Y_i)_{i\in[n]}, b^*(a) = b\big) &= \frac{\mathbb{P}\big(j \in S^{a,b^*(a)}, b^*(a) = b \mid (X_i, Y_i)_{i\in[n]}\big)}{\mathbb{P}\big(b^*(a) = b \mid (X_i, Y_i)_{i\in[n]}\big)} = \frac{|\mathcal{S}_{b,j}|}{|\mathcal{S}_b|} \\
&\ge \frac{|\mathcal{S}_{b,j'}|}{|\mathcal{S}_b|} = \frac{\mathbb{P}\big(j' \in S^{a,b^*(a)}, b^*(a) = b \mid (X_i, Y_i)_{i\in[n]}\big)}{\mathbb{P}\big(b^*(a) = b \mid (X_i, Y_i)_{i\in[n]}\big)} \\
&= \mathbb{P}\big(j' \in S^{a,b^*(a)} \mid (X_i, Y_i)_{i\in[n]}, b^*(a) = b\big),
\end{aligned}$$

这意味着断言（15）。我们指出，（15）的一个推论是，由于 $d \ge s_0$，我们在 Ω 上有

$$q_j \ge \frac{\sum_{\tilde{j}\in([p]\setminus S_0)\cup\{j\}} q_{\tilde{j}}}{p - s_0 + 1} = \frac{d - \sum_{\tilde{j}\in S_0\setminus\{j\}} q_{\tilde{j}}}{p - s_0 + 1} \ge \frac{d - s_0 + 1}{p - s_0 + 1} \ge \frac{1}{p}. \tag{17}$$

再次固定 $j \in S_0$ 和 $j' \notin S_0$，我们在 $\Omega \cap \{j \in S^{a,b^*(a)}\}$ 上观察到，

$$\begin{aligned}
\frac{3}{4}\gamma_{\min} &\le \big[P^{a,b^*(a),\top} Q^{a,b^*(a)} P^{a,b^*(a)}\big]_{j,j} - \big\|\hat{Q}^{a,b^*(a)} - Q^{a,b^*(a)}\big\|_{\mathrm{op}} \le \big[P^{a,b^*(a),\top} \hat{Q}^{a,b^*(a)} P^{a,b^*(a)}\big]_{j,j} \\
&\le [P^{a,b^*(a),\top} Q^{a,b^*(a)} P^{a,b^*(a)}]_{j,j} + \big\|\hat{Q}^{a,b^*(a)} - Q^{a,b^*(a)}\big\|_{\mathrm{op}} \le \frac{5}{4}\gamma_{\max},
\end{aligned}$$

and similarly on $\Omega \cap \{j' \in S^{a,b^*(a)}\}$ that $\big|\big[P^{a,b^*(a)\,\top} \hat{Q}^{a,b^*(a)} P^{a,b^*(a)}\big]_{j',j'}\big| \le \gamma_{\min}/4$. Recall also that $\big[P^{a,b^*(a)\,\top} \hat{Q}^{a,b^*(a)} P^{a,b^*(a)}\big]_{\tilde{j},\tilde{j}} = 0$ for all $\tilde{j} \notin S^{a,b^*(a)}$. Combining the above bounds on the diagonal entries of $\hat{Q}^{a,b^*(a)}$ with (15) and (17), we have on Ω that

$$\begin{aligned}
&\mathbb{E}\big(\big[P^{a,b^*(a),\top} \hat{Q}^{a,b^*(a)} P^{a,b^*(a)}\big]_{j,j} - \big[P^{a,b^*(a),\top} \hat{Q}^{a,b^*(a)} P^{a,b^*(a)}\big]_{j',j'} \mid (X_i, Y_i)_{i\in[n]}\big) \\
&\quad = \mathbb{E}\big(\big[P^{a,b^*(a),\top} \hat{Q}^{a,b^*(a)} P^{a,b^*(a)}\big]_{j,j} \mathbb{1}_{\{j\in S^{a,b^*(a)}\}} \\
&\qquad\qquad - \big[P^{a,b^*(a),\top} \hat{Q}^{a,b^*(a)} P^{a,b^*(a)}\big]_{j',j'} \mathbb{1}_{\{j'\in S^{a,b^*(a)}\}} \mid (X_i, Y_i)_{i\in[n]}\big) \\
&\quad \ge \frac{q_j \gamma_{\min}}{2} \ge \frac{\gamma_{\min}}{2p}.
\end{aligned} \tag{18}$$

Now, let a, j, j' be freely varying again. Since $\hat{w}_j = A^{-1} \sum_{a\in[A]} [P^{a,b^*(a)\,\top} \hat{Q}^{a,b^*(a)} P^{a,b^*(a)}]_{j,j}$, on Ω we have for any $j \in S_0$ and $j' \notin S_0$ that $\mathbb{E}\big(\hat{w}_j - \hat{w}_{j'} \mid (X_i, Y_i)_{i\in[n]}\big) \ge \gamma_{\min}/(2p)$

由（18）。由于 $\ell \geq s_0$，根据Hoeffding不等式可知，在 Ω 上，

$$\begin{aligned}
\mathbb{P}\big(S_0 \not\subseteq \hat{S} \mid (X_i, Y_i)_{i\in[n]}\big) &\leq \mathbb{P}\Big(\min_{j\in S_0} \hat{w}_j \leq \max_{j'\notin S_0} \hat{w}_{j'} \;\Big|\; (X_i, Y_i)_{i\in[n]}\Big) \\
&\leq \sum_{j\in S_0} \mathbb{P}\Big\{\hat{w}_j - \mathbb{E}(\hat{w}_j \mid (X_i, Y_i)_{i\in[n]}) \leq -\frac{\gamma_{\min}}{4p} \;\Big|\; (X_i, Y_i)_{i\in[n]}\Big\} \\
&\quad + \sum_{j\notin S_0} \mathbb{P}\Big\{\hat{w}_j - \mathbb{E}(\hat{w}_j \mid (X_i, Y_i)_{i\in[n]}) \geq \frac{\gamma_{\min}}{4p} \;\Big|\; (X_i, Y_i)_{i\in[n]}\Big\} \\
&\leq p \exp\Big\{-\frac{A}{2}\Big(\frac{\gamma_{\min}}{4p}\Big)^2 \Big/ \Big(\frac{5\gamma_{\max}}{4}\Big)^2\Big\} \leq p e^{-A\gamma_{\min}^2/(50p^2\gamma_{\max}^2)},
\end{aligned}$$

如所需。

5.3 定理的证明 3

定理证明的主要要素 3是以下命题，它控制了分类问题中样本类间和类内协方差矩阵收敛到各自总体版本的收敛速率。

Proposition 8 (Rate of convergence for LDA). *Suppose that* $(Z_1, Y_1) \ldots, (Z_n, Y_n) \in \mathbb{R}^d \times [K]$ *are independent and identically distributed data-label pairs, such that* $\mathbb{P}(Y_1 = k) = \pi_k$ *and* $Z_1 \mid Y_1 = k \sim \mathcal{N}_d(\mu_k, \Sigma_{\mathrm{w}})$ *for* $k \in [K]$. *Write* $\mu := \sum_{k\in[K]} \pi_k \mu_k$ *and* $\Sigma_{\mathrm{b}} := \sum_{k\in[K]} \pi_k (\mu_k - \mu)(\mu_k - \mu)^\top$ *and let* $\hat{\Sigma}_{\mathrm{w}}$ *and* $\hat{\Sigma}_{\mathrm{b}}$ *be computed as in* (3) *applied to* $(Z_1, Y_1) \ldots, (Z_n, Y_n)$. *If* $\|\mu_k - \mu\| \leq R_1$ *for all* $k \in [K]$ *and* $\|\Sigma_{\mathrm{w}}\|_{\mathrm{op}} \leq R_2$ *for some* $R_1, R_2 > 0$, *then for every* $\delta \in (0, 1/4]$, *we have with probability at least* $1 - \delta$ *that*

$$\|\hat{\Sigma}_{\mathrm{b}} - \Sigma_{\mathrm{b}}\|_{\mathrm{op}} \leq \frac{12R_2\{K + \log(8 \cdot 9^d/\delta)\}}{n} + \big(4R_1\sqrt{R_2} + R_1^2 + R_1\big)\sqrt{\frac{2\log(8 \cdot 9^d/\delta)}{n}},$$

$$\|\hat{\Sigma}_{\mathrm{w}} - \Sigma_{\mathrm{w}}\|_{\mathrm{op}} \leq \frac{4R_2\{K + \log(8 \cdot 9^d/\delta)\}}{n} + 4R_2\sqrt{\frac{\log(8 \cdot 9^d/\delta)}{n}}.$$

Proof. We first control the rate of convergence of $\hat{\Sigma}_{\mathrm{b}}$. For $n_k := \sum_{i=1}^n \mathbb{1}_{\{Y_i = k\}}$ and $\tilde{\mu} := \sum_{k\in[K]} (n_k/n)\mu_k$, we define $\tilde{\Sigma}_{\mathrm{b}}^{(1)} := \sum_{k=1}^K (n_k/n)(\mu_k - \mu)(\mu_k - \mu)^\top$ and $\tilde{\Sigma}_{\mathrm{b}}^{(2)} := \sum_{k=1}^K (n_k/n)(\mu_k - \tilde{\mu})(\mu_k - \tilde{\mu})^\top$. We have the following decomposition

$$\|\hat{\Sigma}_{\mathrm{b}} - \Sigma_{\mathrm{b}}\|_{\mathrm{op}} \leq \|\tilde{\Sigma}_{\mathrm{b}}^{(1)} - \Sigma_{\mathrm{b}}\|_{\mathrm{op}} + \|\tilde{\Sigma}_{\mathrm{b}}^{(2)} - \tilde{\Sigma}_{\mathrm{b}}^{(1)}\|_{\mathrm{op}} + \|\hat{\Sigma}_{\mathrm{b}} - \tilde{\Sigma}_{\mathrm{b}}^{(2)}\|_{\mathrm{op}}. \tag{19}$$

We will control the three terms on the right-hand side above separately. For the first term, we note that $\|\mu_k - \mu\| \leq R_1$ for all $k \in [K]$. Since $n_1, \ldots, n_K$ are functions of $Y_1, \ldots, Y_n$, we have by McDiarmid's inequality (see, e.g. Boucheron, Lugosi and Massart, 2013, Theorem 6.2) that with probability at least $1 - \delta/4$,

$$\|\tilde{\Sigma}_{\mathrm{b}}^{(1)} - \Sigma_{\mathrm{b}}\|_{\mathrm{op}} \leq R_1^2 \sum_{k=1}^K \Big|\frac{n_k}{n} - \pi_k\Big| \leq R_1^2\sqrt{\frac{2\log(4/\delta)}{n}}. \tag{20}$$

对于第二项，我们首先再次应用 McDiarmid不等式，以得出以至少 $1 - \delta/4$ 的概率，我们有

$$\|\tilde{\mu} - \mu\| \leq \sum_{k=1}^K \Big|\frac{n_k}{n} - \pi_k\Big| \|\mu_k - \mu\| \leq R_1\sqrt{\frac{2\log(4/\delta)}{n}}.$$

因此，我们以至少 $1-\delta/4$ 的概率有不中返

$$\|\tilde{\Sigma}_{\mathrm{b}}^{(2)}-\tilde{\Sigma}_{\mathrm{b}}^{(1)}\|_{\mathrm{op}} \leq 2\sum_{k=1}^{K}\frac{n_k}{n}\|(\mu_k-\mu)(\mu-\tilde{\mu})^\top\|_{\mathrm{op}}+\sum_{k=1}^{K}\frac{n_k}{n}\|(\mu-\tilde{\mu})(\mu-\tilde{\mu})^\top\|_{\mathrm{op}}$$
$$\leq 2R_1\|\mu-\tilde{\mu}\|+\|\mu-\tilde{\mu}\|^2 \leq 3R_1\|\mu-\tilde{\mu}\| \leq 3R_1^2\sqrt{\frac{2\log(4/\delta)}{n}}. \qquad (21)$$

Finally, for the third term, we write $\hat{\mu}_k := \sum_{i:Y_i=k} Z_i$ and note that $V_k := n_k^{1/2}\hat{\mu}_k$ satisfies $V_k \mid Y_1,\ldots,Y_n \sim \mathcal{N}_d(n_k^{1/2}\mu_k, \Sigma_{\mathrm{w}})$. Defining $N := (n_1^{1/2},\ldots,n_K^{1/2})^\top$ and $P := NN^\top/n \in \mathbb{R}^{K\times K}$, we may write

$$n\hat{\Sigma}_{\mathrm{b}} = \sum_{k=1}^{K} n_k\hat{\mu}_k\hat{\mu}_k^\top - n\hat{\mu}\hat{\mu}^\top = V^\top(I_K-P)V,$$

where $V := (V_1,\ldots,V_K)^\top$, and where $\hat{\mu} := n^{-1}\sum_{i=1}^n Z_i = \sum_{k=1}^K (n_k/n)\hat{\mu}_k$. By Lemma 17, we deduce that $n\hat{\Sigma}_{\mathrm{b}}$ conditional on $Y_1,\ldots,Y_n$ has a d-dimensional non-central Wishart distribution with $K-1$ degrees of freedom, covariance matrix Σ_{w} and non-centrality matrix $n\tilde{\Sigma}_{\mathrm{b}}^{(2)}$, which we denote as

$$n\hat{\Sigma}_{\mathrm{b}} \mid Y_1,\ldots,Y_n \sim \mathcal{W}_d(K-1, \Sigma_{\mathrm{w}}; n\tilde{\Sigma}_{\mathrm{b}}^{(2)});$$

在引理之前给出了一个形式化定义 17。对于任意固定的 $u \in \mathbb{S}^{d-1}$，根据Muirhead(2009年, 定理 10.3.6)，我们有

$$u^\top\hat{\Sigma}_{\mathrm{b}}u \mid Y_1,\ldots,Y_n \sim \frac{u^\top\Sigma_{\mathrm{w}}u}{n}\chi^2_{K-1}\biggl(\frac{nu^\top\tilde{\Sigma}_{\mathrm{b}}^{(2)}u}{u^\top\Sigma_{\mathrm{w}}u}\biggr),$$

where $\chi^2_r(\lambda)$ denotes a non-central chi-squared distribution with r degrees of freedom and non-centrality parameter λ. By Birgé (2001, Lemma 8.1), for every $\delta' \in (0,1/2]$, we have with probability at least $1-2\delta'$ conditional on $Y_1,\ldots,Y_n$, that

$$|u^\top(\hat{\Sigma}_{\mathrm{b}}-\tilde{\Sigma}_{\mathrm{b}}^{(2)})u| \leq \frac{u^\top\Sigma_{\mathrm{w}}u}{n}\biggl\{K+2\sqrt{\biggl(K+\frac{2nu^\top\tilde{\Sigma}_{\mathrm{b}}^{(2)}u}{u^\top\Sigma_{\mathrm{w}}u}\biggr)\log(1/\delta')}+2\log(1/\delta')\biggr\}$$
$$\leq \frac{u^\top\Sigma_{\mathrm{w}}u}{n}\bigl\{2K+3\log(1/\delta')\bigr\}+\sqrt{\frac{8u^\top\Sigma_{\mathrm{w}}uu^\top\tilde{\Sigma}_{\mathrm{b}}^{(2)}u\log(1/\delta')}{n}}$$
$$\leq \frac{3\bigl(K+\log(1/\delta')\bigr)\|\Sigma_{\mathrm{w}}\|_{\mathrm{op}}}{n}+\sqrt{\frac{8\|\Sigma_{\mathrm{w}}\|_{\mathrm{op}}\|\tilde{\Sigma}_{\mathrm{b}}^{(2)}\|_{\mathrm{op}}\log(1/\delta')}{n}}.$$

Let $\mathcal{N}$ be a $1/4$-net of the sphere $\mathbb{S}^{d-1}$, which can be chosen to have cardinality at most 9^d (Vershynin, 2012, Lemma 5.2). Hence, through a union bound, and taking $\delta' := \delta/(8\cdot 9^d)$, we have with probability at least $1-\delta/4$ conditional on $Y_1,\ldots,Y_n$ that

$$\|\hat{\Sigma}_{\mathrm{b}}-\tilde{\Sigma}_{\mathrm{b}}^{(2)}\|_{\mathrm{op}} \leq 2\max_{u\in\mathcal{N}}|u^\top(\hat{\Sigma}_{\mathrm{b}}-\tilde{\Sigma}_{\mathrm{b}}^{(2)})u|$$
$$\leq \frac{6R_2\bigl(K+\log(8\cdot 9^d/\delta)\bigr)}{n}+\sqrt{\frac{32R_1^2R_2\log(8\cdot 9^d/\delta)}{n}}. \qquad (22)$$

ˆ结合（19）、（20）、（21）和（22），我们得到，对 $\|\Sigma_{\mathrm{b}} - \Sigma_{\mathrm{b}}\|_{\mathrm{op}}$所需的界在概率至少为 $1 - 3\delta/4$的事件上成立。

We now turn to control $\|\hat{\Sigma}_{\mathrm{w}} - \Sigma_{\mathrm{w}}\|_{\mathrm{op}}$. Let $Q := \sum_{k\in[K]}(n_k^{-1}\mathbb{1}_{\{Y_i=k,Y_{i'}=k\}})_{i,i'=1}^n \in \mathbb{R}^{n\times n}$, so that

$$n\hat{\Sigma}_{\mathrm{w}} = \sum_{i=1}^n Z_i Z_i^\top - \sum_{k=1}^K n_k \hat{\mu}_k \hat{\mu}_k^\top = Z^\top (I - Q) Z,$$

where $Z := (Z_1, \ldots, Z_n)^\top$. It therefore follows again by Lemma 17 that $\hat{\Sigma}_{\mathrm{w}} \mid Y_1, \ldots, Y_n \sim n^{-1}\mathcal{W}_d(n-K, \Sigma_{\mathrm{w}})$. Another application of Muirhead (2009, Theorem 10.3.6) then yields for any $u \in \mathbb{S}^{d-1}$ that

$$u^\top \hat{\Sigma}_{\mathrm{w}} u \mid Y_1, \ldots, Y_n \sim \frac{u^\top \Sigma_{\mathrm{w}} u}{n} \chi^2_{n-K}.$$

根据Laurent和Massart(2000，引理 1），我们以至少 $1 - 2\delta' = 1 - \delta/(4\cdot 9^d)$ 的概率得到

$$|u^\top(\hat{\Sigma}_{\mathrm{w}} - \Sigma_{\mathrm{w}})u| \le \frac{R_2}{n}\big\{K + 2\sqrt{n\log(1/\delta')} + 2\log(1/\delta')\big\}.$$

再次对 $1/4$-网 $\mathcal{N}$ 应用并集界，我们得出结论：以至少 $1 - \delta/4$ 的概率，有

$$\|\hat{\Sigma}_{\mathrm{w}} - \Sigma_{\mathrm{w}}\|_{\mathrm{op}} \le 2\max_{u\in\mathcal{N}} |u^\top(\hat{\Sigma}_{\mathrm{w}} - \Sigma_{\mathrm{w}})u| \le \frac{4R_2}{n}\Big\{K + \sqrt{n\log(8\cdot 9^d/\delta)} + \log(8\cdot 9^d/\delta)\Big\},$$

如所期望的。 □

Proof of Theorem 3. Define $\delta := \binom{p}{d}^{-1}\varepsilon$. Since Algorithm 2 is permutation-equivariant, by a union bound, it suffices to show that for every $P \in \mathcal{P}_d$, with probability at least $1-\delta$, the desired upper bound holds for $\big\|\psi\big((PX_i, Y_i)_{i\in[n]}\big) - (P\Sigma_{\mathrm{w}}P^\top)^{-1}P\Sigma_{\mathrm{b}}P^\top\big\|_{\mathrm{op}}$. Recall that $n_k := \sum_{i=1}^n \mathbb{1}_{\{Y_i=k\}}$. Write $\Sigma_{\mathrm{w},P} := P\Sigma_{\mathrm{w}}P^\top$ and $\Sigma_{\mathrm{b},P} := P\Sigma_{\mathrm{b}}P^\top$, $\hat{\mu}_{k,P} := n_k^{-1}\sum_{i:y_i=k} PX_i$, $\hat{\mu}_P := n^{-1}\sum_{i=1}^n PX_i$ and

$$\hat{\Sigma}_{\mathrm{w},P} := \frac{1}{n}\sum_{i=1}^n (PX_i - \hat{\mu}_{Y_i,P})(PX_i - \hat{\mu}_{Y_i,P})^\top \quad \text{and} \quad \hat{\Sigma}_{\mathrm{b},P} := \sum_{k=1}^K \frac{n_k}{n}(\hat{\mu}_{k,P} - \hat{\mu}_P)(\hat{\mu}_{k,P} - \hat{\mu}_P)^\top.$$

ˆ 注意到，由于 $n \ge Kd+1$，我们有 $\max_{k\in[K]} n_k \ge d+1$，因此 $\Sigma_{\mathrm{w},P}$ 以概率 1 为正定。因此，根据三角不等式，对于每个 $P \in \mathcal{P}_d$，我们有

$$\begin{aligned}\big\|\psi\big((PX_i,\ Y_i)_{i\in[n]}\big) - (P\Sigma_{\mathrm{w}}P^\top)^{-1}P\Sigma_{\mathrm{b}}P^\top\big\|_{\mathrm{op}} &= \big\|\hat{\Sigma}^{-1}{}_{\mathrm{w}\,P}\hat{\Sigma}_{\mathrm{b},P} - \Sigma^{-1}{}_{\mathrm{w}\,P}\Sigma_{\mathrm{b},P}\big\|_{\mathrm{op}} \\ &\le \big\|\hat{\Sigma}^{-1}{}_{\mathrm{w},P}\hat{\Sigma}_{\mathrm{b},P} - \Sigma^{-1}{}_{\mathrm{w},P}\hat{\Sigma}_{\mathrm{b},P}\big\|_{\mathrm{op}} + \big\|\Sigma^{-1}{}_{\mathrm{w},P}\hat{\Sigma}_{\mathrm{b},P} - \Sigma^{-1}{}_{\mathrm{w},P}\Sigma_{\mathrm{b},P}\big\|_{\mathrm{op}}. \end{aligned} \tag{23}$$

根据命题 8和我们的假设，存在一个事件 Ω_P，其概率至少为 $1-\delta$，在该事件上

$$\begin{aligned}\|\hat{\Sigma}_{\mathrm{w},P} - \Sigma_{\mathrm{w},P}\|_{\mathrm{op}} &\le \frac{4R_2\{K + \log(8\cdot 9^d/\delta)\}}{n} + 4R_2\sqrt{\frac{\log(8\cdot 9^d/\delta)}{n}} \le \frac{1}{2R_2}. \\ \|\hat{\Sigma}_{\mathrm{b},P} - \Sigma_{\mathrm{b},P}\|_{\mathrm{op}} &\lesssim_{R_1,R_2} \frac{K + d + \log(1/\delta)}{n} + \sqrt{\frac{d + \log(1/\delta)}{n}} \lesssim_{R_2} 1.\end{aligned} \tag{24}$$

因此，对于 (23) 中的第一项，由 Weyl 不等式，在 Ω_P 上，我们有

$$\begin{aligned}
\|\hat{\Sigma}_{\mathrm{w},P}^{-1}\hat{\Sigma}_{\mathrm{b},P} - \Sigma_{\mathrm{w},P}^{-1}\hat{\Sigma}_{\mathrm{b},P}\|_{\mathrm{op}} &\le \|\Sigma_{\mathrm{w},P}^{-1} - \hat{\Sigma}_{\mathrm{w},P}^{-1}\|_{\mathrm{op}}\|\hat{\Sigma}_{\mathrm{b},P}\|_{\mathrm{op}} \\
&\le \|\Sigma_{\mathrm{w},P}^{-1}\|_{\mathrm{op}}\|\hat{\Sigma}_{\mathrm{w},P}^{-1}\|_{\mathrm{op}}\|\hat{\Sigma}_{\mathrm{b},P}\|_{\mathrm{op}}\|\hat{\Sigma}_{\mathrm{w},P} - \Sigma_{\mathrm{w},P}\|_{\mathrm{op}} \\
&\le \frac{\big(\|\Sigma_{\mathrm{b},P}\|_{\mathrm{op}} + \|\hat{\Sigma}_{\mathrm{b},P} - \Sigma_{\mathrm{b},P}\|_{\mathrm{op}}\big)\|\hat{\Sigma}_{\mathrm{w},P} - \Sigma_{\mathrm{w},P}\|_{\mathrm{op}}}{\lambda_{\min}(\Sigma_{\mathrm{w},P})\big(\lambda_{\min}(\Sigma_{\mathrm{w},P}) - \|\hat{\Sigma}_{\mathrm{w},P} - \Sigma_{\mathrm{w},P}\|_{\mathrm{op}}\big)} \\
&\le \frac{R_2\big(R_1^2 + \|\hat{\Sigma}_{\mathrm{b},P} - \Sigma_{\mathrm{b},P}\|_{\mathrm{op}}\big)\|\hat{\Sigma}_{\mathrm{w},P} - \Sigma_{\mathrm{w},P}\|_{\mathrm{op}}}{\big(1/R_2 - \|\hat{\Sigma}_{\mathrm{w},P} - \Sigma_{\mathrm{w},P}\|_{\mathrm{op}}\big)} \\
&\lesssim_{R_1,R_2} \|\hat{\Sigma}_{\mathrm{w},P} - \Sigma_{\mathrm{w},P}\|_{\mathrm{op}} \lesssim_{R_2} \frac{K}{n} + \sqrt{\frac{d+\log(1/\delta)}{n}}, \qquad (25)
\end{aligned}$$

其中我们在倒数第二个不等式中使用了（24）。对于（23）中的第二项，在 Ω 上我们同样有$_P$：

$$\big\|\Sigma_{\mathrm{w},P}^{-1}\hat{\Sigma}_{\mathrm{b},P} - \Sigma_{\mathrm{w},P}^{-1}\Sigma_{\mathrm{b},P}\big\|_{\mathrm{op}} \le \|\Sigma_{\mathrm{w},P}^{-1}\|_{\mathrm{op}}\|\hat{\Sigma}_{\mathrm{b},P} - \Sigma_{\mathrm{b},P}\|_{\mathrm{op}} \lesssim_{R_1,R_2} \frac{K}{n} + \sqrt{\frac{d+\log(1/\delta)}{n}}. \quad (26)$$

所需结果由结合（25）和（26），并利用以下事实得到：$\log(1/\delta) \le d\log(ep/d) + \log(1/\varepsilon)$。
□

5.4 命题的证明 5和定理6

In the proof of Proposition 5, we show the convergence of the EM iterates $\hat{\mu}^{(t)}$ by analyzing their components parallel and orthogonal to μ^* separately. Writing $\eta := \mu^*/\|\mu^*\|$, let $\alpha_t \in \mathbb{R}$, $\beta_t \ge 0$ be defined by

$$\hat{\mu}^{(t)} = \alpha_t\eta + \beta_t\xi_t, \qquad (27)$$

其中 $\xi_t \in \mathbb{S}^{d-1}$ 与 η 正交。我们的证明将结合若干命题，在不同条件下控制 α_t 和 β_t。我们首先进行一些铺垫，并定义本小节中将用到的一些量。

First, it will be convenient to relabel the two classes as $\{-1, 1\}$ instead of $\{1, 2\}$. By the rotational symmetry of the problem, we may assume without loss of generality that $\mu^* = (s, 0, \ldots, 0)^\top \in \mathbb{R}^d$ for some $s \ge 0$, and that the first n_{L} observations are labeled (i.e., $Y_i \ne 0$ for $i \in [n_{\mathrm{L}}]$). We assume throughout this section that $s \le r$ and $r \ge 1$. Let $\hat{\mu}_{n_{\mathrm{L}}} := n_{\mathrm{L}}^{-1}\sum_{i=1}^{n_{\mathrm{L}}} Z_iY_i$, with the convention that $\hat{\mu}_{n_{\mathrm{L}}} := 0$ if $n_{\mathrm{L}} = 0$, and define the function $f_{n_{\mathrm{U}}} : \mathbb{R}^d \to \mathbb{R}^d$ by

$$f_{n_{\mathrm{U}}}(v) := \frac{1}{n_{\mathrm{U}}}\sum_{i=n_{\mathrm{L}}+1}^{n} Z_i \tanh\langle Z_i, v\rangle, \qquad (28)$$

与 $f_{n_{\mathrm{U}}} := 0$ 若 $n_{\mathrm{U}} = 0$。在全文中，除非另行说明，我们假设$n = n_{\mathrm{L}} + n_{\mathrm{U}} \ge 2$。采用此记号，EM 更新（12）可改写为，定义函数 $g_n : \mathbb{R}^d \to \mathbb{R}^d$:

$$\hat{\mu}^{(t)} = g_n(\hat{\mu}^{(t-1)}) := \gamma\hat{\mu}_{n_{\mathrm{L}}} + (1-\gamma)f_{n_{\mathrm{U}}}(\hat{\mu}^{(t-1)}).$$

相应的总体量为

$$f(v) := \mathbb{E}Z_1 \tanh\langle v, Z_1\rangle \quad \text{and} \quad g(v) := \gamma\mu^* + (1-\gamma)f(v).$$

令 $\Delta_{n_{\mathrm{U}}} := f_{n_{\mathrm{U}}} - f$，我们有

$$g_n(v) = g(v) + (1-\gamma)\Delta_{n_{\mathrm{U}}}(v) + \gamma(\hat{\mu}_{n_{\mathrm{L}}} - \mu^*). \tag{29}$$

对于 ω， $\phi > 0$ 和 $r \geq 1$，我们定义以下两个事件，分别用于控制 EM迭代中涉及无标签数据和有标签数据的项:

$$\begin{aligned}
\Omega_1(\omega) &:= \left\{ \sup_{v\in\mathbb{R}^d} \|g_n(v)\| \leq 2(r+\sqrt{d}) \right\} \cap \left\{ \sup_{\substack{\|v\|\leq 2(r+\sqrt{d}) \\ v\neq 0}} \frac{\|\Delta_{n_{\mathrm{U}}}(v)\|}{\|v\|} \leq \omega \right\} \\
\Omega_2(\phi) &:= \left\{ \|\hat{\mu}_{n_{\mathrm{L}}} - \mu^*\| \leq \phi \right\}.
\end{aligned} \tag{30}$$

Proposition 9. *There exists $C_r > 0$, depending only on r, such that for any $\delta \in (2e^{-n}, 1]$ and $\omega = C_r\sqrt{\frac{d\log n + \log(1/\delta)}{n_{\mathrm{U}}}}$, we have $\mathbb{P}\big(\Omega_1(\omega)^{\mathrm{c}}\big) \leq \delta$. Moreover, for any $\delta \in (0,1]$ and for $\phi = \sqrt{\frac{2d+3\log(1/\delta)}{n_{\mathrm{L}}}}$, we have $\mathbb{P}\big(\Omega_2(\phi)^{\mathrm{c}}\big) \leq \delta$.*

证明。 对于任意 $v \in \mathbb{R}^d$,

$$\begin{aligned}
\|g_n(v)\| = \|(1-\gamma)f_{n_{\mathrm{U}}}(v) + \gamma\hat{\mu}_{n_{\mathrm{L}}}\| &\leq (1-\gamma)\cdot\frac{1}{n_{\mathrm{U}}}\sum_{i=n_{\mathrm{L}}+1}^{n}\|Z_i\| + \frac{\gamma}{n_{\mathrm{L}}}\sum_{i=1}^{n_{\mathrm{L}}}\|Z_i\| \\
&= \frac{1}{n}\sum_{i=1}^{n}\|Z_i\| \leq \left(\frac{1}{n}\sum_{i=1}^{n}\|Z_i\|^2\right)^{1/2}.
\end{aligned}$$

由于 $\sum_{i=1}^n \|Z_i\|^2 \sim \chi^2_{nd}(ns^2)$ ，根据 Birg´e (2001，引理 8.1），我们以至少 $1-\delta/2$ 的概率

$$\begin{aligned}
\sup_{v\in\mathbb{R}^d}\|g_n(v)\|^2 &\leq d + s^2 + 2\sqrt{\frac{(d+2s^2)\log(2/\delta)}{n}} + \frac{2\log(2/\delta)}{n} \\
&\leq 2d + 3s^2 + \frac{3\log(2/\delta)}{n} \leq 4(r+\sqrt{d})^2.
\end{aligned} \tag{31}$$

此外，采用与证明Wu and Zhou (2022年，定理4）中非常相似的论证，我们以至少 $1-\delta/2$ 的概率有

$$\sup_{\substack{\|v\|\leq 2(r+\sqrt{d}) \\ v\neq 0}} \frac{\|\Delta_{n_{\mathrm{U}}}(v)\|}{\|v\|} \leq C_r\sqrt{\frac{d\log n + \log(1/\delta)}{n_{\mathrm{U}}}}, \tag{32}$$

for some $C_r > 0$ depending only on r. The first claim follows by combining (31) and (32).

For the second claim, we have $\hat{\mu}_{n_{\mathrm{L}}} \sim N_d(\mu^*, n_{\mathrm{L}}^{-1}I_d)$. Hence, by Laurent and Massart (2000, Lemma 1), we have

$$\begin{aligned}
\mathbb{P}\big(\Omega_2(\phi)^{\mathrm{c}}\big) &= \mathbb{P}\big(n_{\mathrm{L}}\|\hat{\mu}_{n_{\mathrm{L}}} - \mu^*\|^2 > n_{\mathrm{L}}\phi^2\big) \\
&\leq \mathbb{P}\big(n_{\mathrm{L}}\|\hat{\mu}_{n_{\mathrm{L}}} - \mu^*\|^2 > d + 2\sqrt{d\log(1/\delta)} + 2\log(1/\delta)\big) \leq \delta,
\end{aligned}$$

按要求。 □

For any $a \in \mathbb{R}$, $b \in [0, \infty)$ and $\xi \in \mathbb{S}^{d-1}$ that is orthogonal to η, we define $F(a, b) := \eta^\top f(a\eta + b\xi)$ and $G(a, b) := \|(I_d - \eta\eta^\top) f(a\eta + b\xi)\|$. Note that the distribution of Z_1 is orthogonally invariant along the axis μ^*; in other words, if $P \in \mathbb{R}^{d\times d}$ is orthogonal and has μ^* as an eigenvector with eigenvalue 1, then $PZ_1 \overset{\mathrm{d}}{=} Z_1$. It follows that $f(a\eta + b\xi)$, and hence $F(a, b)$ and $G(a, b)$, do not depend on ξ. We remark that

$$f(\alpha_t\eta + \beta_t\xi_t) = F(\alpha_t, \beta_t)\eta + G(\alpha_t, \beta_t)\xi'_{t+1}$$

对于某个与 η 正交的 $\xi'_{t+1} \in \mathbb{S}^{d-1}$ 。

命题 10 控制 β_t EM算法迭代中与信号方向 η正交的分量的大小 $\zeta := \omega\gamma^{-1/2} \wedge \omega^{1/2}$。

命题 10。 假设 $\phi\gamma^{1/2} \le \omega \le$最小值$\{1/12, 1/(r+3)\}$且 $\|\hat{}\mu^{(0)}\| \le r+3$。在事件 $\Omega_1(\omega) \cap \Omega_2(\phi)$上，我们有

$$\limsup_{t\to\infty} \beta_t \le 60(\zeta \vee r\omega).$$

此外，在同一事件上，若对于某个 $t_0 \in \mathbb{N}_0$ 有 $\beta_{t_0} \le 60(\zeta \vee r\omega)$ ，则对于所有 $t \ge t_0$, 均有 $\beta_t \le 60(\zeta \vee r\omega)$。

Proof. We first claim that on the event $\Omega_1(\omega) \cap \Omega_2(\phi)$, we have $\|\hat{\mu}^{(t)}\| \le r + 3$ for all $t \in \mathbb{N}_0$. The case $t = 0$ is true by the assumption on the initializer $\hat{\mu}^{(0)}$, and if the claim holds for $t \in \mathbb{N}_0$, then since $2(r + \sqrt{d}) \ge 2(r + 1) \ge r + 3$, we have on $\Omega_1(\omega) \cap \Omega_2(\phi)$ that

$$\begin{aligned}\|\hat{\mu}^{(t+1)}\| &\le (1-\gamma)\big\{|F(\alpha_t, \beta_t)| + G(\alpha_t, \beta_t) + \|\Delta_{n_{\mathrm{U}}}(\hat{\mu}^{(t)})\|\big\} + \gamma(s + \phi)\\ &\le s + 2\sqrt{2/\pi} + \omega\|\hat{\mu}^{(t)}\| + \gamma\phi \le r + 2 + \frac{\|\hat{\mu}^{(t)}\|}{r+3} \le r + 3,\end{aligned}$$

其中第二个不等式用到了Wuand Zhou (2022年，引理5(5))。此外，由于 (29)，在 Ω 上有$_1$ $(\omega) \cap \Omega_2(\phi)$，对于 $t \in \mathbb{N}$,

$$\begin{aligned}\beta_{t+1} &= \big\|(I_d - \eta\eta^\top)\big\{(1-\gamma)\big(f(\hat{\mu}^{(t)}) + \Delta_{n_{\mathrm{U}}}(\hat{\mu}^{(t)})\big) + \gamma\hat{\mu}_{n_{\mathrm{L}}}\big\}\big\|\\ &\le (1-\gamma)\big\{G(\alpha_t, \beta_t) + \omega(|\alpha_t| + \beta_t)\big\} + \gamma\phi\\ &\le \beta_t(1-\gamma)\left\{1 + \omega - \frac{(\alpha_t^2 + \beta_t^2) \wedge 1}{6}\right\} + \gamma\phi + \omega|\alpha_t|, \qquad (33)\end{aligned}$$

其中，最后一个界使用了 Wu and Zhou (2022年，引理5(8))。若 $\alpha_t^2 + \beta_t^2 > 1$ 或 $\gamma > 1/2$, 则利用 $\omega \le 1/12$这一事实，由式（33）可得

$$\beta_{t+1} \le \frac{11}{12}\beta_t + \gamma\phi + (r+3)\omega \le \frac{11}{12}\beta_t + (r+4)\omega. \qquad (34)$$

另一方面，如果 $\alpha_t^2 + \beta_t^2 \le 1$ 和 $\gamma \le 1/2$，那么

$$\beta_{t+1} \le \beta_t\left(1 + \omega - \gamma - \frac{\alpha_t^2 + \beta_t^2}{12}\right) + \gamma\phi + \omega|\alpha_t|. \qquad (35)$$

注意，（34）的右侧关于 β_t 递增，而（35）的右侧关于 β_t 在 $\alpha_t^2 + \beta_t^2 \le 1$ 且 $\gamma \le 1/2$时递增。结合（34）和（35），记 $\beta_\infty :=$

上极限$_{t\to\infty}\beta_t$并利用$0 \le \frac{3}{\beta_\infty}(\omega-|\alpha_t|\beta_\infty/6)^2 = 3\omega^2/\beta_\infty - \omega|\alpha_t| + \alpha_t^2\beta_\infty/12$这一事实，我们有

$$\beta_\infty \le \max\left\{\frac{11}{12}\beta_\infty + (r+4)\omega,\ \beta_\infty\left(1+\omega-\gamma-\frac{\beta_\infty^2}{12}\right)+\gamma\phi+\frac{3\omega^2}{\beta_\infty}\right\}. \tag{36}$$

由（36）中最大值的第一项，我们得到

$$\beta_\infty \le (1r+38)\omega \le 60r\omega. \tag{37}$$

由（36）中的最大值的第二项，我们得到

$$\beta_\infty\left(\gamma-\omega+\frac{\beta_\infty^2}{12}\right) \le \gamma\phi+\frac{3\omega^2}{\beta_\infty}. \tag{38}$$

如果$\gamma < 2\omega$，那么由（38），

$$\beta_\infty \le 5\omega^{1/2} \le 5\sqrt{2}\zeta, \tag{39}$$

since otherwise we would have that the left-hand side would be at least $(-5+125/12)\omega^{3/2}$ and the right-hand side would at most $(\sqrt{2}+3/5)\omega^{3/2}$, contradicting the inequality. On the other hand, if $\gamma \ge 2\omega$, then we derive from (38) that

$$\beta_\infty^2 - 2\phi\beta_\infty - \frac{6\omega^2}{\gamma} \le 0.$$

解这个不等式，我们得到

$$\beta_\infty \le \phi + \sqrt{\phi^2+6\omega^2/\gamma} \le 4\omega\gamma^{-1/2} = 4\zeta. \tag{40}$$

The first claim of the proposition follows by combining (36), (37), (39) and (40). We now prove the second claim by induction on t. The base case $t = t_0$ is true by assumption, so we assume that $\beta_t \le 60(\zeta \vee r\omega)$ for some $t \ge t_0$. Again we consider two cases. If $\beta_t \le 36(\zeta \vee r\omega)$, then from (33) and using that $|\alpha_t| \le \|\hat{\mu}^{(t)}\| \le r+3$ for $t \ge 2$,

$$\beta_{t+1} \le \beta_t(1+\omega) + \gamma\phi + \omega|\alpha_t| \le 44(\zeta \vee r\omega),$$

如所需。另一方面，如果 $\beta_t > 36$（$\zeta \vee r\omega$），那么结合（34）和（35），我们得到

$$\beta_{t+1} \le \max\left\{60(\zeta \vee r\omega),\ \beta_t\left(1+\omega-\gamma-\frac{\beta_t^2}{12}\right)+\gamma\phi+\frac{3\omega^2}{\beta_t}\right\}.$$

只需证明最大值中的第二项不大于 β_t。为此，如果$\gamma \le 2\omega$，那么 $\zeta^2 \le \omega \le 2\zeta^2$，因此

$$\begin{aligned}\beta_t\left(1+\omega-\gamma-\frac{\beta_t^2}{12}\right)+\gamma\phi+\frac{3\omega^2}{\beta_t} &\le \beta_t + 2\zeta^2\beta_t - \frac{\beta_t^3}{12} + \gamma^{1/2}\omega + \frac{3\omega^2}{\beta_t} \\ &\le \beta_t + 120(\zeta \vee r\omega)\zeta^2 - \frac{36^3(\zeta \vee r\omega)^3}{12} + 4\zeta^3 + \frac{\zeta^3}{3} \le \beta_t.\end{aligned}$$

另一方面，如果 $\gamma > 2\omega$，那么 $\zeta = \omega\gamma^{-1/2} \ge \phi$，因此

$$\begin{aligned}\beta_t\left(1+\omega-\gamma-\frac{\beta_t^2}{12}\right)+\gamma\phi+\frac{3\omega^2}{\beta_t} &\le \beta_t - \frac{\gamma\beta_t}{2} + \gamma\phi + \frac{3\omega^2}{\beta_t} \\ &\le \beta_t - 18\gamma\zeta + \gamma\zeta + \frac{\omega^2}{12\zeta} \le \beta_t,\end{aligned}$$

如所愿，这就完成了归纳法。 □

以下结果对 EM迭代的信号分量 α_t 的幅度进行了界定。

Proposition 11. *Assume that* $\phi\gamma^{1/2} \le \omega \le \min\{1/12, 1/(r+3)\}$ *and that* $\|\hat{\mu}^{(0)}\| \le r+3$. *Then there exists* $C_r > 0$, *depending only on* r, *such that on the event* $\Omega_1(\omega) \cap \Omega_2(\phi)$, *we have*

$$\limsup_{t\to\infty} |\alpha_t| \le C_r(\zeta \vee s).$$

证明。 根据 α_{t+1} 和（29），对于每个 $t \in \mathbb{N}_0$ 都有

$$\alpha_{t+1} = \eta^\top\big\{(1-\gamma)\big(f(\hat{\mu}^{(t)}) + \Delta_{n_{\mathrm{U}}}(\hat{\mu}^{(t)})\big) + \gamma\hat{\mu}_{n_{\mathrm{L}}}\big\}.$$

因此，根据命题证明中的第一个论断10，在事件 $\Omega_1(\omega)\cap\Omega_2(\phi)$ 上，有

$$\big|\alpha_{t+1} - (1-\gamma)F(\alpha_t,\beta_t) - \gamma s\big| \le (1-\gamma)\omega(|\alpha_t| + \beta_t) + \gamma\phi \tag{41}$$

for every $t \in \mathbb{N}_0$. From Wu and Zhou (2022, Lemma 5(1) and Lemma 5(7)), $\alpha \mapsto F(\alpha,\beta)$ is an increasing and odd function satisfying $|F(\alpha,\beta) - F(\alpha,0)| \le (1+s^2)|\alpha|\beta^2$ for every $\alpha, \beta \in \mathbb{R}$. Hence, by (41), we have on $\Omega_1(\omega) \cap \Omega_2(\phi)$ that

$$|\alpha_{t+1}| \le (1-\gamma)\Big\{F(|\alpha_t|,0) + (1+s^2)|\alpha_t|\beta_t^2 + \omega(|\alpha_t| + \beta_t)\Big\} + \gamma(s+\phi). \tag{42}$$

Note the right-hand side of (42) is increasing in $|\alpha_t|$. Define $\alpha_\infty := \limsup_{t\to\infty}|\alpha_t|$, so that $\alpha_\infty \le r+3$ on $\Omega_1(\omega)\cap\Omega_2(\phi)$, again by the first claim in the proof of Proposition 10. We may also assume that $\alpha_\infty > s$, because otherwise the result is clear. Since $\alpha \mapsto F(\alpha,0)$ is continuous, we have from (42) that on $\Omega_1(\omega)\cap\Omega_2(\phi)$,

$$\alpha_\infty \le (1-\gamma)\big\{F(\alpha_\infty,0) + (1+r^2)\alpha_\infty\beta_\infty^2 + \omega(\alpha_\infty + \beta_\infty)\big\} + \gamma(s+\phi), \tag{43}$$

其中我们回顾 β_∞:= 上极限$_{t\to\infty}\beta_t$。 定义 $q\colon [0,\infty) \to \mathbb{R}$ 通过

$$q(\alpha) := \begin{cases} F(\alpha,0)/\alpha & \text{if } \alpha \ne 0 \\ 1+s^2 & \text{if } \alpha = 0. \end{cases} \tag{44}$$

By Lemma 20, we have $q(s) = 1$ (which confirms that μ_* is a fixed point of the population EM iteration), and that $q'(\alpha) \le -c_r\alpha$ for all $\alpha \in (0, r]$, where $c_r \in (0,1]$ depends only on r. Thus, dividing both sides of (43) by α_∞, we have

$$\begin{aligned} 1 &\le (1-\gamma)\bigg\{q(s) + \int_s^{\alpha_\infty} q'(\alpha)\,d\alpha + (1+r^2)\beta_\infty^2 + \omega\bigg(1 + \frac{\beta_\infty}{\alpha_\infty}\bigg)\bigg\} + \frac{\gamma(s+\phi)}{\alpha_\infty} \\ &\le (1-\gamma)\bigg\{1 - \frac{c_r}{2}(\alpha_\infty^2 - s^2) + (1+r^2)\beta_\infty^2 + \omega\bigg(1 + \frac{\beta_\infty}{\alpha_\infty}\bigg)\bigg\} + \frac{\gamma(s+\phi)}{\alpha_\infty}. \end{aligned} \tag{45}$$

Now $\beta_\infty \le 60(1+r)\zeta$ by Proposition 10. We now claim that $\alpha_\infty \le 4s + 120c_r^{-1/2}(1+r)^2\zeta$. Indeed, assuming the contrary, we would have $c_r\alpha_\infty^2/4 > c_r s^2/2 + (1+r^2)\beta_\infty^2$ and $\beta_\infty/\alpha_\infty < 1$. Hence from (45), we have

$$1 \le (1-\gamma)\bigg(1 - \frac{c_r\alpha_\infty^2}{4}\bigg) + \frac{\gamma s}{\alpha_\infty} + \bigg(2 + \frac{\gamma^{1/2}}{\alpha_\infty}\bigg)\omega.$$

我们考虑两种情况。首先，如果 $\gamma \le 4\omega$，那么 $2\zeta \ge \omega^{1/2} \ge \gamma^{1/2}/2$，因此

$$\begin{aligned}&(1-\gamma)\left(1-\frac{c_r\alpha_\infty^2}{4}\right)+\frac{\gamma s}{\alpha_\infty}+\left(2+\frac{\gamma^{1/2}}{\alpha_\infty}\right)\omega\\&\le \max\left\{1-\frac{c_r\alpha_\infty^2}{4},0\right\}+\frac{\gamma}{4}+3\omega \le \max\left\{1-\zeta^2,\frac{\gamma}{4}+3\omega\right\}<1,\end{aligned}$$

一个矛盾。其次，如果 $\gamma > 4\omega$，那么 $\zeta = \omega\gamma^{-1/2}$ 和

$$(1-\gamma)\left(1-\frac{c_r\alpha_\infty^2}{4}\right)+\frac{\gamma s}{\alpha_\infty}+\left(2+\frac{\gamma^{1/2}}{\alpha_\infty}\right)\omega \le 1-\gamma+\frac{\gamma}{4}+2\omega+\frac{\gamma\zeta}{\alpha_\infty}<1,$$

再次得到矛盾。这证明了所声称的 α_∞ 上界。 □

Recall the definitions $L(\mu,\mu^*) = \|\mu-\mu^*\| \wedge \|\mu+\mu^*\|$. Our next result shows that if α_t ever becomes sufficiently large, then improved bounds can be derived on the limiting behaviour of α_t, β_t and $L(\hat\mu^{(t)},\mu^*)$.

Proposition 12. *Assume that $\gamma \in [0,1/2)$. Given any $c>0$, there exists $C, c_1 > 0$, depending only on r and c, such that if $|\alpha_{t_0}| \ge cs$, $\beta_{t_0} \le 60(\zeta \vee r\omega)$ for some iteration t_0, and $\phi\gamma^{1/2} \le \omega \le c_1$ and $s \ge C\zeta$, then on the event $\Omega_1(\omega)\cap\Omega_2(\phi)$, we have*

$$\limsup_{t\to\infty}|\alpha_t - s| \lesssim_{r,c} \frac{\omega}{s}\wedge\frac{\omega}{\gamma^{1/2}}, \tag{46}$$

$$\limsup_{t\to\infty}\beta_t \lesssim_{r,c} \frac{\omega}{s}\wedge\frac{\omega}{\gamma^{1/2}}, \tag{47}$$

$$\limsup_{t\to\infty}L(\hat\mu^{(t)},\mu^*) \lesssim_{r,c} \frac{\omega}{s}\wedge\frac{\omega}{\gamma^{1/2}}. \tag{48}$$

证明。如有必要，通过翻转 μ^* 的符号，我们可以不失一般性地假设 $\alpha_{t_0} \ge 0$ 并且 $c_1 \le$ 最小值$\{1/12, 1/(r+3)\}$。由（41）及其下方紧接着的论证，我们有

$$\begin{aligned}(1-\gamma)\Big\{F(\alpha_t,0)-(1+s^2)\alpha_t\beta_t^2-\omega(\alpha_t+\beta_t)\Big\}+\gamma(s-\phi) &\le \alpha_{t+1}\\ &\le (1-\gamma)\Big\{F(\alpha_t,0)+(1+s^2)\alpha_t\beta_t^2+\omega(\alpha_t+\beta_t)\Big\}+\gamma(s+\phi).\end{aligned} \tag{49}$$

对于任意 t 使得 $\alpha_t \ge cs$，由于 $\beta_t \le 60r\omega^{1/2}$ 根据命题 10，我们有 t

$$(1+s^2)\alpha_t\beta_t^2+\omega(\alpha_t+\beta_t) \le \left\{(1+r^2)60^2r^2\omega+\omega\left(1+\frac{60s\zeta}{cs}\right)\right\}\alpha_t \le c''\omega\alpha \tag{50}$$

其中 $c'' := 60^2(1+r^2)r^2 + (1+60rc^{-1}C^{-1})$。此外，如果 $\alpha_t \ge cs$，那么

$$\gamma\phi \le \omega\gamma^{1/2} \le \begin{cases}\gamma\zeta \le C^{-1}\gamma s & \text{if } \gamma \ge \omega\\ \omega^{3/2} = \omega\zeta \le c^{-1}C^{-1}\omega\alpha_t & \text{otherwise.}\end{cases} \tag{51}$$

令 $c' := c'' + 2c^{-1}C^{-1}$，并定义函数 $H, L\colon [0,\infty) \to \mathbb{R}$，其中

$$\begin{aligned}H(\alpha) &:= (1-\gamma)\big\{F(\alpha,0)+c'\omega\alpha\big\}+\gamma(s+\phi),\\ L(\alpha) &:= (1-\gamma)\big\{F(\alpha,0)-c'\omega\alpha\big\}+\gamma\max(s-\phi, s/2),\end{aligned}$$

由（49）、（50）和（51），我们得到对于 $\alpha_t \geq cs$ 和 $C \geq 2$,

$$L(\alpha_t) \leq \alpha_{t+1} \leq H(\alpha_t). \tag{52}$$

Defi由 $\alpha_{t_0}^+ := \alpha_{t_0} =: \alpha_{t_0}^-$ 定义的辅助序列 $(\alpha_t^+)_{t\geq t_0}$ 和 $(\alpha_t^-)_{t\geq 0}$，以及对于 $t \geq t_0$,

$$\alpha_{t+1}^+ := H(\alpha_t^+) \quad \text{and} \quad \alpha_{t+1}^- := L(\alpha_t^-).$$

We first derive some properties of the two recursion maps H and L. For the former, we have by Wu and Zhou (2022, Lemma 3) that F, and hence H, is increasing and concave on $[0,\infty)$ with $H(0) > 0$ when $\gamma > 0$ and $H'(0) > \partial_1 F(0,0) > 1$ when $\gamma = 0$. Moreover, since F is bounded, we can choose $c_1 > 0$, depending only on r and c, such that $\lim_{\alpha\to\infty} H'(\alpha) = (1-\gamma)c'\omega \leq (1-\gamma)c'c_1 < 1/2$. On the other hand, we have $L(0) > 0$ when $\gamma > 0$. When $\gamma = 0$, we have $\omega^{1/2} = \zeta \leq s/C$, which means that after increasing $C \equiv C(r,c) > 0$ if necessary, $L'(0) = \partial_1 F(0,0) - c'\omega \geq 1 + s^2 - c's^2/C^2 > 1$. By Wu and Zhou (2022, Lemma 3), $\alpha \mapsto F(\alpha, 0)$ is differentiable, increasing and concave for $\alpha \in [0,\infty)$. Reducing $c_1 \equiv c_1(r,c) > 0$ if necessary to ensure that $c_1 \leq \partial_1 F(r+3,0)/c'$, we have for $\alpha \in [0, r+3]$ that

$$L'(\alpha) = \partial_1 F(\alpha, 0) - c'\omega \geq \partial_1 F(r+3, 0) - c'c_1 \geq 0.$$

In other words, L is increasing on $[0, r+3]$, and moreover, similarly to H, it is also concave on this interval. Finally, we claim that for $\tilde{c} := \min\{c, 32(3+r^4)/3\}^{-1/2}\}$, and $\tilde{\alpha} := \tilde{c}s$, we have $L(\tilde{\alpha}) \geq \tilde{\alpha}$. To verify this, we note by Wu and Zhou (2022, Lemma 3(3) and Equation (98)), we have

$$L(\tilde{\alpha}) \geq (1-\gamma)\tilde{\alpha}\left\{1 + s^2 - \frac{8}{3}(3+r^4)\tilde{\alpha}^2 - c'\omega\right\} + 2\gamma\tilde{\alpha} \geq (1-\gamma)\tilde{\alpha}\left(1 + \frac{3s^2}{4} - c'\omega\right) + 2\gamma\tilde{\alpha}. \tag{53}$$

To control the right-hand side of (53), if $\gamma \leq c'\omega$, we have $\zeta = \omega^{1/2} \wedge \omega\gamma^{-1/2} \geq (\omega/c')^{1/2}$. Hence, from the condition $s \geq C\zeta$, if we choose $C > 2c'$ (which is possible because c' is a decreasing function of C), we have $c'\omega \leq (c's/C)^2 \leq s^2/4$ and consequently the right-hand side of (53) is at least $\tilde{\alpha}$. If $\gamma > c'\omega$, then we have $L(\tilde{\alpha}) \geq (1-\gamma)^2\tilde{\alpha} + 2\gamma\tilde{\alpha} \geq \tilde{\alpha}$ as desired. This establishes the claim.

现在我们通过归纳法证明对所有 $t \geq t_0$,

$$\tilde{\alpha} \leq \alpha_t^- \leq \alpha_t \leq \alpha_t^+. \tag{54}$$

基例根据 $\alpha_{t_0}^-$ 和 $\alpha_{t_0}^+$ 的上述定义显然成立。现在假设（54）对某次迭代 $t \geq 0$成立，因此特别地，（52）适用。

Using the monotonicity of H and (52), we have $\alpha_{t+1} \leq H(\alpha_t) \leq H(\alpha_t^+) = \alpha_{t+1}^+$. Observe that $\alpha_t \leq \|\hat{\mu}^{(t)}\| \leq r+3$ by the proof of Proposition 10. Using the monotonicity of L on $[0, r+3]$, we find that $\alpha_{t+1} \geq L(\alpha_t) \geq L(\alpha_t^-) = \alpha_{t+1}^-$. Moreover,

$$\alpha_{t+1}^- = L(\alpha_t^-) \geq L(\tilde{\alpha}) \geq \tilde{\alpha},$$

这就完成了归纳法。

To prove (46), we will analyze the sequences $(\alpha_t^+)_{t\geq t_0}$ and $(\alpha_t^-)_{t\geq t_0}$, which sandwich $(\alpha_t)_{t\geq t_0}$. We start by considering the behaviour of $(\alpha_t^+)_{t\geq t_0}$. The properties of H derived above mean that we can apply Lemma 19 to obtain that α_t^+ converges to a limit, denoted α^+, satisfying $\alpha^+ = H(\alpha^+)$. By Lemma 20, we have $F(s,0) = s$ and so $H(s) = (1-$

$\gamma)(s + c'\omega s) + \gamma(s + \phi) > s$. Hence from Lemma 19 again, we have $\alpha^+ > s$. On the other hand, since $F(\alpha, 0) \le \mathbb{E}|Z_{1,1}| \le (\mathbb{E}Z_{1,1}^2)^{1/2} \le (1 + s^2)^{1/2} \le 1 + r$, we have $\alpha^+ = H(\alpha^+) \le (1-\gamma)(1+r) + \alpha^+/2 + \gamma r + 1 \le r + 2 + \alpha^+/2$, so $\alpha^+ \le 2r + 4$. Recalling the definition of q from (44), by Lemma 20 again, we have

$$q(\alpha^+) = q(s) + \int_s^{\alpha^+} q'(\alpha)\, d\alpha \le 1 - c_2\big((\alpha^+)^2 - s^2\big)$$

对于某个仅取决于 r 的 $c_2 > 0$。因此，

$$\begin{aligned}\alpha^+ = H(\alpha^+) &= (1-\gamma)\alpha^+\big\{q(\alpha^+) + c'\omega\big\} + \gamma(s+\phi)\\ &\le (1-\gamma)\alpha^+\big\{1 - c_2\big((\alpha^+)^2 - s^2\big) + c'\omega\big\} + \gamma(s + \omega\gamma^{-1/2}),\end{aligned}$$

so

$$(\alpha^+)^2 - s^2 \le \frac{c'\omega}{c_2} - \frac{\gamma}{(1-\gamma)c_2}\frac{\alpha^+ - s - \omega\gamma^{-1/2}}{\alpha^+}. \tag{55}$$

我们现在来证明，

$$\alpha^+ - s \lesssim_{r,c} \frac{\omega}{s} \wedge \frac{\omega}{\gamma^{1/2}} \tag{56}$$

通过考虑两种情况。如果 $\alpha^+ \le 2s$，则根据（55），我们有

$$\alpha^+ - s \le \frac{(1-\gamma)c'\omega + \gamma^{1/2}\omega/\alpha^+}{(1-\gamma)c_2(\alpha^+ + s) + \gamma/\alpha^+} \lesssim_{r,c} \frac{\omega}{s}\cdot\frac{1 + \gamma^{1/2}/s}{1 + \gamma/s^2} \lesssim \frac{\omega}{s} \wedge \frac{\omega}{\gamma^{1/2}}.$$

另一方面，如果 $\alpha^+ > 2s$，那么我们再次由（55）得到

$$\frac{3(\alpha^+)^2}{4} + \frac{\gamma}{2c_2} \le (\alpha^+)^2 - s^2 + \frac{\gamma(\alpha^+ - s)}{(1-\gamma)c_2\alpha^+} \le \frac{(c' + 2\gamma^{1/2}/\alpha^+)\omega}{c_2}. \tag{57}$$

In particular, $\gamma/2 \le (c' + \gamma^{1/2}/s)\omega \le c'\omega + C^{-1}(\gamma^{1/2}\omega^{1/2} \vee \gamma)$, so $\gamma \lesssim_{r,c} \omega$ and $\alpha^+ \gtrsim_{r,c} \zeta \gtrsim_{r,c} \gamma^{1/2}$. Consequently from (57),

$$\alpha^+ - s \le \alpha^+ \lesssim_{r,c} \omega^{1/2} \lesssim_{r,c} \frac{\omega}{\alpha^+} \wedge \frac{\omega}{\gamma^{1/2}} \lesssim \frac{\omega}{s} \wedge \frac{\omega}{\gamma^{1/2}},$$

which establishes (56). We now consider $(\alpha_t^-)_{t \ge t_0}$. Define $\tilde{L} : [0,\infty) \to [0,\infty)$ by $\tilde{L}(\alpha) := L\big(\alpha \wedge (r+3)\big)$. Since $\alpha_t^- \le \alpha_t \le r + 3$ for all $t \ge 0$, we have $\alpha_{t+1}^- = \tilde{L}(\alpha_t^-)$ for all $t \ge t_0$. From the properties of L derived above, we see that $\tilde{L}$ satisfies the conditions of Lemma 19, and hence α_t^- converges to a limit, denoted α^-, satisfying $\alpha^- = \tilde{L}(\alpha^-) = L(\alpha^-)$. By Lemma 20, $F(s,0) = s$, so we have $\tilde{L}(s) = L(s) \le (1-\gamma)(s - c'\omega s) + \gamma s < s$, so by Lemma 19, we must have $\alpha^- < s$. By Lemma 20 again, we have

$$q(\alpha^-) = q(s) - \int_{\alpha^-}^{s} q'(\alpha)\, d\alpha \ge 1 + c_2'\big(s^2 - (\alpha^-)^2\big),$$

其中 $c_2' > 0$ 仅取决于 r。因此，我们有

$$\alpha^- = L(\alpha^-) \ge (1-\gamma)\alpha^-\big\{1 + c_2'\big(s^2 - (\alpha^-)^2\big) - c'\omega\big\} + \gamma(s - \omega\gamma^{-1/2}),$$

整理并利用 $\alpha^- \ge \tilde{\alpha} \gtrsim_r s$ 得到

$$s - \alpha^- \le \frac{(1-\gamma)c'\omega + \gamma^{1/2}\omega/\alpha^-}{(1-\gamma)c_2'(s + \alpha^-) + \gamma/\alpha^-} \lesssim_{r,c} \frac{\omega}{s}\cdot\frac{1 + \gamma^{1/2}/s}{1 + \gamma/s^2} \lesssim \frac{\omega}{s} \wedge \frac{\omega}{\gamma^{1/2}}. \tag{58}$$

结合（54）、（56）和（58），我们得到（46）。

现在我们转而证明（47）。如有必要，通过递增 C，我们有对于所有充分大的 t，$s/2 \le \alpha_t \le 2s$。因此，根据（33），对于所有足够大的 t,

$$\beta_{t+1} \le \beta_t(1-\gamma)\left(1+\omega-\frac{s^2/4 \wedge 1}{6}\right)+\gamma\phi+2s\omega \le \beta_t(1-\gamma)(1+\omega-c_3 s^2)+(\gamma^{1/2}+2s)\omega, \quad (59)$$

with $c_3 := 1/(6r^2+24)$. Denote $\beta_\infty := \limsup_{t\to\infty} \beta_t$. If $\gamma \le 2\omega$, then $\zeta \in [(\omega/2)^{1/2}, \omega^{1/2}]$ so $\omega \le 2s^2/C^2 \le c_3 s^2/2$ for C sufficiently large. Hence, from (59), and the fact that $\gamma^{1/2} \le 2\zeta \le s$,

$$\beta_\infty \le \frac{(\gamma^{1/2}+2s)\omega}{c_3 s^2/2} \lesssim_r \frac{\omega}{s} = \frac{\omega}{s} \wedge \frac{\omega}{\gamma^{1/2}}.$$

另一方面，如果 $\gamma > 2\omega$，那么 $(1-\gamma)$ $(1+\omega-c_3 s^2)$ $\le 1-\gamma/2-c_3 s^2/2$并且由（59），我们得到

$$\beta_\infty \le \frac{(\gamma^{1/2}+2s)\omega}{(\gamma+c_3 s^2)/2} \lesssim_r \frac{\omega}{s}\cdot\frac{1+\gamma^{1/2}/s}{1+\gamma/s^2} \lesssim \frac{\omega}{s} \wedge \frac{\omega}{\gamma^{1/2}}.$$

结合上述两种情况可得出（47）。

最后，回顾$\mu^{(t)}$ 在（27）中的分解，可知（48）可由（46）和（47）立即得到。

□

接下来，我们证明，只要初始化与真实参数的相关性不至于太低，$|\alpha_t|$ 就会达到足以使命题 12 在经过足够多次迭代后适用的水平。

Proposition 13. *Assume that $n \ge 3$, that $\phi\gamma^{1/2} \le \omega \le \min\{1/12, 1/(r+3)\}$ and that $\gamma \in [0, 1/2)$. Suppose that $\hat{\mu}^{(0)}$ is chosen such that $c'(\zeta \vee r\omega) \le \|\hat{\mu}^{(0)}\| \le 60(\zeta \vee r\omega)$ for some $c' \in (0,1)$ and that $|\langle \hat{\mu}^{(0)}/\|\hat{\mu}^{(0)}\|, \eta\rangle| \ge \sqrt{1/(d \log n_{\mathrm{U}})}$. Then there exist $c_4, c_5 > 0$, depending only on r and c', such that if $s \ge c_4\zeta\sqrt{d \log n_{\mathrm{U}}}$, then on $\Omega_1(\omega) \cap \Omega_2(\phi)$, we have*

$$|\alpha_t| \ge c_5 s$$

对于某个 $t > 0$。

Proof. By flipping the sign of μ^* if necessary, we may assume without loss of generality that $\alpha_0 \ge 0$. Assuming that the desired result is not true, we will prove by induction that on $\Omega_1(\omega) \cap \Omega_2(\phi)$, (a) $\alpha_t/\beta_t \ge c'/(60\sqrt{d \log n_{\mathrm{U}}})$ and (b) $\alpha_{t+1} \ge (1+\omega\sqrt{d \log n_{\mathrm{U}}})\alpha_t$ for all $t \ge 0$. We show this by first verifying the base case of (a), then proving that (a) implies (b) for each t, and finally proving that $\alpha_{t+1}/\beta_{t+1} \ge 1/(60\sqrt{d \log n_{\mathrm{U}}})$ once (b) holds for a given t.

对于基例，根据对 $\hat{}\mu^{(0)}$ 的假设，我们有

$$\frac{\alpha_0}{\beta_0} \ge \frac{\alpha_0}{\|\hat{\mu}^{(0)}\|} = \left\langle \frac{\hat{\mu}^{(0)}}{\|\hat{\mu}^{(0)}\|}, \eta \right\rangle \ge \frac{1}{\sqrt{d \log n_{\mathrm{U}}}} \ge \frac{c'}{60\sqrt{d \log n_{\mathrm{U}}}}$$

因为 $c' \le 60$。

Now assume that $\alpha_t/\beta_t \ge c'/(60\sqrt{d \log n_{\mathrm{U}}})$ and that $\alpha_0 \le \alpha_t < c_5 s$ for some $t \ge 0$. We aim to show that (b) holds for the same t, and start by controlling $e_1^\top f(\hat{\mu}^{(t)})$. Let $W = (W_1, \ldots, W_d)^\top$ be an independent copy of Z_1, independent of all other randomness in the problem, and define $u_t := \{\tanh(W_1\alpha_t + W_{-1}^\top \hat{\mu}_{-1}^{(t)}) - \tanh(W_1\alpha_t - W_{-1}^\top \hat{\mu}_{-1}^{(t)})\}/2$.

然后通过应用引理的第二部分18，并令 $a = W_1\alpha_t$ 和 $b = W_{-1}^\top \hat{\mu}_{-1}^{(t)}$（使得$a + b = W^\top \hat{\mu}^{(t)}$），我们得到

$$\begin{aligned} e_1^\top f(\hat{\mu}^{(t)}) &= \mathbb{E}\big\{W_1 \tanh(W_1\alpha_t + W_{-1}^\top \hat{\mu}_{-1}^{(t)})\big\} \\ &\geq \mathbb{E}\bigg\{\alpha_t W_1^2 - \frac{\alpha_t^3 W_1^4}{3} - \alpha_t W_1^2 (W_{-1}^\top \hat{\mu}_{-1}^{(t)})^2\bigg\} + \mathbb{E}(W_1 u_t) \\ &= \alpha_t(1+s^2)(1-\beta_t^2) - \alpha_t^3(1 + 2s^2 + s^4/3), \end{aligned}$$

where in the final step we have used the fact that u_t is an odd function of $W_{-1}^\top \hat{\mu}_{-1}^{(t)}$, which has a symmetric distribution about 0, conditional on $(W_1, \hat{\mu}_{-1}^{(t)})$, and hence $\mathbb{E}(W_1 u_t) = \mathbb{E}\big\{\mathbb{E}(W_1 u_t \mid W_1, \hat{\mu}_{-1}^{(t)})\big\} = 0$. From the assumption $s \geq c_4\zeta$, and using $\beta_t \leq 60(1+r)\zeta$ from Proposition 10, we have for sufficiently large c_4 that $\beta_t^2(1+s^2) \leq \{60(1+r)\}^2(1+r^2)s^2/c_4^2 \leq s^2/4$. By choosing $c_5 > 0$, depending only on r, sufficiently small, we may assume that $\alpha_t^2(1 + 2s^2 + s^4/3) < c_5^2(1 + 2r^2 + r^4/3)s^2 \leq s^2/4$. Recall the definition of f_{n_U} from (28). Since on the event $\Omega_1(\omega) \cap \Omega_2(\phi)$, we have $\|\hat{\mu}^{(t)}\| \leq r + 3 \leq 2(r + \sqrt{d})$ as in the first line of the proof of Proposition 10, we have on the event $\Omega_1(\omega) \cap \Omega_2(\phi)$ that

$$\begin{aligned} \alpha_{t+1} &= (1-\gamma) e_1^\top f_{n_U}(\hat{\mu}^{(t)}) + \gamma e_1^\top \hat{\mu}_{n_L} \\ &\geq (1-\gamma)\big\{e_1^\top f(\hat{\mu}^{(t)}) - \omega\|\hat{\mu}^{(t)}\|\big\} + \gamma(s - \phi) \\ &\geq (1-\gamma)\big\{\alpha_t(1 + s^2/2) - \omega(\alpha_t + \beta_t)\big\} + \gamma(s - \phi). \end{aligned}$$

如果 $\gamma \geq \omega$，则 $\phi \leq \omega\gamma^{-1/2} = \zeta \leq s/2$（假设 $c_4 \geq 2$）。如果 $\gamma < \omega$，则

$$\gamma\phi \leq \omega\gamma^{1/2} < \omega^{3/2} = \omega\zeta \leq \frac{\alpha_0\omega}{c'}\sqrt{d \log n_U} \leq \frac{2(1-\gamma)\alpha_t\omega}{c'}\sqrt{d \log n_U}.$$

因此，在这两种情况下，我们都有 $\Omega_1(\omega) \cap \Omega_2(\phi)$，使得

$$\begin{aligned} \alpha_{t+1} &\geq (1-\gamma)\alpha_t\big\{1 + s^2/2 - \omega(1 + 62/c')\sqrt{d \log n_U})\big\} + \frac{\gamma s}{2} \\ &\geq (1-\gamma)\alpha_t\big\{1 + s^2/2 + (1 + 62/c')(2\gamma - \omega\sqrt{d \log n_U})\big\}, \end{aligned}$$

where the final bound holds provided we reduce c_5 to be at most $1/(2+124/c')$ if necessary.

Now, when $\gamma \leq \omega\sqrt{d \log n_U}$, we have $\zeta = \omega^{1/2} \wedge \omega\gamma^{-1/2} \geq \omega^{1/2}(d \log n_U)^{-1/4}$ and hence by the condition on s in the proposition, we have $s^2 \geq c_4^2\zeta^2 d \log n_U \geq c_4^2\omega\sqrt{d \log n_U}$. Thus, by increasing c_4 to be at least $\sqrt{4 + 248/c'}$ if necessary, we have $(1 + 62/c')\omega\sqrt{d \log n_U} \leq s^2/4$. Hence, in this case, and on the event $\Omega_1(\omega) \cap \Omega_2(\phi)$,

$$\alpha_{t+1} \geq (1-\gamma)(1 + s^2/4 + (2 + 124/c')\gamma)\alpha_t \geq \bigg(1 + \frac{s^2}{8}\bigg)\alpha_t \geq (1 + \omega\sqrt{d \log n_U})\alpha_t.$$

另一方面，当 $\gamma > \omega\sqrt{d \log n_U}$ 时，我们有

$$\alpha_{t+1} \geq (1-\gamma)(1 + (1 + 62/c')\gamma)\alpha_t \geq (1+\gamma)\alpha_t \geq (1 + \omega\sqrt{d \log n_U})\alpha_t.$$

结合上述两个界，即可证明给定 t 的 (b)。

It remains to verify (a) for $t+1$, assuming that (a) and (b) hold up to and including t. Since $\beta_0 \leq \|\hat{\mu}^{(0)}\| \leq 60(\zeta \vee r\omega)$, we have by Proposition 10 that $\beta_{t+1} \leq 60(\zeta \vee r\omega) \leq 60(c')^{-1}\|\hat{\mu}^{(0)}\|$. Thus,

$$\frac{\alpha_{t+1}}{\beta_{t+1}} \geq \frac{\alpha_0}{60\|\hat{\mu}^{(0)}\|/c'} \geq \frac{c'}{60\sqrt{d \log n_U}},$$

由此完成归纳法。特别是，(b) 所蕴含的 α_t 的几何增长意味着，当 $t > 0$ 足够大时，α_t 将超过 $c_5 s$ 。这构成了我们所需的矛盾，从而证明了该结论。

□

Proof of Proposition 5. Define $\phi_0 := \omega_0 \gamma^{-1/2}$, and recall the definitions of $\Omega_1(\omega)$ and $\Omega_2(\phi)$ from (30). By Proposition 9, there exists $C_r \geq 1$, depending only on r, such that for $\omega = C_r \omega_0$ and $\phi = C_r \phi_0$, we have $\mathbb{P}\big(\Omega_1(\omega) \cap \Omega_2(\phi)\big) \geq 1 - 2\delta$.

(i) By the definition of ω and ϕ, we have $\omega = \phi\gamma^{1/2}$. If we choose c such that $c \leq C_r^{-1} \min\{1/12, 1/(r+3)\}$, then $\omega \leq \min\{1/12, 1/(r+3)\}$. Thus, we may apply Propositions 10 and 11 to obtain that on $\Omega_1(\omega) \cap \Omega_2(\phi)$, we have

$$\limsup_{t\to\infty} \|\hat{\mu}^{(t)} - \mu^*\| \leq \limsup_{t\to\infty} (|\alpha_t| + \beta_t + \|\mu^*\|) \lesssim_r \zeta \vee \|\mu^*\|.$$

第一个论断得证。

(ii) 由引理 21 并单独考虑 $d = 1$ 的情形，对于所选的 η_0，我们有

$$\mathbb{P}\big(|e_1^\top \eta_0| \leq 1/\sqrt{d \log n_{\mathrm{U}}}\big) \leq \sqrt{\frac{2}{\pi \log n_{\mathrm{U}}}}.$$

Again, if we choose $c \leq C_r^{-1} \min\{1/12, 1/(r+3)\}$, then $\phi\gamma^{1/2} = \omega \leq \min\{1/12, 1/(r+3)\}$. Also, $\|\hat{\mu}^{(0)}\| = \zeta_0 \vee r\omega_0 \in [C_r^{-1}(\zeta \vee r\omega) \ \zeta \vee r\omega]$. Thus, applying Proposition 13 with $c' = 1/C_r$, there exists $c > 0$, depending only on r, and $t_0 \in \mathbb{N}$ such that on $\Omega_1(\omega) \cap \Omega_2(\phi) \cap \{|e_1^\top \eta_0| > 1/\sqrt{d \log n_{\mathrm{U}}}\}$, we have $|\alpha_{t_0}| \geq cs$.

Since $\beta_0 \leq \|\hat{\mu}^{(0)}\| \leq \zeta \vee r\omega$, we can apply Proposition 10 to obtain that $\beta_t \leq 60(\zeta \vee r\omega)$ for all $t \geq 0$. Hence all conditions of Proposition 12 are satisfied, and the desired result then follows from (48). □

为证明定理6，我们需要以下命题，该命题将估计 μ^* 的损失与估计 $\mu^*\mu^{*\top}$的算子范数损失联系起来。

命题 14。 假设 $n \geq 3$，且 $(Z_1, Y_1, Y_1^*), \ldots, (Z_n, Y_n, Y_n^*)$ 相互独立，且

$$Y_i^* \sim \mathrm{Unif}\{-1, 1\}, \ Z_i \mid Y_i^* \sim \mathcal{N}_d(Y_i^* \mu^*, I_d), \ Y_i = Y_i^* \mathbb{1}_{\{i \leq n_{\mathrm{L}}\}} \quad \text{for } i \in [n].$$

For $\mu \in \mathbb{R}^d$ and $i \in [n]$, let $L_i(\mu) := Y_i \mathbb{1}_{\{Y_i \neq 0\}} + \tanh\langle Z_i, \mu\rangle \mathbb{1}_{\{Y_i = 0\}}$, $\mu_{\mathrm{tot}}(\mu) := \mu n^{-1} \sum_{i=1}^n L_i(\mu)$ and $\Sigma_{\mathrm{b}}(\mu) := \mu\mu^\top - \mu_{\mathrm{tot}}(\mu)\mu_{\mathrm{tot}}(\mu)^\top$. For any $\delta \in (0,1)$ and $B > 0$, we have with probability at least $1 - \delta$ that

$$\sup_{\mu : \|\mu\| \leq B} \big\{\|\Sigma_{\mathrm{b}}(\mu) - \mu^*\mu^{*\top}\|_{\mathrm{op}} - (B+s)L(\mu, \mu^*)\big\} \lesssim B^2(s^2 \vee 1)\left(\frac{d \log(2Bn + e) + \log(1/\delta)}{n}\right).$$

证明。 对于任意满足 $\|\mu\| \leq B$ 的 $\mu \in \mathbb{R}^d$，我们有

$$\begin{aligned}\|\Sigma_{\mathrm{b}}(\mu) - \mu^*\mu^{*\top}\|_{\mathrm{op}} &\leq \|\mu\mu^\top - \mu^*\mu^{*\top}\|_{\mathrm{op}} + \|\mu_{\mathrm{tot}}(\mu)\mu_{\mathrm{tot}}(\mu)^\top\|_{\mathrm{op}} \\ &\leq (B+s)L(\mu, \mu^*) + B^2\left(\frac{1}{n}\sum_{i=1}^n L_i(\mu)\right)^2.\end{aligned}$$

Thus, it is enough to show that $\sup_{\mu:\|\mu\|\le B} n^{-1}\sum_{i=1}^n L_i(\mu) \lesssim \sqrt{\frac{(s^2\vee 1)\{d\log(2Bn+e)+\log(1/\delta)\}}{n}}$ with probability at least $1-\delta$. To this end, we have

$$\sup_{\mu:\|\mu\|\le B}\frac{1}{n}\sum_{i=1}^n L_i(\mu) = \frac{1}{n}\sum_{i=1}^{n_{\mathrm{L}}} Y_i + \frac{1}{n}\sup_{\mu:\|\mu\|\le B}\sum_{i=n_{\mathrm{L}}+1}^n \tanh\langle Z_i,\mu\rangle. \tag{60}$$

对于60右侧的第一项，根据Hoeffding不等式，我们有

$$\mathbb{P}\biggl(\frac{1}{n}\sum_{i=1}^{n_{\mathrm{L}}} Y_i > \frac{\sqrt{2n_{\mathrm{L}}\log(3/\delta)}}{n}\biggr) \le \frac{\delta}{3}. \tag{61}$$

For the second term on the right-hand side of (60), let $\mathcal{N}$ be a ε-net of $\{v : \|v\| \le B\}$ with respect to the Euclidean distance, for some $\varepsilon \in (0, 1/2]$ to be specified later. Since a maximal ε-packing set is an ε-net, we may assume that $|\mathcal{N}| \le (B+\varepsilon/2)^d/(\varepsilon/2)^d = (2B/\varepsilon+1)^d$. Using the fact that $x \mapsto \tanh x$ is 1-Lipschitz, together with the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
&\sup_{v:\|v\|\le B}\sum_{i=n_{\mathrm{L}}+1}^n \tanh\langle Z_i, v\rangle \\
&\qquad\le \sup_{v\in\mathcal{N}}\sum_{i=n_{\mathrm{L}}+1}^n \tanh\langle Z_i, v\rangle + \sup_{u,v:\|u-v\|\le\varepsilon}\sum_{i=n_{\mathrm{L}}+1}^n (\tanh\langle Z_i,u\rangle - \tanh\langle Z_i,v\rangle) \\
&\qquad\le \sup_{v\in\mathcal{N}}\sum_{i=n_{\mathrm{L}}+1}^n \tanh\langle Z_i, v\rangle + \varepsilon\sum_{i=n_{\mathrm{L}}+1}^n \|Z_i\|.
\end{aligned}$$

因此取 $\varepsilon = 1/n$，并定义 $\tau := \log(3/\delta)\ d\log(2Bn+e) > 0$，我们有

$$\begin{aligned}
&\mathbb{P}\biggl(\frac{1}{n}\sup_{v:\|v\|\le B}\sum_{i=n_{\mathrm{L}}+1}^n \tanh\langle Z_i,v\rangle > 2\sqrt{\frac{2(s^2\vee 1)(1+\tau)d\log(2Bn+e)}{n}}\biggr) \\
&\le \mathbb{P}\biggl(\frac{1}{n}\sup_{v\in\mathcal{N}}\sum_{i=n_{\mathrm{L}}+1}^n \tanh\langle Z_i,v\rangle > \sqrt{\frac{2(s^2\vee 1)(1+\tau)d\log(2Bn+e)}{n}}\biggr) \\
&\qquad + \mathbb{P}\biggl(\frac{1}{n}\sum_{i=n_{\mathrm{L}}+1}^n \|Z_i\| \ge \sqrt{2(s^2\vee 1)(1+\tau)nd\log(2Bn+e)}\biggr) \\
&\le \frac{|\mathcal{N}|}{e^{(1+\tau)d\log(2Bn+e)}} + \mathbb{P}\biggl(\frac{1}{n}\sum_{i=n_{\mathrm{L}}+1}^n \|Z_i\|^2 \ge 2(s^2\vee 1)(1+\tau)nd\log(2Bn+e)\biggr) \le \frac{2\delta}{3},
\end{aligned} \tag{62}$$

where the penultimate bound uses Hoeffding's inequality and the Cauchy–Schwarz inequality and the final bound uses the fact that $\sum_{i=n_{\mathrm{L}}+1}^n \|Z_i\|^2 \sim \chi^2_{n_{\mathrm{U}}d}(n_{\mathrm{U}}s^2)$ and Birgé (2001, Lemma 8.1). Combining (60), (61) and (62), we have with probability at least $1-\delta$ that

$$\begin{aligned}
\sup_{\mu:\|\mu\|\le B}\frac{1}{n}\sum_{i=1}^n L_i(\mu) &\le \frac{\sqrt{2n_{\mathrm{L}}\log(3/\delta)}}{n} + 2\sqrt{\frac{2(s^2\vee 1)(1+\tau)d\log(2Bn+e)}{n}} \\
&\lesssim \sqrt{\frac{(s^2\vee 1)\{d\log(2Bn+e)+\log(1/\delta)\}}{n}},
\end{aligned}$$

如所需。 □

Proof of Theorem 6. We write $\hat{\mu}_{[m]} \equiv \hat{\mu}_{[m]}^{(T)}$ for the Tth (final) iterate of the EM update in Algorithm 3 starting from the mth random initializer $\hat{\mu}_{[m]}^{(0)}$. Let $\omega = C_r\omega_0$, $\phi = C_r\phi_0$, $\Omega_1(\omega)$ and $\Omega_2(\phi)$ be defined as in the proof of Proposition 5. Further, let $\Sigma_{\mathrm{b}}(\mu)$ be defined as in Proposition 14. By Proposition 9, the first claim in the proof of Proposition 10 and Proposition 14, we have for some $C > 0$ depending only on r that

$$\mathbb{P}\biggl[\max_{m\in[M]}\sup_{T\in\mathbb{N}}\Bigl\{\|\Sigma_{\mathrm{b}}(\hat{\mu}_{[m]}^{(T)}) - \mu^*\mu^{*\top}\|_{\mathrm{op}} - (2r+3)L(\hat{\mu}_{[m]}^{(T)}, \mu^*)\Bigr\} \le \frac{C\{d\log(rn) + \log(1/\delta)\}}{n}\biggr] \ge 1-\delta.$$

In this balanced two-cluster setup, for the tth EM iteration starting from the mth random initializer, we have $-\hat{\mu}_1 = \hat{\mu}_2 = \hat{\mu}$, where we suppress the dependence on t and m for convenience. For $i \ge n_{\mathrm{L}} + 1$, we have $L_{i,1} = e^{z_i^\top\hat{\mu}_1}/(e^{z_i^\top\hat{\mu}_1} + e^{z_i^\top\hat{\mu}_2})$, $L_{i,2} = e^{z_i^\top\hat{\mu}_2}/(e^{z_i^\top\hat{\mu}_1} + e^{z_i^\top\hat{\mu}_2})$ and hence $L_{i,2} - L_{i,1} = \tanh\langle Z_i, \hat{\mu}\rangle$. Thus, $\hat{\mu}_{\mathrm{tot}} = n^{-1}\hat{\mu}\sum_{i=1}^n\{\tanh\langle Z_i,\hat{\mu}\rangle\mathbb{1}_{\{Y_i=0\}} + Y_i\mathbb{1}_{\{Y_i\neq 0\}}\}$. Also, we note that

$$\hat{\Sigma}_{\mathrm{b}} = \frac{1}{n}\sum_{i=1}^n\sum_{k=1}^2 L_{i,k}(\hat{\mu}_k - \hat{\mu}_{\mathrm{tot}})(\hat{\mu}_k - \hat{\mu}_{\mathrm{tot}})^\top = \frac{1}{n}\sum_{i=1}^n\sum_{k=1}^2 L_{i,k}\hat{\mu}_k\hat{\mu}_k^\top - \hat{\mu}_{\mathrm{tot}}\hat{\mu}_{\mathrm{tot}}^\top = \hat{\mu}\hat{\mu}^\top - \hat{\mu}_{\mathrm{tot}}\hat{\mu}_{\mathrm{tot}}^\top.$$

因此，使用命题 14和算法 3的记号，我们有 $\hat{Q} \equiv \hat{Q}^{(T)} \equiv \hat{Q}^{[\hat{m}]} = \Sigma_{\mathrm{b}}\bigl(\hat{\mu}_{[\hat{m}]}^{(T)}\bigr)$。

We consider two cases. If $\|\mu^*\| \le \omega_0^{1/3} \wedge \zeta_0^{1/2}$, then by the proof of Proposition 5*(i)*, we have on the event $\Omega_1(\omega)\cap\Omega_2(\phi)$ that $\limsup_{T\to\infty} L(\hat{\mu}_{[m]}^{(T)}, \mu^*) \lesssim_r \zeta_0 \vee \|\mu^*\| \lesssim_r \omega_0^{1/3}\wedge\zeta_0^{1/2}$ for every $m \in [M]$. Thus, by Proposition 14, with probability at least $\mathbb{P}\bigl(\Omega_1(\omega)\cap\Omega_2(\phi)\bigr) - \delta \ge 1 - 3\delta$, we have

$$\begin{aligned}\limsup_{T\to\infty}\|\hat{Q} - \mu^*\mu^{*\top}\|_{\mathrm{op}} &\lesssim_r (\omega_0^{1/3}\wedge\zeta_0^{1/2})\limsup_{T\to\infty}\max_{m\in[M]} L(\hat{\mu}_{[m]}^{(T)}, \mu^*) \\ &\quad + (\omega_0^{2/3}\wedge\zeta_0)\biggl(\frac{d\log n + \log(1/\delta)}{n}\biggr) \\ &\lesssim_r \omega_0^{2/3}\wedge\zeta_0 = \frac{\omega_0}{\omega_0^{1/3}\wedge\zeta_0^{1/2}}\wedge\zeta_0 \le \frac{\omega_0}{\|\mu^*\|}\wedge\zeta_0. \end{aligned} \tag{63}$$

We now turn to the case where $\|\mu^*\| > \omega_0^{1/3}\wedge\zeta_0^{1/2}$. Let $\mathcal{M}_0$ be the set of $m \in [M]$ such that $|\langle\hat{\mu}_{[m]}^{(0)}, \mu^*\rangle|/(\|\mu^*\|\|\hat{\mu}_{[m]}^{(0)}\|) \ge \sqrt{1/(d\log n_{\mathrm{U}})}$ and let $M_0 := |\mathcal{M}_0|$. By definition of the EM initializers, the random variables $\bigl\{\langle\hat{\mu}_{[m]}^{(0)}, \mu^*\rangle/(\|\mu^*\|\|\hat{\mu}_{[m]}^{(0)}\|) : m \in [M]\bigr\}$ are independent, and moreover, by Lemma 21 we have

$$\mathbb{P}\biggl(\frac{|\langle\hat{\mu}_{[m]}^{(0)}, \mu^*\rangle|}{\|\mu^*\|\|\hat{\mu}_{[m]}^{(0)}\|} \ge \sqrt{\frac{1}{d\log n_{\mathrm{U}}}}\biggr) \ge 1 - \sqrt{\frac{2}{\pi\log n_{\mathrm{U}}}} > \frac{3}{5}.$$

定义 $\Omega_3 := \{M_0 > M/2\}$，根据 Hoeffding不等式，我们有

$$\mathbb{P}(\Omega_3^{\mathrm{c}}) \le e^{-M/50}.$$

Let

$$\mathcal{M}_1 := \bigl\{m \in [M]\setminus\{\hat{m}\} : \|\hat{Q}^{[m]} - \hat{Q}^{[\hat{m}]}\|_{\mathrm{op}} \le \mathrm{median}(\|\hat{Q}^{[m']} - \hat{Q}^{[\hat{m}]}\|_{\mathrm{op}} : m' \in [M]\setminus\{\hat{m}\})\bigr\}.$$

Since $|\mathcal{M}_1 \cup \{\hat{m}\}| \geq \lceil (M-1)/2 \rceil + 1 > M/2$, we have on Ω_3 that $\mathcal{M}_0 \cap (\mathcal{M}_1 \cup \{\hat{m}\}) \neq \emptyset$. Thus, on the event $\Omega_1(\omega) \cap \Omega_2(\phi) \cap \Omega_3$, we can let $\tilde{m} := \min\big(\mathcal{M}_0 \cap (\mathcal{M}_1 \cup \{\hat{m}\})\big)$, so by definition of $\hat{m}$, we have

$$\begin{aligned}
\|\hat{Q}^{[\hat{m}]} - \mu^*\mu^{*\top}\|_{\mathrm{op}} &\leq \|\hat{Q}^{[\hat{m}]} - \hat{Q}^{[\tilde{m}]}\|_{\mathrm{op}} + \|\hat{Q}^{[\tilde{m}]} - \mu^*\mu^{*\top}\|_{\mathrm{op}} \\
&\leq \mathrm{median}(\|\hat{Q}^{[m']} - \hat{Q}^{[\hat{m}]}\|_{\mathrm{op}} : m' \in [M] \setminus \{\hat{m}\}) + \|\hat{Q}^{[\tilde{m}]} - \mu^*\mu^{*\top}\|_{\mathrm{op}} \\
&\leq \mathrm{median}(\|\hat{Q}^{[m']} - \hat{Q}^{[\tilde{m}]}\|_{\mathrm{op}} : m' \in [M] \setminus \{\tilde{m}\}) + \|\hat{Q}^{[\tilde{m}]} - \mu^*\mu^{*\top}\|_{\mathrm{op}} \\
&\leq \max_{m,m' \in \mathcal{M}_0} \|\hat{Q}^{[m]} - \hat{Q}^{[m']}\|_{\mathrm{op}} + \|\hat{Q}^{[\tilde{m}]} - \mu^*\mu^{*\top}\|_{\mathrm{op}} \\
&\leq 3 \max_{m \in \mathcal{M}_0} \|\hat{Q}^{[m]} - \mu^*\mu^{*\top}\|_{\mathrm{op}}.
\end{aligned}$$

Since $\omega_0 \leq (d \log n)^{-3}$, by discussing cases of $\gamma < \omega$, $\omega \leq \gamma \leq \omega^{2/3}$ and $\gamma > \omega^{2/3}$, we see that $\omega_0^{1/3} \wedge \zeta_0^{1/2} \geq \zeta_0 \sqrt{d \log n}$. From the proof of Proposition 5*(ii)*, we have on the event $\Omega_1(\omega) \cap \Omega_2(\phi)$ that $\limsup_{T\to\infty} \max_{m \in \mathcal{M}_0} L(\hat{\mu}^{(T)}_{[m]}, \mu^*) \lesssim_r \frac{\omega_0}{\|\mu^*\|} \wedge (\omega_0 \gamma^{-1/2})$. Let Ω_4 be the event on which the conclusion of Proposition 14 holds. Then on $\Omega_1(\omega) \cap \Omega_2(\phi) \cap \Omega_3 \cap \Omega_4$, we therefore have

$$\begin{aligned}
\limsup_{T\to\infty} \|\hat{Q}^{[\hat{m}]} - \mu^*\mu^{*\top}\|_{\mathrm{op}} &\lesssim_r \limsup_{T\to\infty} \max_{m \in \mathcal{M}_0} L(\hat{\mu}^{(T)}_{[m]}, \mu^*) + \frac{d \log n + \log(1/\delta)}{n} \\
&\lesssim_r \left(\frac{\omega_0}{\|\mu^*\|} \wedge \frac{\omega_0}{\gamma^{1/2}} \right) + \frac{d \log n + \log(1/\delta)}{n} \\
&\lesssim_r \frac{\omega_0}{\|\mu^*\|} \wedge \zeta_0. \qquad (64)
\end{aligned}$$

所需结果可通过结合（63）和（64），以及以下事实得到：$\mathbb{P}\big(\Omega_1(\omega) \cap \Omega_2(\phi) \cap \Omega_3 \cap \Omega_4\big) \geq 1 - 3\delta - e^{-M/50}$。 □

5.5 推论的证明 7

Proof of Corollary 7. Fix $P \in \mathcal{P}_d$, define $Z_i := PX_i$ for $i \in [n]$, $\mu^* := P\nu^*$, $\delta := \varepsilon / \big\{4\binom{p}{d}\big\}$, $\omega_0 := \sqrt{\frac{d \log n + \log(1/\delta)}{n_{\mathrm{U}}}}$ and $\zeta_0 := \omega_0^{1/2} \wedge \omega_0 \gamma^{-1/2}$. Then, provided $C_1 > 2$, we have $\delta \geq 2e^{-n/2}/p^d > 2e^{-n}$, and $\|\mu^*\| \leq \|\nu^*\| \leq r$. Let $c > 0$ be chosen, depending only on r, to satisfy Theorem 6. By increasing $C_1 > 0$, depending only on r, if necessary, we may assume that $\omega_0 \leq \min\{c, (d \log n)^{-3}\}$. Hence, since $(P\Sigma_{\mathrm{w}}P^\top)^{-1} P\Sigma_{\mathrm{b}}P^\top = \mu^*\mu^{*\top}$, we can apply Theorem 6 to obtain that for some $C_2' > 0$ depending only on r, with probability at least $1 - 3\delta - e^{-M/50}$ we have that

$$\begin{aligned}
&\limsup_{T\to\infty} \big\| \psi^{(T)}\big((PX_i, Y_i)_{i \in [n]}\big) - (P\Sigma_{\mathrm{w}}P^\top)^{-1} P\Sigma_{\mathrm{b}}P^\top \big\|_{\mathrm{op}} \leq C_2' \zeta_0 \\
&\qquad \leq C_2 \min\left[\left\{ \frac{d \log(p \vee n) + \log(1/\varepsilon)}{n} \right\}^{1/4}, \sqrt{\frac{d \log(p \vee n) + \log(1/\varepsilon)}{n_{\mathrm{L}}}} \right] \leq \frac{(\nu^*_{\min})^2}{4}.
\end{aligned}$$

Since $\psi^{(T)}$ is permutation equivariant for each $T \geq 0$, by Fatou's lemma and a union bound, we have that

$$\begin{aligned}
&\limsup_{T\to\infty} \mathbb{P}\left(\max_{P\in\mathcal{P}_d} \left\| \psi^{(T)}\big((PX_i, Y_i)_{i\in[n]}\big) - (P\Sigma_{\mathrm{w}}P^\top)^{-1}P\Sigma_{\mathrm{b}}P^\top \right\|_{\mathrm{op}} > \frac{(\nu^*_{\min})^2}{4} \right) \\
&\quad \leq \mathbb{P}\left(\limsup_{T\to\infty} \max_{P\in\mathcal{P}_d} \left\| \psi^{(T)}\big((PX_i, Y_i)_{i\in[n]}\big) - (P\Sigma_{\mathrm{w}}P^\top)^{-1}P\Sigma_{\mathrm{b}}P^\top \right\|_{\mathrm{op}} > \frac{(\nu^*_{\min})^2}{4} \right) \\
&\quad \leq \sum_{P\in\mathcal{P}_d} \mathbb{P}\left(\limsup_{T\to\infty} \left\| \psi^{(T)}\big((PX_i, Y_i)_{i\in[n]}\big) - (P\Sigma_{\mathrm{w}}P^\top)^{-1}P\Sigma_{\mathrm{b}}P^\top \right\|_{\mathrm{op}} > \frac{(\nu^*_{\min})^2}{4} \right) \\
&\quad \leq \binom{p}{d}(3\delta + e^{-M/50}) \leq \frac{3}{4}\varepsilon + e^{-M/50 + d\log p} \leq \varepsilon.
\end{aligned}$$

现在的结果由定理 2 得出，注意到 $\gamma_{\min} = (\nu^*_{\min})^2$ 且 $\gamma_{\max} = (\nu^*_{\max})^2$。

□

6 辅助引理

Lemma 15. *Suppose that $K = 2$ and $\mathcal{C}$ is defined as in (11). Let $(-\hat{\mu}^{(t)}, \hat{\mu}^{(t)}, I_d) \in \mathcal{C}$ be the tth iterate of the EM iteration described in (4) and (5) with data $(Z_1, Y_1) \ldots, (Z_n, Y_n)$, starting from $(-\hat{\mu}^{(0)}, \hat{\mu}^{(0)}, I_d)$. Then for all $t \geq 1$, we have*

$$\hat{\mu}^{(t)} = \frac{1}{n}\left\{ \sum_{i:Y_i\neq 0} (-1)^{Y_i} Z_i + \sum_{i:Y_i=0} Z_i \tanh\langle Z_i, \hat{\mu}^{(t-1)}\rangle \right\}.$$

证明。 在步骤 $t \geq 1$ 中，在 E步中，根据（4），我们有对于 $k \in \{1, 2\}$，$L_{i,\,k} = \mathbb{1}_{\{Y_i=k\}}$ 如果$Y_i \neq 0$ 且

$$L_{i,k} = \frac{e^{-\|Z_i - (-1)^k \hat{\mu}^{(t-1)}\|^2/2}}{e^{-\|Z_i - \hat{\mu}^{(t-1)}\|^2/2} + e^{-\|Z_i + \hat{\mu}^{(t-1)}\|^2/2}}$$

否则。在M步中，定义

$$Q(\mu \mid \hat{\mu}^{(t-1)}) := \frac{1}{n}\sum_{i=1}^{n} (L_{i,1}\|Z_i + \mu\|^2 + L_{i,2}\|Z_i - \mu\|^2),$$

我们有 $\hat{}\mu^{(t)} = \mathsf{argmin}_{\mu\in\mathbb{R}^d} Q(\mu \mid \hat{}\mu^{(t-1)})$。对 $Q(\mu \mid \hat{}\mu^{(t-1)})$ 关于 μ 求导，我们得到

$$\hat{\mu}^{(t)} = \frac{1}{n}\sum_{i=1}^{n} (L_{i,2} - L_{i,1}) Z_i.$$

所需结果随之成立，因为 $L_{i,\,2} - L_{i,\,1} = (-1)^{Y_i}$ 若 $Y_i \in \{1, 2\}$，且

$$L_{i,2} - L_{i,1} = \frac{e^{-\|Z_i - \hat{\mu}^{(t-1)}\|^2/2} - e^{-\|Z_i + \hat{\mu}^{(t-1)}\|^2/2}}{e^{-\|Z_i - \hat{\mu}^{(t-1)}\|^2/2} + e^{-\|Z_i + \hat{\mu}^{(t-1)}\|^2/2}} = \frac{e^{\langle Z_i, \hat{\mu}^{(t-1)}\rangle} - e^{-\langle Z_i, \hat{\mu}^{(t-1)}\rangle}}{e^{\langle Z_i, \hat{\mu}^{(t-1)}\rangle} + e^{-\langle Z_i, \hat{\mu}^{(t-1)}\rangle}} = \tanh\langle Z_i, \hat{\mu}^{(t-1)}\rangle$$

如果 $Y_i = 0$。

□

Lemma 16. *Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$ for some distribution P on $\mathbb{R}^d$. If $\|\mu^*\| \leq n^{-1/4}$, then for any Borel measurable function $\psi : (\mathbb{R}^d)^n \to \{0, 1\}$ of the null hypothesis $H_0 : P = \mathcal{N}_d(0, I_d)$ against the alternative $H_1 : P = \frac{1}{2}\mathcal{N}_d(\mu^*, I_d) + \frac{1}{2}\mathcal{N}_d(-\mu^*, I_d)$, we have*

$$\mathbb{P}_{H_0}\big(\psi(X_1, \ldots, X_n) = 1\big) + \mathbb{P}_{H_1}\big(\psi(X_1, \ldots, X_n) = 0\big) > 1/2.$$

证明。 写作 $X = (X_1, \ldots, X_n)^\top$。注意，记 d_{TV} 为概率测度之间的全变差距离，

$$\begin{aligned}
\mathbb{P}_{H_0}\big(\psi(X)=1\big) + \mathbb{P}_{H_1}\big(\psi(X)=0\big) &\geq 1 - d_{\mathrm{TV}}(\mathbb{P}_{H_0}, \mathbb{P}_{H_1}) \\
&= 1 - \frac{1}{2}\int \left| \frac{d\mathbb{P}_{H_1}}{d\mathbb{P}_{H_0}} - 1 \right| d\mathbb{P}_{H_0} \geq 1 - \frac{1}{2}\left\{ \int \left(\frac{d\mathbb{P}_{H_1}}{d\mathbb{P}_{H_0}} - 1 \right)^2 d\mathbb{P}_{H_0} \right\}^{1/2} \\
&= 1 - \frac{1}{2}\left\{ \int \left(\frac{d\mathbb{P}_{H_1}}{d\mathbb{P}_{H_0}} \right)^2 d\mathbb{P}_{H_0} - 1 \right\}^{1/2}. \qquad (65)
\end{aligned}$$

To control the chi-squared divergence in the right-hand side of (65) above, we let $\xi = (\xi_1, \ldots, \xi_n)^\top$ have independent Rademacher components and $W = (W_{i,j})_{i\in[n], j\in[d]}$ be a random matrix with independent $N(0,1)$ entries, independent of ξ. Then $X \overset{\mathrm{d}}{=} W$ under H_0 and $X \overset{\mathrm{d}}{=} \xi\mu^{*\top} + W$ under H_1. Let $\tilde{\xi}$ be an independent copy of ξ. Using the Ingster–Suslina device, see, e.g., Ingster and Suslina (2012), Liu, Gao and Samworth (2021), Lemma 21, we have that

$$\int \left(\frac{d\mathbb{P}_{H_1}}{d\mathbb{P}_{H_0}} \right)^2 d\mathbb{P}_{H_0} = \mathbb{E}\exp\langle \xi\mu^{*\top}, \tilde{\xi}\mu^{*\top} \rangle = \cosh^n(\|\mu^*\|^2) \leq e^{n\|\mu^*\|^4/2} \leq e^{1/2},$$

其中，我们在倒数第二步中用到了双曲余弦函数 $x \leq e^{x^2/2}$ 对所有 $x \in \mathbb{R}$ 成立这一事实。将上述界代入（65）并利用 $1 - (e^{1/2} - 1)^{1/2}/2 > 1/2$，即可得到所需结果。 □

We prove a generalization of Cochran's theorem for quadratic forms of independent Gaussian random vectors with a common covariance matrix, which result in independent noncentral Wishart distributions. Recall that if X is a matrix, then $\mathrm{vec}(X)$ is the vectorization of X, obtained by stacking its columns on top of each other. The Kronecker product between matrices $A = (A_{i,j})_{i\in[m], j\in[n]}$ and B is defined as

$$A \otimes B := \begin{pmatrix} A_{1,1}B & \cdots & A_{1,n}B \\ \vdots & \ddots & \vdots \\ A_{m,1}B & \cdots & A_{m,n}B \end{pmatrix}.$$

Recall also that when $X_1, \ldots, X_n \overset{\mathrm{iid}}{\sim} \mathcal{N}_d(0, \Sigma)$, the matrix $\sum_{i=1}^n X_i X_i^\top$ has a d-dimensional Wishart distribution with n degrees of freedom and covariance matrix $\Sigma \in \mathbb{S}^{d\times d}$, denoted $\mathcal{W}_d(n, \Sigma)$. More generally, $\sum_{i=1}^n (X_i + \mu_i)(X_i + \mu_i)^\top$ has a non-central Wishart distribution with n degrees of freedom, covariance matrix Σ and non-centrality matrix $\Omega = \sum_{i=1}^n \mu_i \mu_i^\top$, written $\mathcal{W}_d(n, \Sigma; \Omega)$. Thus $\mathcal{W}_d(n, \Sigma; 0) \overset{d}{=} \mathcal{W}_d(n, \Sigma)$.

Lemma 17. *Let $Z_1, \ldots, Z_n$ be independent with $Z_i \sim \mathcal{N}_d(\mu_i, \Sigma)$ for $i \in [n]$, and write $Z := (Z_1, \ldots, Z_n)^\top \in \mathbb{R}^{n\times d}$ and $M := \mathbb{E}(Z)$. If $P_1, \ldots, P_k \in \mathbb{R}^{n\times n}$ are positive semidefinite matrices such that $P_1 + \cdots + P_k = I_n$ and $\mathrm{rank}(P_1) + \cdots + \mathrm{rank}(P_k) = n$, then $Z^\top P_1 Z, \ldots, Z^\top P_k Z$ are independent with $Z^\top P_r Z \sim \mathcal{W}_d\big(\mathrm{rank}(P_r)\ \Sigma; M^\top P_r M\big)$.*

证明。 与经典的Cochran定理（Cochran，1934）的证明一样，我们首先注意到 $P_1, \ldots, P_k$ 可以同时对角化，使得

$$P_r = Q D_r Q^\top,$$

for some $Q \in \mathbb{O}^{n\times n}$ and $D_r = \mathrm{diag}\big((\mathbb{1}_{\{j\in S_r\}})_{j\in[n]}\big)$, where $S_r \subseteq [n]$, $|S_r| = \mathrm{rank}(P_r)$ and $S_r \cap S_{r'} = \emptyset$ for all $r \neq r'$. In particular, $P_1, \ldots, P_k$ satisfy $P_r^2 = P_r$ for $r \in [k]$ and $P_r P_{r'} = 0$ for all $r \neq r'$. Since $P_1 Z, \ldots, P_k Z$ are jointly Gaussian, with

$$\begin{aligned}\mathrm{Cov}\big(\mathrm{vec}(P_r Z), \mathrm{vec}(P_{r'} Z)\big) &= \mathrm{Cov}\big((I_d \otimes P_r)\mathrm{vec}(Z), (I_d \otimes P_{r'})\mathrm{vec}(Z)\big) \\ &= (I_d \otimes P_r)(\Sigma \otimes I_n)(I_d \otimes P_{r'})^\top = 0,\end{aligned}$$

we have that $P_1 Z, \ldots, P_k Z$ are independent. But $Z^\top P_r Z = (P_r Z)^\top P_r Z$, so it follows that $Z^\top P_1 Z, \ldots, Z^\top P_k Z$ are independent. Moreover, writing $W = (W_1, \ldots, W_n)^\top := Q^\top Z$, we have $\mathrm{vec}(Z) \sim \mathcal{N}_{nd}\big(\mathrm{vec}(M)\ \Sigma \otimes I_n\big)$, so

向量化(W)(I_d ⊗ Q^⊤)vec(Z) ∼ N_{nd}((I_d ⊗ Q^⊤)vec(M), Σ ⊗ I_n) =^d N_{nd}(vec(Q^⊤M), Σ ⊗ I_n

因此，

$$\begin{aligned}Z^\top P_r Z = W^\top D_r W = \sum_{i\in S_r} W_i W_i^\top &\sim \mathcal{W}_d\bigg(|S_r|, \Sigma; \sum_{i\in S_r} \mathbb{E}(W_i)\mathbb{E}(W_i)^\top\bigg) \\ &\overset{d}{=} \mathcal{W}_d\big(\mathrm{rank}(P_r), \Sigma; M^\top P_r M\big),\end{aligned}$$

如所需。 □

引理 18。 对于任意 a、$b \in \mathbb{R}$，我们有

$$\frac{1}{2}\{\tanh(a+b) - \tanh(a-b)\} \leq |b|$$

and

$$\frac{a}{2}\{\tanh(a+b) + \tanh(a-b)\} \geq a^2 - \frac{a^4}{3} - a^2 b^2$$

Proof. For the first inequality, since the left-hand side is an increasing function of b, and an even function of a, we may assume that $a \geq 0$ and $b \geq 0$. Notice that $\frac{\partial}{\partial a}\big(\tanh(a+b) - \tanh(a-b)\big) = 1/\cosh^2(a+b) - 1/\cosh^2(|a-b|) \leq 0$, since $x \mapsto \cosh(x)$ is an increasing function on $[0, \infty)$. Hence

$$\frac{1}{2}\{\tanh(a+b) - \tanh(a-b)\} \leq \tanh b \leq b.$$

如愿。

对于第二个不等式，由于两边都是关于 a 和 b 的偶函数，我们再次可以不失一般性地假设 $a > 0$ 和 $b \geq 0$。我们还可以假设 $b \leq 1$，因为否则右侧为负，不等式显然成立。但此时

$$\begin{aligned}\frac{1}{2}\{\tanh(a+b) + \tanh(a-b)\} &= \frac{1}{2}\bigg(\frac{\tanh a + \tanh b}{1 + \tanh a \tanh b} + \frac{\tanh a - \tanh b}{1 - \tanh a \tanh b}\bigg) \\ &= \frac{\tanh a}{(1 - \tanh^2 a \tanh^2 b)\cosh^2 b} \geq \frac{\tanh a}{\cosh^2 b} \\ &\geq (1 - b^2)\tanh a \geq (1 - b^2)\bigg(a - \frac{a^3}{3}\bigg) \geq a - \frac{a^3}{3} - ab^2,\end{aligned}$$

如愿 d. 此处，第二个不等式成立是因为 $(1-b^2)$双曲余弦函数$^2\, b \leq (1-b^2)e^{b^2} \leq 1$。 □

Lemma 19. *Let $H : [0,\infty) \to [0,\infty)$ be an increasing, concave function with $H'(x_0) < 1$ for some $x_0 \geq 0$ and either $H(0) > 0$ or both $H(0) = 0$ and $H'(0) > 1$. Then there exists a unique $\alpha^* > 0$ such that*

$$H(\alpha) - \alpha \begin{cases} > 0 & \alpha \in (0, \alpha^*) \\ = 0 & \alpha = \alpha^* \\ < 0 & \alpha \in (\alpha^*, \infty). \end{cases}$$

此外，如果 $\alpha_0 > 0$，则由 $\alpha_t := H(\alpha_{t-1})$ 给出的序列 $(\alpha_t)_{t\geq 0}$ 单调收敛到 α^*。

Proof. For the first claim, consider the concave function $\tilde{H}(x) := H(x) - x$, which satisfies $\tilde{H}(x) > 0$ for sufficiently small $x > 0$, and for $x \geq x_0$, we have that any supergradient $v_x \in \mathbb{R}$ of $\tilde{H}$ at x satisfies $v_x \leq -\{1 - H'(x_0)\} < 0$. It follows that $\tilde{H}(x) \to -\infty$ as $x \to \infty$, so by the intermediate value theorem, there exists $\alpha^* \in (0,\infty)$ such that $\tilde{H}(\alpha^*) = 0$, i.e. $H(\alpha^*) = \alpha^*$. Again using the facts that $\tilde{H}(x) > 0$ for sufficiently small $x > 0$, and $\tilde{H}(x) \to -\infty$ as $x \to \infty$, we see that the concave function $\tilde{H}$ can only cross the x-axis at one positive value α^*, and $\tilde{H}(\alpha) > 0$ for $\alpha \in (0, \alpha^*)$ and $\tilde{H}(\alpha) < 0$ for $\alpha \in (\alpha^*, \infty)$.

Next, note that if $\alpha \in (0, \alpha^*)$, then $\alpha < H(\alpha) < H(\alpha^*) = \alpha^*$. Thus, if $\alpha_0 < \alpha^*$, then $(\alpha_t)_{t\geq 0}$ is an increasing sequence, bounded above by α^*, so it converges to a limit. But then, taking limits on both sides of the recursion $\alpha_t := H(\alpha_{t-1})$, we deduce that this limit must be α^*. A similar argument can be used to show that if $\alpha_0 \in (\alpha^*, \infty)$ then $(\alpha_t)_{t\geq 0}$ decreases down to the limit α^*, while if $\alpha_0 = \alpha^*$, then $\alpha_t = \alpha^*$ for all t. □

Lemma 20. *Let μ^* be a non-zero vector in $\mathbb{R}^d$, let $Z \sim \frac{1}{2}\mathcal{N}_d(-\mu^*, I_d) + \frac{1}{2}\mathcal{N}_d(\mu^*, I_d)$, let $\eta := \mu^*/\|\mu^*\|$, and define $q : [0,\infty) \to [0,\infty)$ by*

$$q(\alpha) := \begin{cases} \alpha^{-1}\eta^\top \mathbb{E}(Z \tanh\langle \alpha\eta, Z\rangle) & \text{if } \alpha > 0 \\ 1 + \|\mu^*\|^2 & \text{if } \alpha = 0. \end{cases}$$

则 q 是一个可微函数，满足 $q(\|\mu^*\|) = 1$，并且对于任意 $h \geq \|\mu^*\|$，我们有

$$\sup_{\alpha\in[0,h]} \frac{q'(\alpha)}{\alpha} \leq -\frac{e^{-h^2/2}}{3 \cdot 2^{11}\sqrt{2\pi}(h^5 \vee 1)}.$$

Proof. Write $s := \|\mu^*\|$. The fact that $q(s) = 1$ follows from Xu, Hsu and Maleki (2016, Theorem 1). By Wu and Zhou (2022, Lemma 3(4)), q is differentiable with $q'(\alpha) \leq -(2\alpha/3) \cdot \mathbb{E}\big(\tilde{Z}^4/\cosh^2(\alpha\tilde{Z})\big)$ for $\alpha \in [0,\infty)$, where $\tilde{Z} \sim \frac{1}{2}\mathcal{N}(-s, 1) + \frac{1}{2}\mathcal{N}(s, 1)$. We can

现在计算 α， $s \in [0, h]$,

$$\begin{aligned}
\mathbb{E}\left(\frac{\tilde{Z}^4}{\cosh^2(\alpha\tilde{Z})}\right) &\geq \mathbb{E}(\tilde{Z}^4 e^{-2\alpha|\tilde{Z}|}) \geq \frac{1}{2\sqrt{2\pi}} \int_0^\infty y^4 e^{-2\alpha y} e^{-(y-s)^2/2}\, dy \\
&= \frac{1}{2\sqrt{2\pi}} \int_0^\infty y^4 e^{-(y-s+2\alpha)^2/2 - 2\alpha s + 2\alpha^2}\, dy \\
&\geq \begin{cases} \frac{e^{-2\alpha s+2\alpha^2}}{2\sqrt{2\pi}} \int_0^{\frac{1}{2(2\alpha-s)}} \frac{1}{2} y^4 e^{-(2\alpha-s)^2/2}\, dy & \text{if } s+1 \leq 2\alpha \\ \frac{e^{-2\alpha s+2\alpha^2}}{2\sqrt{2\pi}} \int_{s-2\alpha+1}^{s-2\alpha+2} y^4 e^{-2}\, dy & \text{if } s+1 > 2\alpha \end{cases} \\
&\geq \begin{cases} \frac{e^{-s^2/2}}{2^7 \cdot 5\sqrt{2\pi}} \left(\frac{1}{2\alpha-s}\right)^5 & \text{if } s+1 \leq 2\alpha \\ \frac{e^{-2\alpha s+2\alpha^2-2}}{10\sqrt{2\pi}} & \text{if } s+1 > 2\alpha \end{cases} \\
&\geq \frac{e^{-h^2/2}}{2^{12} \cdot 5\sqrt{2\pi}(h^5 \vee 1)},
\end{aligned}$$

从而建立了所需的界限。 □

引理 21。 设 $d \geq 2$，且 $\eta = (\eta_1, \ldots, \eta_d)^\top \sim$均匀分布$(\mathbb{S}^{d-1})$。那么对于任意 $a > 0$，有

$$\mathbb{P}\left(|\eta_1| \leq \frac{a}{\sqrt{d}}\right) \leq \sqrt{\frac{2}{\pi}} a.$$

Proof. Leting $Z = (Z_1, \ldots, Z_d)^\top \sim N_d(0, I_d)$, we have $\eta \overset{\mathrm{d}}{=} Z/\|Z\|$ and in particular $\eta_1^2 \overset{\mathrm{d}}{=} Z_1^2/(Z_1^2 + \cdots + Z_d^2) \sim \mathrm{Beta}\big(1/2, (d-1)/2\big)$. Thus,

$$\begin{aligned}
\mathbb{P}\left(|\eta_1| \leq \frac{a}{\sqrt{d}}\right) = \mathbb{P}\left(\frac{Z_1^2}{\|Z\|^2} \leq \frac{a^2}{d}\right) &= \frac{\Gamma(d/2)}{\Gamma(1/2)\Gamma\big((d-1)/2\big)} \int_0^{a^2/d} t^{-1/2}(1-t)^{(d-3)/2}\, dt \\
&\leq \frac{2\Gamma(d/2)a}{\sqrt{d}\Gamma(1/2)\Gamma((d-1)/2)} \leq \sqrt{\frac{2}{\pi}} a,
\end{aligned}$$

其中最终的界限使用了例如D¨umbgen、Samworth 和 Wellner(2021年，推论 11）。 □

参考文献

Ahfock, D. C., Astle, W.J. 和Richardson, S. (2021年) 草图算法的统计性质。 *Bio metrika*, 108, 283–297.

Akcay, S.、Atapour-Abarghouei, A. 和 Breckon, T. P.（2019年）Ganomaly：通过对抗训练进行半监督异常检测。载于第*14*届亚洲计算机视觉会议，修订精选论文，第三部分 *14*，622–637，Springer。

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Nat. Acad. Sci.*, **96**, 6745–6750.

Anderlucci, L., Fortunato, F. and Montanari, A. (2022) High-dimensional clustering via random projections. *J. Classification*, **39**, 191–216.

Anderson, T. W. (2003) *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics.

Azizyan, M., Singh, A. and Wasserman, L. (2013) Minimax theory for high-dimensional Gaussian mixtures with sparse mean separation. In *Adv. Neur. Inf. Proc. Syst.*, 2139–2147.

Azizyan, M., Singh, A. and Wasserman, L. (2015) Efficient sparse clustering of high-dimensional non-spherical Gaussian mixtures. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, 37–45.

Balakrishnan, S., Wainwright, M. J. and Yu, B. (2017) Statistical guarantees for the EM algorithm: from population to sample-based analysis. *Ann. Statist.*, **45**, 77–120.

Bingham, E. and Mannila, H. (2001) Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 245–250.

Birgé, L. (2001) An alternative point of view on Lepski's method. In de Gunst, M., Klaassen, C. and van der Vaart, A. Eds,. *Lecture Notes-Monograph Series 36*, 113–133.

Boucheron, S., Lugosi, G. and Massart, P. (2013) *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E. and Satija, R. (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotech.*, **36**, 411–420.

Cai, T. T. and Liu, W. (2011) A direct estimation approach to sparse linear discriminant analysis. *J. Amer. Statist. Assoc.*, **106**, 1566–1577.

Cai, T. T. and Zhang, L. (2019) High dimensional linear discriminant analysis: optimality, adaptive algorithm and missing data. *J. Roy. Statist. Soc., Ser. B*, **81**, 675–705.

Cannings, T. I. (2021) Random projections: Data perturbation for classification problems. *Wiley Interdisciplinary Reviews: Computational Statistics*, **13**, e1499.

Cannings, T. I., Berrett, T. B. and Samworth, R. J. (2020) Local nearest neighbour classification with applications to semi-supervised learning. *Ann. Statist.*, **48**, 1789–1814.

Cannings, T. I. and Samworth, R. J. (2017) Random-projection ensemble classification. *J. Roy. Statist. Soc., Ser. B*, **79**, 959–1035.

Chakrabortty, A. and Cai, T. (2018) Efficient and adaptive linear regression in semi-supervised settings. *Ann. Statist.*, **46**, 1541–1572.

Chapelle, O., Schölkopf, B. and Zien, A. (Eds.) (2006) *Semi-Supervised Learning*. The MIT Press.

Cheplygina, V., de Bruijne, M. and Pluim, J. P. (2019) Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, **54**, 280–296.

Cochran, W. G. (1934) The distribution of quadratic forms in a normal system, with applications to the analysis of covariance. *Mathematical Proceedings of the Cambridge Philosophical Society*, **30**, 178–191.

Dasgupta, S. (1999) Learning mixtures of Gaussians. In *The 40th Annual Symposium on Foundations of Computer Science*, 634–644.

Dasgupta, S. and Gupta, A. (2003) An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, **22**, 60–65.

Daskalakis, C., Tzamos, C. and Zampetakis, M. (2017) Ten steps of EM suffice for mixtures of two Gaussians. In *Conference on Learning Theory*, 704–710, PMLR.

Davis, D., Diaz, M. and Wang, K. (2021) Clustering a mixture of Gaussians with unknown covariance. *arXiv preprint arXiv:2110.01602*.

de Souto, M. C., Costa, I. G., de Araujo, D. S., Ludermir, T. B. and Schliep, A. (2008) Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics*, **9**, 497.

Devroye, L., Györfi, L. and Lugosi, G. (2013) *A Probabilistic Theory of Pattern Recognition*, vol. 31. Springer Science & Business Media.

Dimitriadou, E., Weingessel, A. and Hornik, K. (2002) A combination scheme for fuzzy clustering. *International Journal of Pattern Recognition and Artificial Intelligence*, **16**, 901–912.

Dobriban, E. and Liu, S. (2019) Asymptotics for sketching in least squares regression. In *Adv. Neur. Inf. Proc. Syst.*, 3675–3685.

Doss, N., Wu, Y., Yang, P. and Zhou, H. H. (2023) Optimal estimation of high-dimensional Gaussian mixtures. *Ann. Statist.*, **51**, 62–95.

Dümbgen, L., Samworth, R. J. and Wellner, J. A. (2021) Bounding distributional errors via density ratios. *Bernoulli*, **27**, 818–852.

Durrant, R. J. and Kabán, A. (2015) Random projections as regularizers: learning a linear discriminant from fewer observations than dimensions. *Machine Learning*, **99**, 257–286.

Dwivedi, R., Ho, N., Khamaru, K., Wainwright, M., Jordan, M. and Yu, B. (2020a) Sharp analysis of expectation-maximization for weakly identifiable models. In *International Conference on Artificial Intelligence and Statistics*, 1866–1876, PMLR.

Dwivedi, R., Ho, N., Khamaru, K., Wainwright, M. J., Jordan, M. I. and Yu, B. (2020b) Singularity, misspecification and the convergence rate of EM. *Ann. Statist.*, **48**, 3161–3182.

Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, **95**, 14863–14868.

Fern, X. Z. and Brodley, C. E. (2003) Random projection for high dimensional data clustering: A cluster ensemble approach. In *Proceedings of the 20th International Conference on Machine Learning*, 186–193.

Fraley, C. and Raftery, A. (1998) MCLUST: Software for model-based cluster and discriminant analysis. *Department of Statistics, University of Washington: Technical Report*, **342**, 1312.

Gataric, M., Wang, T. and Samworth, R. J. (2020) Sparse principal component analysis via axis-aligned random projections. *J. Roy. Statist. Soc., Ser. B*, **82**, 329–359.

Han, S. and Boutin, M. (2015) The hidden structure of image datasets. In *2015 IEEE International Conference on Image Processing (ICIP)*, 1095–1099, IEEE.

Hastie, T., Tibshirani, R. and Friedman, J. H. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, vol. 2. Springer.

Ho, N., Khamaru, K., Dwivedi, R., Wainwright, M. J., Jordan, M. I. and Yu, B. (2020) Instability, computational efficiency and statistical accuracy. *arXiv preprint arXiv:2005.11411*.

Ingster, Y. and Suslina, I. A. (2012) *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*, vol. 169. Springer Science & Business Media.

Jain, A. K. and Flynn, P. J. (1996) Image segmentation using clustering. In *Advances in image understanding: A Festschrift for Azriel Rosenfeld*, 65–83, IEEE Press, Piscataway, NJ.

Jin, J. and Wang, W. (2016) Influential features PCA for high dimensional clustering. *Ann. Statist.*, **44**, 2323–2359.

Johnson, W. B. and Lindenstrauss, J. (1984) Extensions of Lipschitz maps into a Hilbert space. *Contemp. Math.*, **26**, 189–206.

Kaufman, L. and Rousseeuw, P. J. (2009) *Finding Groups in Data: an Introduction to Cluster Analysis*, vol. 344. John Wiley & Sons.

Laurent, B. and Massart, P. (2000) Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, **28**, 1302–1338.

Liang, P. (2005) *Semi-supervised learning for natural language*. Ph.D. thesis, Massachusetts Institute of Technology.

Liu, H., Gao, C. and Samworth, R. J. (2021) Minimax rates in sparse, high-dimensional change point detection. *Ann. Statist.*, **49**, 1081–1112.

Lloyd, S. (1982) Least squares quantization in PCM. *IEEE Transactions on Information Theory*, **28**, 129–137.

Löffler, M., Wein, A. S. and Bandeira, A. S. (2022) Computationally efficient sparse clustering. *Information and Inference: A Journal of the IMA*, **11**, 1255–1286.

Löffler, M., Zhang, A. Y. and Zhou, H. H. (2021) Optimality of spectral clustering in the Gaussian mixture model. *Ann. Statist.*, **49**, 2506–2530.

Lopes, M., Jacob, L. and Wainwright, M. J. (2011) A more powerful two-sample test in high dimensions using random projection. In *Advances in Neural Information Processing Systems*, 1206–1214.

Mai, Q., Zou, H. and Yuan, M. (2012) A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, **99**, 29–42.

Marzetta, T. L., Tucci, G. H. and Simon, S. H. (2011) A random matrix-theoretic approach to handling singular covariance estimates. *IEEE Transactions on Information Theory*, **57**, 6256–6271.

Muirhead, R. J. (2009) *Aspects of Multivariate Statistical Theory*. John Wiley & Sons.

Ndaoud, M. (2022) Sharp optimal recovery in the two component Gaussian mixture model. *Ann. Statist.*, **50**, 2096–2126.

Oymak, S. and Gulcu, T. C. (2020) Statistical and algorithmic insights for semi-supervised learning with self-training. *Preprint*, arxiv:2006.11006.

Ramey, J. A. (2016) `Datamicroarray`: Collection of data sets for classification. *R Package*, https://rdrr.io/github/ramhiser/datamicroarray/.

Reeve, H. W., Kabán, A. and Bootkrajang, J. (2022) Heterogeneous sets in dimensionality reduction and ensemble learning. *Machine Learning*, 1–22.

Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. d. F. and Rodrigues, F. A. (2019) Clustering algorithms: a comparative approach. *PLoS ONE*, **14**, e0210236.

Slawski, M. (2018) On principal components regression, random projections, and column subsampling. *Electron. J. Statist.*, **12**, 3673–3712.

Stewart, G. W. and Sun, J. (1990) *Matrix Perturbation Theory*. Academic Press, Inc., San Diego, CA.

Thanei, G.-A., Heinze, C. and Meinshausen, N. (2017) Random projections for large-scale regression. In *Big and Complex Data Analysis*, 51–68, Springer.

Turian, J., Ratinov, L. and Bengio, Y. (2010) Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 384–394.

Van Engelen, J. E. and Hoos, H. H. (2020) A survey on semi-supervised learning. *Machine Learning*, **109**, 373–440.

Vershynin, R. (2012) How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, **25**, 655–686.

Verzelen, N. and Arias-Castro, E. (2017) Detection and feature selection in sparse mixture models. *Ann. Statist.*, **45**, 1920–1950.

von Luxburg, U. (2007) A tutorial on spectral clustering. *Statistics and Computing*, **17**, 395–416.

Wang, D., Lin, J., Cui, P., Jia, Q., Wang, Z., Fang, Y., Yu, Q., Zhou, J., Yang, S. and Qi, Y. (2019) A semi-supervised graph attentive network for financial fraud detection. In *2019 IEEE International Conference on Data Mining (ICDM)*, 598–607, IEEE.

Wasserman, L., Azizyan, M. and Singh, A. (2014) Feature selection for high-dimensional clustering. *Preprint*, arxiv:1406.2240.

Weyl, H. (1912) Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf der Theorie der Hohlraumstrahlung). *Math. Ann.*, **71**, 441–479.

Witten, D. M. and Tibshirani, R. (2010) A framework for feature selection in clustering. *J. Amer. Statist. Assoc.*, **105**, 713–726.

Witten, D. M. and Tibshirani, R. (2011) Penalized classification using Fisher's linear discriminant. *J. Roy. Statist. Soc., Ser. B*, **73**, 753–772.

Wu, Y. and Zhou, H. H. (2022) Randomly initialised EM algorithm for two-component Gaussian mixture achieves near optimality in $O(\sqrt{n})$ iterations. *Mathematical Statistics and Learning*, **4**, 143–220.

Xu, D. and Tian, Y. (2015) A comprehensive survey of clustering algorithms. *Annals of Data Science*, **2**, 165–193.

Xu, J., Hsu, D. J. and Maleki, A. (2016) Global analysis of Expectation Maximization for mixtures of two Gaussians. *Adv. Neur. Inf. Proc. Syst.*, **29**.

Xu, R. and Wunsch, D. (2005) Survey of clustering algorithms. *IEEE Transactions on neural networks*, **16**, 645–678.

Yan, B., Yin, M. and Sarkar, P. (2017) Convergence of gradient EM on multi-component mixture of Gaussians. In *Adv. Neur. Inf. Proc. Syst.*, 6956–6966.

Yang, F., Liu, S., Dobriban, E. and Woodruff, D. P. (2021) How to reduce dimension with PCA and random projections? *IEEE Transactions on Information Theory*, **67**, 8154–8189.

Yellamraju, T. and Boutin, M. (2018) Clusterability and clustering of images and other "real" high-dimensional data. *IEEE Transactions on Image Processing*, **27**, 1927–1938.

Zhang, A., Brown, L. D. and Cai, T. T. (2019) Semi-supervised inference: general theory and estimation of means. *Ann. Statist.*, **47**, 2538–2566.

Zhu, X. and Goldberg, A. B. (2009) Introduction to semi-supervised learning. In *Synthesis Lectures on Artificial Intelligence and Machine Learning* (Brachman, R. J. and Dietterich, T., eds.), 1–130, Morgan & Claypool Publishers.

Zhu, X. J. (2005) Semi-supervised learning literature survey. Tech. rep., University of Wisconsin-Madison Department of Computer Sciences.