Learning Mixtures of Linear Classifiers

Yuekai Sun YUEKAI@STANFORD.EDU
Institute for Computational and Mathematical Engineering, Stanford University, 475 Via Ortega, Stanford, CA 94305

Stratis Ioannidis STRATIS.IOANNIDIS@TECHNICOLOR.COM
Technicolor, 175 S San Antonio Rd, Los Altos, CA 94022

Andrea Montanari MONTANARI@STANFORD.EDU
Dept. of Electrical Engineering & Dept. of Statistics, Stanford University, 350 Serra Mall, Stanford, CA 94305

Abstract

We consider a discriminative learning (regression) problem, whereby the regression function is a convex combination of k linear classifiers. Existing approaches are based on the EM algorithm, or similar techniques, without provable guarantees. We develop a simple method based on spectral techniques and a 'mirroring' trick, that discovers the subspace spanned by the classifiers' parameter vectors. Under a probabilistic assumption on the feature vector distribution, we prove that this approach has nearly optimal statistical efficiency.

1. Introduction

Since Pearson's seminal contribution (Pearson, 1894), and most notably after the introduction of the EM algorithm (Dempster et al., 1977), mixture models and latent variable models have played a central role in statistics and machine learning, with numerous applications—see, e.g., McLachlan & Peel (2004), Bishop (1998), and Bartholomew et al. (2011). Despite their ubiquity, fitting the parameters of a mixture model remains a challenging task. The most popular methods (e.g., the EM algorithm or likelihood maximization by gradient ascent) are plagued by local optima and come with little or no guarantees. Computationally efficient algorithms with provable guarantees are an exception in this area. Even the idealized problem of learning mixtures of Gaussians has motivated a copious theoretical literature (Arora & Kannan, 2001; Moitra & Valiant, 2010).

In this paper we consider the problem of modeling a regression function as a mixture of k components. Namely, we are given labels $Y_i \in \mathbb{R}$ and feature vectors $X_i \in \mathbb{R}^d$, $i \in [n] \equiv \{1, 2, \dots, n\}$, and we seek estimates of the parameters of a mixture model

$$Y_i\big|_{X_i=x_i} \sim \sum_{\ell=1}^{k} p_\ell \, f(y_i|x_i, u_\ell)\,. \qquad (1)$$

Here k is the number of components, $(p_\ell)_{\ell\in[k]}$ are weights of the components, and u_ℓ is a vector of parameters for the ℓ-th component. Models of this type have been intensely studied in the neural network literature since the early nineties (Jordan & Jacobs, 1994; Bishop, 1998). They have also found numerous applications ranging from object recognition (Quattoni et al., 2004) to machine translation (Liang et al., 2006). These studies are largely based on learning algorithms without consistency guarantees.

Recently, Chaganty & Liang (2013) considered mixtures of linear regressions, whereby the relation between labels and feature vectors is linear within each component; i.e., $Y_i = \langle u_\ell, X_i\rangle + \text{noise}$ (here and below $\langle a, b\rangle$ denotes the standard inner product in $\mathbb{R}^m$). Equivalently, $f(y_i|x_i, u_\ell) = f_0(y_i - \langle x_i, u_\ell\rangle)$ with $f_0(\,\cdot\,)$ a density of mean zero. Building on a new approach developed by Hsu et al. (2012) and Anandkumar et al. (2012), these authors propose an algorithm for fitting mixtures of linear regressions with provable guarantees. The main idea is to regress Y_i^q, for $q \in \{1, 2, 3\}$ against the tensors X_i, $X_i \otimes X_i$, $X_i \otimes X_i \otimes X_i$. The coefficients of these regressions are tensors whose decomposition yields the parameters u_ℓ, p_ℓ.

While the work of Chaganty & Liang (2013) is a significant step forward, it leaves several open problems:

Statistical efficiency. Consider a standard scaling of the feature vectors, whereby the components $(X_{i,j})_{j\in[p]}$ are of order one. Then, the mathematical guarantees of Chaganty & Liang (2013) require a sample size $n \gg d^6$. This is substantially larger than the 'information-theoretic' optimal scaling, and is an unrealistic requirement in high-dimension (large d). As noted in (Chaganty & Liang,

Proceedings of the 31^st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

2013), this scaling is an intrinsic drawback of the tensor approach as it operates in a higher-dimensional space (tensor space) than the space in which data naturally live.

Linear regression versus classification. In virtually all applications of the mixture model (1), labels Y_i are categorical—see, e.g., Jordan & Jacobs (1994), Bishop (1998), Quattoni et al. (2004), Liang et al. (2006). In this case, the very first step of Chaganty & Liang, namely, regressing Y_i^2 on $X_i^{\otimes 2}$ and Y_i^3 on $X_i^{\otimes 3}$, breaks down. Consider—to be definite—the important case of binary labels (e.g., $Y_i \in \{0, 1\}$ or $Y_i \in \{+1, -1\}$). Then powers of the labels do not provide additional information (e.g., if $Y_i \in \{0, 1\}$, then $Y_i = Y_i^2$). Also, since Y_i is non-linearly related to u_ℓ, Y_i^2 does not depend only on $u_\ell^{\otimes 2}$.

Computational complexity. The method of Chaganty & Liang (2013) solves a regularized linear regression in d^3 dimensions and factorizes a third order tensor in d dimensions. Even under optimistic assumptions (finite convergence of iterative schemes), this requires $O(d^3 n + d^4)$ operations.

In this paper, we develop a spectral approach to learning mixtures of linear classifiers in high dimension. For the sake of simplicity, we shall focus on the case of binary labels $Y_i \in \{+1, -1\}$, but we expect our ideas to be more broadly applicable. We consider regression functions of the form $f(y_i|x_i, u_\ell) = f(y_i|\langle x_i, u_\ell\rangle)$, i.e., each component corresponds to a generalized linear model with parameter vector $u_\ell \in \mathbb{R}^d$. In a nutshell, our method constructs a symmetric matrix $\hat{Q} \in \mathbb{R}^{d\times d}$ by taking a suitable empirical average of the data. The matrix $\hat{Q}$ has the following property: $(d - k)$ of its eigenvalues are roughly degenerate. The remaining k eigenvalues correspond to eigenvectors that—approximately—span the same subspace as u_1, $\dots$, u_k. Once this space is accurately estimated, the problem dimensionality is reduced to k; as such, it is easy to come up with effective prediction methods (as a matter of fact, simple K-nearest neighbors works very well).

The resulting algorithm is *computationally efficient*, as its most expensive step is computing the eigenvector decomposition of a $d \times d$ matrix (which takes $O(d^3)$ operations). Assuming Gaussian feature vectors $X_i \in \mathbb{R}^d$, we prove that our method is also *statistically efficient*, i.e., it only requires $n \geq d$ samples to accurately reconstruct the subspace spanned by $u_1, \dots, u_k$. This is the same amount of data needed to estimate the covariance of the feature vectors X_i or a parameter vector $u_1 \in \mathbb{R}^d$ in the trivial case of a mixture with a single component, $k = 1$. It is unlikely that a significantly better efficiency can be achieved without additional structure.

The assumption of Gaussian feature vectors X_i's is admittedly restrictive. On one hand, as for the problem of learning mixtures of Gaussians (Arora & Kannan, 2001; Moitra & Valiant, 2010), we believe that useful insights can be gained by studying this simple setting. On the other, and as discussed below, our proof does not really require the distribution of the X_i's to be Gaussian, and a strictly weaker assumption is sufficient. We expect that future work will succeed in further relaxing this assumption.

1.1. Technical contribution and related work

Our approach is related to the principal Hessian directions (pHd) method proposed by Li (1992) and further developed by Cook (1998) and co-workers. PHd is an approach to dimensionality reduction and data visualization. It generalizes principal component analysis to the regression (discriminative) setting, whereby each data point consists of a feature vector $X_i \in \mathbb{R}^d$ and a label $Y_i \in \mathbb{R}$. Summarizing, the idea is to form the 'Hessian' matrix $\hat{H} = n^{-1}\sum_{i=1}^n Y_i X_i X_i^T \in \mathbb{R}^{d\times d}$. (We assume here, for ease of exposition, that the X_i's have zero mean and unit covariance.) The eigenvectors associated to eigenvalues with largest magnitude are used to identify a subspace in $\mathbb{R}^d$ onto which to project the feature vectors X_i's.

Unfortunately, the pHd approach fails in general for the mixture models of interest here, namely, mixtures of linear classifiers. For instance, it fails when each component of (1) is described by a logistic model $f(y_i = +1|z) = (1 + e^{-z})^{-1}$, when features are centered at $\mathbb{E}(X_i) = 0$; a proof can be found in the extended version of this paper (Sun et al., 2013).

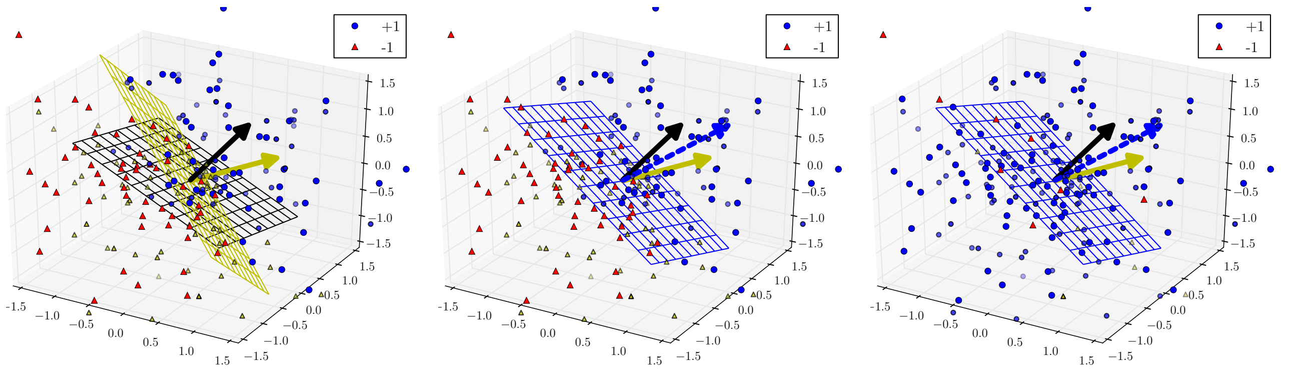
Our approach overcomes this problem by constructing $\hat{Q} = n^{-1}\sum_{i=1}^n Z_i X_i X_i^T \in \mathbb{R}^{d\times d}$. The Z_i's are pseudo-labels obtained by applying a 'mirroring' transformation to the Y_i's. Unlike with $\hat{H}$, the eigenvector structure of $\hat{Q}$ enables us to estimate the span of $u_1, \dots, u_k$.

As an additional technical contribution, we establish non-asymptotic bounds on the estimation error that allow to characterize the trade-off between the data dimension d and the sample size n. In contrast, rigorous analysis on pHd is limited to the low-dimensional regime of d fixed as $n \to \infty$. It would be interesting to generalize the analysis developed here to characterize the high-dimensional properties of pHd as well.

2. Problem Formulation

2.1. Model

Consider a dataset comprising n i.i.d. pairs $(X_i, Y_i) \in \mathbb{R}^d \times \{-1, +1\}$, $i \in [n]$. We refer to the vectors $X_i \in \mathbb{R}^d$ as *features* and to the binary variables as *labels*. We assume that the features $X_i \in \mathbb{R}^d$ are sampled from a Gaussian dis-



(a) (b) (c)

Figure 1. The mirroring process applied to a mixture of two 3-dimensional classifiers. Figure (a) shows labels generated by two classifiers in $\mathbb{R}^3$; the figure includes the parameter profiles as well as the corresponding classification surfaces. Figure (b) shows the mirroring direction $\hat{r}$ as a dashed vector, computed by (5), as well as the plane it defines; note that $\hat{r}$ lies within the positive cone spanned by the two classifier profiles, approximately. Finally, Figure (c) shows the result of the mirroring process: the region of points that was predominantly positive has remained unaltered, while the region of points that was predominantly negative has been flipped.

tribution with mean $\mu \in \mathbb{R}^d$ and a positive definite covariance $\Sigma \in \mathbb{R}^{d\times d}$. The labels $Y_i \in \{-1,+1\}$ are generated by a *mixture of linear classifiers*, i.e.,

$$\mathbf{Pr}(Y_i = +1 \mid X_i) = \textstyle\sum_{\ell=1}^{k} p_\ell\, f(\langle u_\ell, X_i\rangle)\,. \qquad (2)$$

Here, $k \geq 2$ is the number of components in the mixture; $(p_\ell)_{\ell\in[k]}$ are the weights, satisfying of course $p_\ell > 0$, $\sum_{\ell=1}^{k} p_\ell = 1$; and $(u_\ell)_{\ell\in[k]}$, $u_\ell \in \mathbb{R}^d$ are the normals to the planes defining the k linear classifiers. We refer to each normal u_ℓ as the *parameter profile* of the ℓ-th classifier; we assume that the profiles u_ℓ, $\ell \in [k]$, are linearly independent, and that $k < n/2$.

We assume that the function $f : \mathbb{R} \to [0,1]$, characterizing the classifier response, is analytic, non-decreasing, strictly concave in $[0,+\infty)$, and satisfies:

$$\lim_{t\to\infty} f(t)=1, \quad \lim_{t\to-\infty} f(t)=0, \quad 1-f(t)=f(-t). \quad (3)$$

As an example, it is useful to keep in mind the logistic function $f(t) = (1+e^{-t})^{-1}$. Fig. 1(a) illustrates a mixture of $k = 2$ classifiers over $d = 3$ dimensions.

2.2. Subspace Estimation, Prediction and Clustering

Our main focus is the following task:

> **Subspace Estimation:** After observing (X_i, Y_i), $i \in [n]$, estimate the subspace spanned by the profiles of the k classifiers, i.e., $U \equiv \text{span}(u_1, \ldots, u_k)$.

For $\widehat{U}$ an estimate of U, we characterize performance via the *principal angle* between the two spaces, namely

$$d_P(U, \widehat{U}) = \max_{x\in U, y\in\widehat{U}} \arccos\left(\tfrac{\langle x,y\rangle}{\|x\|\|y\|}\right).$$

Notice that projecting the features X_i on U entails no loss of information w.r.t. (2). This can be exploited to improve the performance of several learning tasks through dimensionality reduction, by projecting the features to the estimate of the subspace U. Two such tasks are:

> **Prediction**: Given a new feature vector X_{n+1}, predict the corresponding label Y_{n+1}.
>
> **Clustering**: Given a new feature vector and label pair (X_{n+1}, Y_{n+1}), identify the classifier that generated the label.

As we will see in Section 5, our subspace estimate can be used to significantly improve the performance of both prediction and clustering.

2.3. Technical Preliminary

We review here a few definitions used in our exposition. The *sub-gaussian norm* of a random variable X is:

$$\|X\|_{\psi_2} = \sup_{p\geq 1} \frac{1}{\sqrt{p}} (\mathbb{E}[|X|^p])^{1/p}.$$

We say X is sub-gaussian if $\|X\|_{\psi_2} < \infty$. We say that a random vector $X \in \mathbb{R}^d$ is sub-gaussian if $\langle y, X\rangle$ is sub-gaussian for any y on the unit sphere $\mathbf{S}^{d-1}$.

We use the following variant of Stein's identity (Stein, 1973; Liu, 1994). Let $X \in \mathbb{R}^d$, $X' \in \mathbb{R}^{d'}$ be jointly Gaussian random vectors, and consider a function $h : \mathbb{R}^{d'} \to \mathbb{R}$ that is almost everywhere (a.e.) differentiable and satisfies $\mathbb{E}[|\partial h(X')/\partial x_i|] < \infty$, $i \in [d']$. Then, the following identity holds:

$$\text{Cov}(X, h(X')) = \text{Cov}(X, X')\mathbb{E}[\nabla h(X')]. \qquad (4)$$

Algorithm 1 SPECTRALMIRROR

```
Input: Pairs $(X_i, Y_i)$, $i \in [n]$
Output: Subspace estimate $\widehat{U}$
1: $\hat{\mu} \leftarrow \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} X_i$
2: $\hat{\Sigma} \leftarrow \frac{1}{\lceil n/2 \rceil} \sum_{i=1}^{\lfloor n/2 \rfloor} (X_i - \hat{\mu})(X_i - \hat{\mu})^T$
3: $\hat{r} \leftarrow \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} Y_i \hat{\Sigma}^{-1}(X_i - \hat{\mu})$
4: for each $i \in \{\lfloor n/2 \rfloor + 1, \ldots, n\}$:
       $Z_i \leftarrow Y_i \,\mathrm{sgn}\langle \hat{r}, X_i \rangle$
5: $\hat{Q} \leftarrow \frac{1}{\lceil n/2 \rceil} \sum_{i=\lfloor n/2 \rfloor + 1}^{n} Z_i \hat{\Sigma}^{-1/2}(X_i - \hat{\mu})(X_i - \hat{\mu})^T \hat{\Sigma}^{-1/2}$
6: Find eigendecomposition $\sum_{\ell=1}^{d} \lambda_\ell w_\ell w_\ell^T$ of $\hat{Q}$
7: Let $\lambda_{(1)}, \ldots, \lambda_{(k)}$ be the $k$ eigenvalues furthest from
   the median.
8: $\hat{U} \leftarrow \mathrm{span}\left(\hat{\Sigma}^{-1/2} w_{(1)}, \ldots, \hat{\Sigma}^{-1/2} w_{(k)}\right)$
```

3. Subspace Estimation

In this section, we present our algorithm for subspace estimation, which we refer to as SPECTRALMIRROR. Our main technical contribution, stated formally below, is that the output $\hat{U}$ of SPECTRALMIRROR is a consistent estimator of the subspace U as soon as $n \geq C\,d$, for a sufficiently large constant C.

3.1. Spectral Mirror Algorithm

We begin by presenting our algorithm for estimating the subspace span U. Our algorithm consists of three main steps. First, as pre-processing, we estimate the mean and covariance of the underlying features X_i. Second, using these estimates, we identify a vector $\hat{r}$ that concentrates near the convex cone spanned by the profiles $(u_\ell)_{\ell \in [k]}$. We use this vector to perform an operation we call *mirroring*: we 'flip' all labels lying in the negative halfspace determined by $\hat{r}$. Finally, we compute a weighted covariance matrix $\hat{Q}$ over all X_i, where each point's contribution is weighed by the mirrored labels: the eigenvectors of this matrix, appropriately transformed, yield the span U.

These operations are summarized in Algorithm 1. We discuss each of the main steps in more detail below:

Pre-processing. (Lines 1–2) We split the dataset into two halves. Using the first half (*i.e.*, all X_i with $1 \leq i \leq \lfloor \frac{n}{2} \rfloor$), we construct estimates $\hat{\mu} \in \mathbb{R}^d$ and $\hat{\Sigma} \in \mathbb{R}^{d \times d}$ of the feature mean and covariance, respectively. Standard Gaussian (i.e., 'whitened') versions of features X_i can be constructed as $\hat{\Sigma}^{-1/2}(X_i - \hat{\mu})$.

Mirroring. (Lines 3–4) We compute the vector:

$$\hat{r} = \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} Y_i \hat{\Sigma}^{-1}(X_i - \hat{\mu}) \in \mathbb{R}^d. \qquad (5)$$

We refer to $\hat{r}$ as the *mirroring direction.* In Section 4, we show that $\hat{r}$ concentrates around its population ($n = \infty$) version $r \equiv \mathbb{E}[Y \Sigma^{-1}(X - \mu)]$. Crucially, r lies in the interior of the convex cone spanned by the parameter profiles, i.e., $r = \sum_{\ell=1}^{k} \alpha_\ell u_\ell$, for some positive $\alpha_\ell > 0$, $\ell \in [k]$ (see Lemma 2 and Fig. 1(b)). Using this $\hat{r}$, we 'mirror' the labels in the second part of the dataset:

$$Z_i = Y_i \,\mathrm{sgn}\langle \hat{r}, X_i \rangle, \quad \text{for } \lfloor n/2 \rfloor < i \leq n.$$

In words, Z_i equals Y_i for all i in the positive half-space defined by the mirroring direction; instead, all labels for points i in the negative half-space are flipped (i.e., $Z_i = -Y_i$). This is illustrated in Figure 1(c).

Spectral Decomposition. (Lines 5–8) The mirrored labels are used to compute a weighted covariance matrix over whitened features as follows:

$$\hat{Q} = \frac{1}{\lceil \frac{n}{2} \rceil} \sum_{i=\lfloor n/2 \rfloor + 1}^{n} Z_i \hat{\Sigma}^{-1/2}(X_i - \hat{\mu})(X_i - \hat{\mu})^T \hat{\Sigma}^{-1/2}$$

The spectrum of $\hat{Q}$ has a specific structure, that reveals the span U. In particular, as we will see in Section 4, $\hat{Q}$ converges to a matrix Q that contains an eigenvalue with multiplicity $n - k$; crucially, the eigenvectors corresponding to the remaining k eigenvalues, subject to the linear transform $\hat{\Sigma}^{-1/2}$, span the subspace U. As such, the final steps of the algorithm amount to discovering the eigenvalues that 'stand out' (i.e., are different from the eigenvalue with multiplicity $n - k$), and rotating the corresponding eigenvectors to obtain $\hat{U}$. More specifically, let $(\lambda_\ell, w_\ell)_{\ell \in [d]}$ be the eigenvalues and eigenvectors of $\hat{Q}$. Recall that $k < n/2$. The algorithm computes the median of all eigenvalues, and identifies the k eigenvalues furthest from this median; these are the 'outliers'. The corresponding k eigenvectors, multiplied by $\hat{\Sigma}^{-1/2}$, yield the subspace estimate $\hat{U}$.

The algorithm *does not require* knowledge of the classifier response function f. Also, while we assume knowledge of k, an eigenvalue/eigenvectors statistic (see, e.g., Zelnik-Manor & Perona (2004)) can be used to estimate k, as the number of 'outlier' eigenvalues.

3.2. Main Result

Our main result states that SPECTRALMIRROR is a consistent estimator of the subspace spanned by $(u_\ell)_{\ell \in [k]}$. This is true for 'most' $\mu \in \mathbb{R}^d$. Formally, we say that an event occurs for *generic* μ if adding an arbitrarily small random perturbation to μ, the event occurs with probability 1 w.r.t. this perturbation.

Theorem 1. *Denote by $\hat{U}$ the output of* SPECTRALMIRROR, *and let $P_r^\perp \equiv I - rr^T/\|r\|^2$ be the projector orthogonal to r, given by (6). Then, for generic μ, as well as for $\mu = 0$, there exists $\epsilon_0 > 0$ such that, for all $\epsilon \in [0, \epsilon_0)$,*

$$\mathbf{Pr}(d_P(P_r^\perp U, \hat{U}) > \epsilon) \le C_1 \exp(-C_2 \frac{n\epsilon^2}{d}).$$

Here C_1 is an absolute constant, and $C_2 > 0$ depends on μ, Σ, f and $(u_\ell)_{\ell \in [k]}$.

In other words, $\hat{U}$ provides an accurate estimate of $P_r^\perp U$ as soon as n is significantly larger than d. This holds for generic μ, but we also prove that it holds for the specific and important case where $\mu = 0$; in fact, it also holds for all small-enough μ. Note that this does not guarantee that $\hat{U}$ spans the direction $r \in U$; nevertheless, as shown below, the latter is accurately estimated by $\hat{r}$ (see Lemma 1) and can be added to the span, if necessary. Moreover, our experiments suggest this is rarely the case in practice, as $\hat{U}$ indeed includes the direction r (see Section 5).

4. Proof of Theorem 1

Recall that we denote by r the population ($n = \infty$) version of $\hat{r}$. Let $g(s) \equiv 2f(s) - 1$, for $s \in \mathbb{R}$, and observe that $\mathbb{E}[Y \mid X = x] = \sum_{\ell=1}^k p_\ell g(\langle u_\ell, x \rangle)$. Hence,

$$r = \mathbb{E}\left[\Sigma^{-1}(X - \mu) \cdot \left(\sum_{\ell=1}^k p_\ell g(\langle u_\ell, X \rangle)\right)\right]. \qquad (6)$$

Then, the following concentration result holds:

Lemma 1. *There exist an absolute constant $C > 0$ and c_1, c_1', c_2' that depend on $\|X\|_{\psi_2}$ such that:*

$$\mathbf{Pr}(\|\hat{r} - r\|_2 \ge \epsilon) \le C e^{-\min\left\{\frac{c_2 n\epsilon^2}{d}, \left(c_1'\sqrt{n}\epsilon - c_2'\sqrt{d}\right)^2\right\}}.$$

The proof of Lemma 1 relies on a large deviation inequality for sub-Gaussian vectors, and is provided in (Sun et al., 2013). Crucially, r lies in the interior of the convex cone spanned by the parameter profiles:

Lemma 2. $r = \sum_{\ell=1}^k \alpha_\ell u_\ell$ *for some $\alpha_\ell > 0$, $\ell \in [k]$.*

Proof. From (6),

$$r = \sum_{\ell=1}^k p_\ell \Sigma^{-1} \mathbb{E}[(X - \mu) g(\langle u_\ell, X \rangle)].$$

It thus suffices to show that $\Sigma^{-1}\mathbb{E}[(X - \mu)g(\langle u, X \rangle)] = \alpha u$, for some $\alpha > 0$. Note that $X' = \langle u, X \rangle$ is normal with mean $\mu_0 = u^T\mu$ and variance $\sigma_0^2 = u^T \Sigma u > 0$. Since f is analytic and non-decreasing, so is g; moreover, $g' \ge 0$. This, and the fact that g is non-constant, implies $\mathbb{E}[g'(X')] > 0$. On the other hand, from Stein's identity (4), $\mathbb{E}[g'(X')] = \frac{1}{\sigma_0^2}\mathbb{E}[X' g(X')] < \infty$, as g is bounded. Hence:

$$\begin{aligned}
&\Sigma^{-1}\mathbb{E}[(X - \mu)g(\langle u, X \rangle)] \\
&\overset{(4)}{=} \Sigma^{-1}\mathrm{Cov}(X, \langle u, X \rangle)\mathbb{E}[g'(X')], \text{ where } X' \sim \mathcal{N}(\mu_0, \sigma_0^2) \\
&= \Sigma^{-1} \cdot \mathbb{E}[(X - \mu)X^T u] \cdot \mathbb{E}[g'(X')] \\
&= \Sigma^{-1} \cdot \Sigma u \cdot \mathbb{E}[g'(X')] = \mathbb{E}[g'(X')] \cdot u
\end{aligned}$$

and the lemma follows. □

For r and $(\alpha_\ell)_{\ell \in [k]}$ as in Lemma 2, define

$$\begin{aligned}
z(x) &= \mathbb{E}[Y \,\mathrm{sgn}(\langle r, X \rangle) \mid X = x] \\
&= \left(\sum_{\ell=1}^k p_\ell g(\langle x, u_\ell \rangle)\right) \cdot \mathrm{sgn}\left(\sum_{\ell=1}^k \alpha_\ell \langle x, u_\ell \rangle\right).
\end{aligned}$$

Observe that $z(x)$ is the expectation of the mirrored label at a point x *presuming that the mirroring direction is exactly* r. Let $Q \in \mathbb{R}^{d \times d}$ be the matrix:

$$Q = \mathbb{E}[z(X)\Sigma^{-1/2}(X - \mu)(X - \mu)^T \Sigma^{-1/2}].$$

Then $\hat{Q}$ concentrates around Q, as stated below.

Lemma 3. *Let $\epsilon_0 \equiv \min\{\alpha_1, \ldots, \alpha_k\}\sigma_{\min}(U)$, where the $\alpha_\ell > 0$ are defined as per Lemma 2 and $\sigma_{\min}(U)$ is the smallest non-zero singular value of U. Then for $\epsilon < \min(\epsilon_0, \|r\|/2)$:*

$$\mathbf{Pr}(\|\hat{Q} - Q\|_2 > \epsilon) \le C \exp\{-F(\epsilon^2)\},$$

where $F(\epsilon) \equiv \min\left\{\frac{c_1 n\epsilon^2}{d}, \left(c_1'\sqrt{n}\epsilon - c_2'\sqrt{d}\right)^2\right\}$, C an absolute constant, and c_1, c_1', c_2' depend on μ, Σ, and $\|r\|$.

The proof of Lemma 3 is also provided in (Sun et al., 2013). We again rely on large deviation bounds for sub-gaussian random variables; nevertheless, our proof diverges from standard arguments because $\hat{r}$, rather than r, is used as a mirroring direction. Additional care is needed to ensure that (a) when $\hat{r}$ is close enough to r, its projection to U lies in the interior of the convex cone spanned by the profiles, and (b) although $\hat{r}$ may have a (vanishing) component outside the convex cone, the effect this has on $\hat{Q}$ is negligible, for n large enough.

An immediate consequence of Lemma 2 is that r reveals a direction in the span U. The following lemma states that the eigenvectors of Q, subject to a rotation, yield the remaining $k - 1$ directions:

Lemma 4. *Matrix Q has at most $k + 1$ distinct eigenvalues. One eigenvalue, termed λ_0, has multiplicity $d - k$. For generic μ, as well as for $\mu = 0$, the eigenvectors $w_1, \ldots, w_k$ corresponding to the remaining eigenvalues $\lambda_1, \ldots, \lambda_k$ are such that*

$$P_r^\perp U = \mathrm{span}(P_r^\perp \Sigma^{-1/2} w_1, \ldots, P_r^\perp \Sigma^{-1/2} w_k),$$

where $P_r^\perp$ is the projection orthogonal to r.

Proof. Note that

$$\begin{aligned} Q &= \mathbb{E}[z(X)\Sigma^{-\frac{1}{2}}(X-\mu)(X-\mu)^T\Sigma^{-\frac{1}{2}}] \\ &= \mathbb{E}[z(\Sigma^{1/2}W+\mu)WW^T], \quad \text{where } W \sim \mathcal{N}(0,I) \\ &= \mathbb{E}\big[\sum_{\ell=1}^{k} p_\ell g(\langle\Sigma^{\frac{1}{2}}W{+}\mu{,}u_\ell\rangle)\,\mathrm{sgn}(\langle\Sigma^{\frac{1}{2}}W{+}\mu{,}r\rangle)WW^T\big] \\ &= \mathbb{E}\big[\sum_{\ell=1}^{k} p_\ell g(\langle W+\tilde{\mu},\tilde{u}_\ell\rangle)\,\mathrm{sgn}(\langle W+\tilde{\mu},\tilde{r}\rangle)WW^T\big] \end{aligned}$$

for $\tilde{u}_\ell \equiv \Sigma^{\frac{1}{2}}u_\ell$, $\tilde{r} \equiv \Sigma^{\frac{1}{2}}r$, and $\tilde{\mu} \equiv \Sigma^{-\frac{1}{2}}\mu$. Hence $Q = \sum_{\ell=1}^{k} p_\ell Q_\ell$ where

$$Q_\ell = \mathbb{E}[g(\langle\tilde{u}_\ell, W+\tilde{\mu}\rangle)\,\mathrm{sgn}(\langle\tilde{r}, W+\tilde{\mu}\rangle)WW^T].$$

By a rotation invariance argument, Q_ℓ can be written as

$$Q_\ell = a_\ell I + b_\ell(\tilde{u}_\ell\tilde{r}^T + \tilde{r}\tilde{u}_\ell^T) + c_\ell\tilde{u}_\ell\tilde{u}_\ell^T + d_\ell\tilde{r}\tilde{r}^T \tag{7}$$

for some $a_\ell, b_\ell, c_\ell, d_\ell \in \mathbb{R}$. To see this, let $\tilde{Q}_\ell = [\tilde{q}_{ij}]_{i,j\in[d]}$, and suppose first that

$$\tilde{r} = [\tilde{r}_1, \tilde{r}_2, 0, \dots, 0] \text{ and } \tilde{u}_\ell = [\tilde{u}_{\ell 1}, \tilde{u}_{\ell 2}, 0, \dots, 0]. \tag{8}$$

Since W is whitened, its coordinates are independent. Thus, under (8), $\tilde{q}_{ij} = 0$ for all $i \neq j$ s.t. $i, j > 2$, and $\tilde{q}_{ii} = a_\ell$ for $i > 2$, for some a_ℓ. Thus $\tilde{Q}_\ell = a_\ell I + B$, where B is symmetric and 0 everywhere except perhaps on $B_{11}, B_{12}, B_{21}, B_{22}$ (the top left block). Since the profiles u_ℓ are linearly independent, so are $\tilde{u}_\ell$ and $\tilde{r}$, by Lemma 2. Hence, matrices $\tilde{u}_\ell\tilde{r}^T + \tilde{r}\tilde{u}_\ell^T, \tilde{u}_\ell\tilde{u}_\ell^T, \tilde{r}\tilde{r}^T$ span all such B, so (7) follows. Moreover, since W is whitened, $\tilde{Q}_\ell$ is rotation invariant and thus (7) extends beyond (8); indeed, if $\tilde{r}' = R\tilde{r}$, $\tilde{u}_\ell' = R\tilde{u}_\ell$, $\tilde{\mu}' = R\tilde{\mu}$ where R a rotation matrix (i.e. $RR^T = I$), then $Q' = RQR^T$. Hence, as (8) holds for some orthonormal basis, (7) holds for all bases.

Let $a = \sum_{\ell=1}^{k} p_\ell a_\ell$. Then

$$\begin{aligned} Q - aI = &\sum_{\ell=1}^{k} p_\ell d_\ell \tilde{r}\tilde{r}^T + \tilde{r}(\sum_{\ell=1}^{k} p_\ell b_\ell \tilde{u}_\ell)^T + \\ &+ (\sum_{\ell=1}^{k} p_\ell b_\ell \tilde{u}_\ell)\tilde{r}^T + \sum_{\ell=1}^{k} p_\ell c_\ell \tilde{u}_\ell\tilde{u}_\ell^T. \end{aligned}$$

Let $P_{\tilde{r}}^{\perp}$ be the projector orthogonal to $\tilde{r}$, i.e., $P_{\tilde{r}}^{\perp} = I - \frac{\tilde{r}\tilde{r}^T}{\|\tilde{r}\|_2^2}$. Let $v_\ell \equiv P_{\tilde{r}}^{\perp}\tilde{u}_\ell$. Lemma 2 and the linear independence of $\tilde{u}_\ell$ imply that $v_\ell \neq 0$, for all $\ell \in [k]$. Define $R \equiv P_{\tilde{r}}^{\perp}(Q - aI)P_{\tilde{r}}^{\perp} = \sum_{\ell=1}^{k} \gamma_\ell v_\ell v_\ell^T$, where $\gamma_\ell = p_\ell c_\ell$, $\ell \in [k]$. We will show below that for generic μ, as well as for $\mu = 0$, $\gamma_\ell \neq 0$ for all $\ell \in [k]$. This implies that $\mathrm{rank}(R) = k - 1$. Indeed, $R = P_{\tilde{r}}^{\perp} \sum \gamma_\ell \tilde{u}_\ell \tilde{u}_\ell^T P_{\tilde{r}}^{\perp} = P_{\tilde{r}}^{\perp} \tilde{R} P_{\tilde{r}}^{\perp}$, where $\tilde{R}$ has rank k by the linear independence of profiles. As $P_\perp$ is a projector orthogonal to a 1-dimensional space, R has rank at least $k-1$. On the other hand, $\mathrm{range}(R) \subseteq \tilde{U}$, for $\tilde{U} = \mathrm{span}(\tilde{u}_1, \dots, \tilde{u}_\ell)$, and $\tilde{r}^T R \tilde{r} = 0$ where $\tilde{r} \in \tilde{U} \setminus \{0\}$), so $\mathrm{rank}(R) = k - 1$. The latter also implies that $\mathrm{range}(R) = P_{\tilde{r}}^{\perp}\tilde{U}$, as $\mathrm{range}(R) \perp \tilde{r}$, $\mathrm{range}(R) \subseteq \tilde{U}$, and $\dim(\mathrm{range}(R)) = k - 1$.

The above imply that Q has one eigenvalue of multiplicity $n - k$, namely a. Moreover, the eigenvectors $w_1, \dots, w_k$ corresponding to the remaining eigenvalues (or, the non-zero eigenvalues of $Q - aI$) are such that

$$P_{\tilde{r}}^{\perp}\Sigma^{1/2}U = P_{\tilde{r}}^{\perp}\,\mathrm{span}(w_1, \dots, w_k).$$

The lemma thus follows by multiplying both sides of the above equality with $P_r^{\perp}\Sigma^{-1/2}$, and using the fact that $P_r^{\perp}\Sigma^{-1/2}P_{\tilde{r}}^{\perp} = P_r^{\perp}\Sigma^{-1/2}$.

It remains to show that $\gamma_\ell \neq 0$, for all $\ell \in [k]$, when μ is generic or 0. Note that

$$\begin{aligned} &c_\ell\langle\tilde{u}_\ell, v_\ell\rangle^2 \overset{(7)}{=} \langle v_\ell, (Q_\ell - a_\ell I)v_\ell\rangle = \\ &\quad \mathrm{Cov}(g(\langle\tilde{u}_\ell, W+\tilde{\mu}\rangle)\,\mathrm{sgn}(\langle\tilde{r}, W+\tilde{\mu}\rangle); \langle W, v_\ell\rangle^2) \equiv \tilde{c}_\ell \end{aligned} \tag{9}$$

It thus suffices to show that $\tilde{c}_\ell \neq 0$. Lemma 2 implies that $\tilde{u}_\ell = v_\ell + c\tilde{r}$ for some $c > 0$, hence

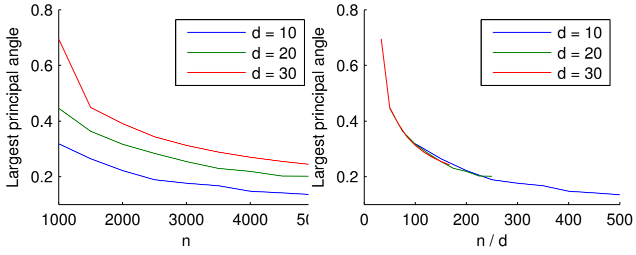
$$\tilde{c}_\ell = \mathrm{Cov}[g(X + cY + z_\ell(\mu)\rangle)\,\mathrm{sgn}(Y + z_0(\mu)); X^2],$$

where $X \equiv \langle v_\ell, W\rangle$ and $Y \equiv \langle\tilde{r}, W\rangle$ are independent Gaussians with mean 0, and $z_\ell(\mu) \equiv \langle\tilde{u}_\ell, \tilde{\mu}\rangle$, $z_0(\mu) \equiv \langle\tilde{r}, \tilde{\mu}\rangle$. Hence, $\tilde{c}_\ell = \mathrm{Cov}[F(X); X^2]$ where

$$\begin{aligned} &F(x) = \mathbb{E}_Y[g(x + cY + z_\ell(\mu))\,\mathrm{sgn}(Y + z_0(\mu))] \\ &= \int_{-z_0(\mu)}^{\infty} g(x{+}cy + z_\ell(\mu))\phi(y)dy - \int_{-\infty}^{-z_0(\mu)} g(x + cy + z_\ell(\mu)\phi(y)dy \end{aligned}$$

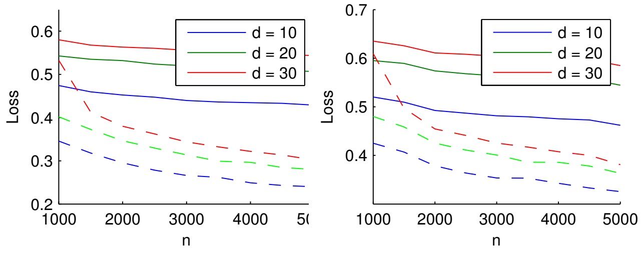
where ϕ the normal p.d.f. Assume first that $\mu = 0$. By (3), g is anti-symmetric, i.e., $g(-x) = -g(x)$. Thus, $F(-x) = \mathbb{E}_Y[g(-x + cY)\,\mathrm{sgn}(Y)] \overset{Y' \equiv -Y}{=} \mathbb{E}_{Y'}[g(-x - cY')\,\mathrm{sgn}(-Y')] = F(x)$, i.e., F is symmetric. Further, $F'(x) = \mathbb{E}_y[g'(x + cY)\,\mathrm{sgn}(Y)] = \int_0^\infty (g'(x + cy) - g'(x - cy))\phi(y)dy$. The strict concavity of g in $[0, \infty)$ implies that g' is decreasing in $[0, +\infty)$, and the anti-symmetry of g implies that g' is symmetric. Take $x > 0$: if $x > cy \geq 0$, $g'(x{+}cy) > g'(x{-}cy)$, while if $x \leq cy$, then $g'(x - cy) = g'(cy - x) > g'(cy + x)$, so $F'(x)$ is negative for $x > 0$. By the symmetry of F, $F'(x)$ is positive for $x < 0$. As such, $F(x) = G(x^2)$ for some strictly decreasing G, and $\tilde{c}_\ell = \mathrm{Cov}(G(Z); Z)$ for $Z = X^2$; hence, $\tilde{c}_\ell < 0$, for all $\ell \in [k]$.

To see that $\tilde{c}_\ell \neq 0$ for generic μ, recall that f is analytic and hence so is g. Hence, $\tilde{c}_\ell$ is an analytic function of μ, for every $\ell \in [k]$; also, as $\tilde{c}_\ell(\mu) < 0$ for $\mu = 0$, it is not identically 0. Hence, the sets $\{\mu \in \mathbb{R}^d : \tilde{c}_\ell(\mu) = 0\}$, $\ell \in [k]$, have Lebesgue measure 0 (see, e.g., pg. 83 in (Krantz & Parks, 2002)), and so does their union Z. As such, $\tilde{c}_\ell \neq 0$ for generic μ; if not, there exists a ball $B \subset \mathbb{R}^d$ such that $B \cap Z$ has positive Lebesgue measure, a contradiction. □



(a) $\sin(d_P)$ vs. n (b) $\sin(d_P)$ vs. n/d

Figure 2. Convergence of $\hat{U}$ to U.



(a) $K = \sqrt{n}$ (b) $K = \log(n)$

Figure 3. Predicting the expected label given features using K-NN (RMSE). Dotted lines are for K-NN after projecting the features X_i onto $\hat{U}$.

Denote by λ_0 the eigenvalue of multiplicity $d-k$ in Lemma 4. Let $\Delta = \min_{\ell\in[k]} |\lambda_0 - \lambda_\ell|$ be the gap between λ_0 and the remaining eigenvalues. Then, the following lemma holds; this, along with Lemma 4, yields Theorem 1.

Lemma 5. *Let $\hat{U}$ be our estimate for U. If $\lambda_1,\ldots,\lambda_k$ are separated from λ_0 by at least Δ, then for $\epsilon \le \min(\epsilon_0/\Delta, \frac{1}{4})$, we have*

$$\mathbf{Pr}(d_P(U,\hat{U}) > \epsilon) \le C \exp\big(-F(\Delta\epsilon)\big),$$

where ϵ_0, F are defined as per Lemma 3.

Proof. If we ensure $\|\hat{Q}-Q\| \le \Delta/4$, then, by Weyl's theorem (Horn & Johnson, 2012), $d-k$ eigenvalues of $\hat{Q}$ are contained in $[\lambda_{k+1}-\Delta/4, \lambda_{k+1}+\Delta/4]$, and the remaining eigenvalues are outside this set, and will be detected by SPECTRALMIRROR. Moreover, by the Davis-Kahan $\sin(\theta)$ theorem,

$$d_p(\text{range}(Q), \text{range}(\hat{Q})) \le \frac{\|\hat{Q}-Q\|_2}{\Delta - \|\hat{Q}-Q\|_2} = \frac{1}{\frac{\Delta}{\|\hat{Q}-Q\|_2}-1}.$$

Thus the event $d_p(U,\hat{U}) \le \epsilon$ is implied by $\|\hat{Q}-Q\|_2 \le \frac{\Delta\epsilon}{1+\epsilon} \le \Delta\epsilon$. Moreover, this implies that sufficient condition for $\|\hat{Q}-Q\|_2 \le \Delta/4$ (which is required for SPECTRALMIRROR to detect the correct eigenvalues) is that $\epsilon \le \frac{1}{4}$. The lemma thus follows from Lemma 3. □

Note that the Gaussianity of X is crucially used in the fact that the 'whitened' features W are uncorrelated, which in turn yields Eq. (7). We believe that the theorem can be extended to more general distributions, provided that the transform $\Sigma^{-\frac{1}{2}}$ de-correlates the coordinates of X.

5. Experiments

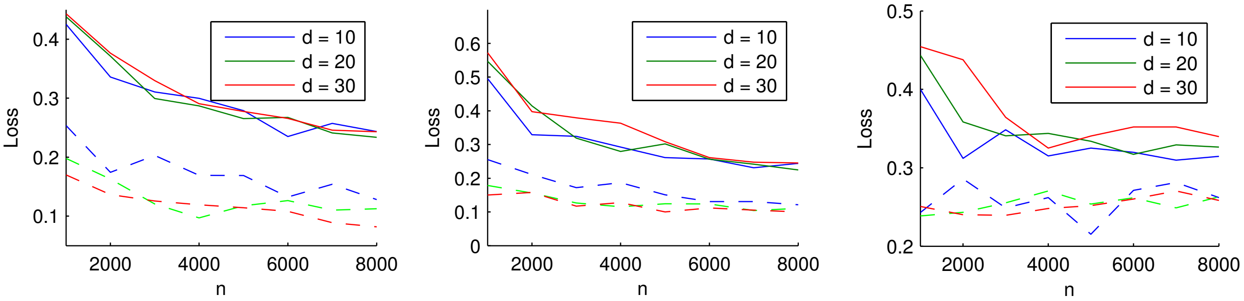
We conduct computational experiments to validate the performance of SPECRALMIRROR on subspace estimation, prediction, and clustering. We generate synthetic data using $k=2$, with profiles $u_\ell \sim \mathcal{N}(0,I)$, $\ell = 1,2$ and mixture weights p_ℓ sampled uniformly at random from the k-dimensional simplex. Features are also Gaussian: $X_i \sim \mathcal{N}(0,I)$, $i=1,\ldots,n$; labels generated by the ℓ-th classifier are given by $y_i = \text{sgn}(u_\ell^T X_i)$, $i=1,\ldots,n$.

Convergence. We study first how well SPECTRALMIRROR estimates the span U. Figure 2(a) shows the convergence of $\hat{U}$ to U in terms of (the sin of) the largest principal angle between the subspaces versus the sample size n. We also plot the convergence versus the effective sample size n/d (Figure 2(a)). The curves for different values of d align in Figure 2, indicating that the upper bound in Thm. 1 correctly predicts the sample complexity as $n \approx \Theta(d)$. Though not guaranteed by Theorem 1, in all experiments r was indeed spanned by $\hat{U}$, so the addition of $\hat{r}$ to $\hat{U}$ was not necessary.

Prediction through K-NN. Next, we use the estimated subspace to aid in the prediction of expected labels. Given a new feature vector X, we use the average label of its K nearest neighbors (K-NN) in the training set to predict its expected label. We do this for two settings: once over the raw data (the 'ambient' space), and once over data for which the features X are first projected to $\hat{U}$, the estimated span (of dimension 2). For each n, we repeat this procedure 25 times with $K=\sqrt{n}$ and $K=\log n$. We record the average root mean squared error between predicted and expected labels over the 25 runs. Figures 3(a) and 3(b) show that, despite the error in $\hat{U}$, using K-NN on this subspace outperforms K-NN on the ambient space.

Prediction and Clustering through EM. We next study the performance of prediction and clustering using the Expectation-Maximization (EM) algorithm. We use EM to fit the individual profiles both over the training set, as well as on the dataset projected to the estimated subspace $\hat{U}$. We conducted two experiments in this setting: (a) initialize EM close to the true profiles u_ℓ, $\ell \in [k]$, and (b) randomly initialize EM and choose the best set of profiles from 30 runs. For each n we run EM 10 times.

The first set of prediction experiments, we again compare expected labels to the predicted labels, using for the latter profiles u_ℓ and mixture probabilities p_ℓ as estimated by



(a) EM Prediction (close to gr. truth) (b) EM Prediction (random) (c) Clustering (random)

Figure 4. (a) Predicting the label given features and the classifier using using EM (normalized 0-1 loss) from a starting point close to ground truth. Dotted lines are for kNN after projecting the features onto the estimated subspace. (b) Predicting the label given features and the classifier using using EM (normalized 0-1 loss) from a random starting point. (c) Predicting the classifier given features and the label.

EM. Figure 4(a) measures the statistical efficiency of EM over the estimated subspace versus EM over the ambient space, when EM is initialized close to the true profiles. The second set of experiments, illustrated in Figure 4(b), aims to capture the additional improvement due to the reduction in the number of local minima in the reduced space. In both cases we see that constraining the estimated profiles to lie in the estimated subspace improves the statistical efficiency of EM; in the more realistic random start experiments, enforcing the subspace constraint also improves the performance of EM by reducing the number of local minima. We also observe an overall improvement compared to prediction through K-NN.

Finally, we use the fitted profiles u_ℓ to identify the classifier generating a label given the features and the label. To do this, once the profiles u_ℓ have been detected by EM, we use a logistic model margin condition to identify the classifier who generated a label, given the label and its features. Figure 4(c) shows the result for EM initialized at a random point, after choosing the best set of profiles from out of 30 runs. We evaluate the performance of this clustering procedure using the normalized 0-1 loss. Again, constraining the estimated profiles to the estimated subspace significantly improves the performance of this clustering task.

6. Conclusions

We have proposed SPECTRALMIRROR, a method for discovering the span of a mixture of linear classifiers. Our method relies on a non-linear transform of the labels, which we refer to as 'mirroring'. Moreover, we have provided consistency guarantees and non-asymptotic bounds, that also imply the near optimal statistical efficiency of the method. Finally, we have shown that, despite the fact that SPECTRALMIRROR discovers the span only approximately, this is sufficient to allow for a significant improvement in both prediction and clustering, when the features are projected to the estimated span.

We have already discussed several technical issues that remain open, and that we believe are amenable to further analysis. These include amending the Gaussianity assumption, and applying our bounds to other pHd-inspired methods. An additional research topic is to further improve the computational complexity of the estimation of the eigenvectors of the 'mirrored' matrix $\hat{Q}$. This is of greatest interest in cases where the covariance Σ and mean μ are a priori known. This would be the case when, e.g., the method is applied repeatedly and, although the features X are sampled from the same distribution each time, labels Y are generated from a different mixture of classifiers. In this case, SPECTRALMIRROR lacks the pre-processing step, that requires estimating Σ and is thus computationally intensive; the remaining operations amount to discovering the spectrum of $\hat{Q}$, an operation that can be performed more efficiently. For example, we can use a regularized M-estimator to exploit the fact that $\Sigma^{-1/2}\hat{Q}\Sigma^{-1/2}$ should be the sum of a multiple of the identity and a low rank matrix—see, e.g., Negahban et al. (2012).

References

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. Tensor decompositions for learning latent variable models, 2012. arXiv preprint, arXiv:1210.7559.

Arora, S. and Kannan, R. Learning mixtures of arbitrary Gaussians. In *Proceedings of the 33rd annual ACM Symposium on Theory of Computing*, pp. 247–257. ACM, 2001.

Bartholomew, D. J., Knott, M., and Moustaki, I. *Latent Variable Models and Factor Analysis: a Unified Approach*, volume 899. Wiley & Sons, 2011.

Bishop, C. M. Latent variable models. In *Learning in Graphical Models*, pp. 371–403. Springer, 1998.

Chaganty, A. T. and Liang, P. Spectral experts for estimating mixtures of linear regressions. In *ICML*, 2013.

Cook, R. D. Principal Hessian directions revisited. *Journal of the American Statistical Association*, 93(441):84–94, 1998.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

Horn, R. A. and Johnson, C. R. *Matrix Analysis*. Cambridge University Press, 2012.

Hsu, D., Kakade, S. M., and Zhang, T. A spectral algorithm for learning hidden Markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

Jordan, M. I. and Jacobs, R. A. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6 (2):181–214, 1994.

Krantz, S. G. and Parks, H. R. *A primer of real analysis and functions*. Springer, 2002.

Li, K.-C. On principal Hessian directions for data visualization and dimension reduction: another application of Stein's Lemma. *Journal of the American Statistical Association*, 87(420):1025–1039, 1992.

Liang, P., Bouchard-Côté, A., Klein, D., and Taskar, B. An end-to-end discriminative approach to machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pp. 761–768. Association for Computational Linguistics, 2006.

Liu, J. S. Siegel's formula via Stein's identities. *Statistics & Probability Letters*, 21(3):247–251, 1994.

McLachlan, G. and Peel, D. *Finite Mixture Models*. Wiley & Sons, 2004.

Moitra, A. and Valiant, G. Settling the polynomial learnability of mixtures of Gaussians. In *51st Annual IEEE Symposium on Foundations of Computer Science*, pp. 93–102. IEEE, 2010.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Pearson, Karl. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

Quattoni, A., Collins, M., and Darrell, T. Conditional random fields for object recognition. In *Advances in Neural Information Processing Systems*, pp. 1097–1104, 2004.

Stein, C. M. Estimation of the mean of a multivariate normal distribution. In *Prague Symposium on Asymptotic Statistics*, 1973.

Sun, Y., Ioannidis, S., and Montanari, A. Learning mixtures of linear classifiers, 2013. arXiv:1311.2547.

Zelnik-Manor, L. and Perona, P. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems*, 2004.