Electronic Journal of Statistics
Vol. 8 (2014) 3004–3030
ISSN: 1935-7524
DOI: 10.1214/14-EJS981

# Learning mixtures of Bernoulli templates by two-round EM with performance guarantee

Adrian Barbu, Tianfu Wu and Ying Nian Wu

**Abstract:** Dasgupta and Shulman [1] showed that a two-round variant of the EM algorithm can learn mixture of Gaussian distributions with near optimal precision with high probability if the Gaussian distributions are well separated and if the dimension is sufficiently high. In this paper, we generalize their theory to learning mixture of high-dimensional Bernoulli templates. Each template is a binary vector, and a template generates examples by randomly switching its binary components independently with a certain probability. In computer vision applications, a binary vector is a feature map of an image, where each binary component indicates whether a local feature or structure is present or absent within a certain cell of the image domain. A Bernoulli template can be considered as a statistical model for images of objects (or parts of objects) from the same category. We show that the two-round EM algorithm can learn mixture of Bernoulli templates with near optimal precision with high probability, if the Bernoulli templates are sufficiently different and if the number of features is sufficiently high. We illustrate the theoretical results by synthetic and real examples.

**Keywords and phrases:** Clustering, performance bounds, unsupervised learning.

Received December 2013.

## 1. Introduction

During the past decades, a large number of theoretical results have been obtained for supervised learning such as classification and regression [9]. For unsupervised learning, however, relatively few theoretical results are available. A main difficulty is that the objective functions in unsupervised learning are usually non-convex and multi-modal, so the optimization algorithms usually cannot find the global optima. As a result, it is generally difficult to obtain theoretical guarantees for the performances of the unsupervised learning algorithms. A simple and typical example of unsupervised learning is clustering or learning mixture models, and a typical algorithm for fitting the mixture models is the EM algorithm [3], which is a statistical counterpart of the k-means algorithm for clustering. Although the EM algorithm is simple and interpretable, and is known to converge monotonically to a local mode of the observed-data log-likelihood, little is known about its theoretical performance in terms of correctly recovering the mixture components. As such, the EM algorithm is often considered a heuristic algorithm.

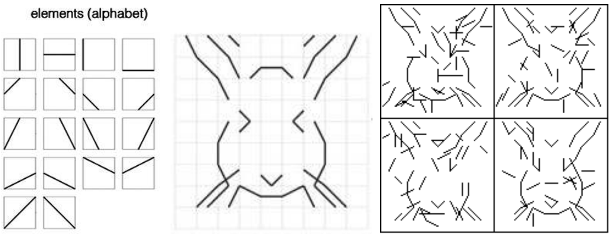

FIG 1. *Left: An alphabet of 18 sketch patterns. These sketch patterns are edge segments that connect the corners and mid-points of the sides of a squared cell. Middle: The image domain is partitioned into squared cells. Within each cell, any of the sketch patterns can be present or absent. The whole feature map can be represented by a binary vector, where each component is a binary decision on whether a certain sketch pattern in the alphabet is present or absent within a certain cell. Right: Some examples generated by the template in the middle by randomly switching the binary components with a certain probability.*

A major recent advance in the theoretical understanding of the EM algorithm for fitting mixture models was made by Dasgupta and Shulman [1]. They proposed a two-round variant of the EM algorithm that consists of only two iterations of EM: the first iteration is initialized from a number of randomly selected training examples as the centers of the Gaussian distributions, and the second iteration is carried out after pruning the clusters learned from the first iteration. They showed that the two-round EM can learn the mixture of Gaussian distributions with near optimal precision with high probability if the Gaussian distributions are well separated and if the dimensionality of the Gaussian distributions is sufficiently high. Here near optimal precision means that one can estimate the parameters of the Gaussian distributions as if the memberships of the observations are known.

In this paper, we generalize the theory of Dasgupta and Shulman [1] to learning mixture of Bernoulli templates. Each template is a binary vector, and it generates examples by independently switching its binary components with a certain probability. So the observed examples are also binary vectors. This setup is a version of the latent class model of [5] restricted to binary data. In potential applications in computer vision, a binary vector is a feature map of an image, where each binary component indicates whether a local feature or structure is present or absent within a certain cell of the image domain. Fig. 1 illustrates the basic idea by a synthetic example. The image domain is equally partitioned into squared cells (in the example in Fig. 1, there are a total of $9 \times 9 = 81$ cells in the image domain). There is an alphabet of sketch patterns that can appear in these cells (Fig. 1 shows an alphabet of 18 types of sketch patterns). Each cell may contain one or more sketch patterns, so the binary vector for each image consists of $9 \times 9 \times 18$ binary components, each component indicates whether a certain sketch pattern is present or not within a certain cell. Specifically, each component is a binary decision that can be made based on local edge detection, Gabor filter responses [2], beamlet transformation [6] or a pre-trained classifier. A Gabor filter is a 2D linear filter that has a prefered orientation. Along

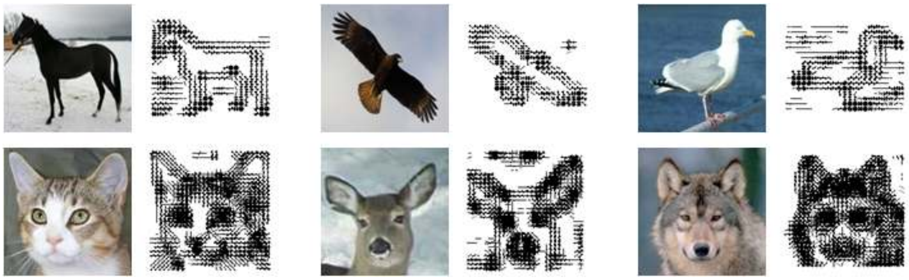

Fig 2. *Real images and their binary sketches. Each bar in the sketch image indicates the existence of a Gabor filter response above a threshold within a local cell of the image.*

that orientation the Gabor filter resembles a Gaussian and along the perpendicular direction it resembles the derivative of a Gaussian. It was shown in [2] that the Gabor filters are a good approximation of the receptive field profiles of orientation-sensitive neurons in a cat's visual cortex.

The formulation is very general. One can design any alphabet of local features or patterns, and one can use any binary detector or classifier to decide the presence or absence of these features within each cell. The whole feature map is a composition of local image features and is in the form of a binary vector, usually high dimensional (on the order of $10^3 - 10^5$). A template itself is a binary vector that is subject to component-wise switching or Bernoulli noise to account for the variations of the feature maps of individual images. The reason we focus on binary feature maps in this article is that they are easy to design and we do not need to make strong assumptions on their distributions such as Gaussianity.

As another illustration, Fig. 2 displays some examples of real images and their binary sketches based on a simple design of image features and binary decision rules. We partition the image domain into squared cells of equal size (in these images, the cells are relatively small, ranging from $5 \times 5$ pixels to $7 \times 7$ pixels). We convolve the image with Gabor filters [2] at 8 orientations. Within each cell, at each orientation, we pool a local maximum of the Gabor filter responses (in absolute values). If the local maximum is above a threshold, we then declare that there is a sketch within this cell at this orientation, and the sketch is depicted by a bar in the corresponding binary sketch image in Fig. 2. Clearly the sketch image captures a lot of information in the corresponding original image.

Now back to the issue of learning mixture models by EM. We assume that there are $k$ Bernoulli templates, and each observed example is a noisy observation of one of the $k$ template. The question we want to answer is: given a number of training examples that are noisy observations of the $k$ templates, can an EM-type algorithm reliably recover these $k$ templates with high probability? The reason we are interested in this question is that it will shed light on unsupervised learning of templates of objects (or their parts) from real images, which is a crucial task for object modeling and recognition in computer vision. Many learning methods are based on fitting mixture models by EM-type algorithms, including the Active Basis model [8]. In the language of the And-Or graph [10] for object modeling, each template is an And-node, which is a composition of a number of sketches. The mixture of $k$ templates is an Or-node, with each template being its child node. So the mixture of the templates is an Or-And structure. The theoretical results in this paper will be useful for us to understand the learning of the Or-And structure from training images.

To answer the above question, we shall generalize the theory of Dasgupta and Shulman [1] to Bernoulli distributions, and we shall show that the two-round EM algorithm can learn mixtures of Bernoulli templates with near optimal precision with high probability if the templates are sufficiently different and if the dimensions are sufficiently high.

Generalizing the theory of [1] from Gaussian mixtures to the mixtures of Bernoulli distributions is far from being straightforward. The sample space is no longer Euclidean, and some results for Gaussian distributions cannot be translated directly into those for the Bernoulli models. So we have to establish a theoretical foundation that is suitable for our purpose. For example, we will need bounds on the tails of the distribution of distances between a template $\mathbf{P}$ and the mean of $m$ binary vectors obtained by perturbing $\mathbf{P}$ by Bernoulli noise. Similar bounds for the Gaussian case are easy to obtain because the moment generating function of $\|X\|^2$ is known when $X$ is an isotropic Gaussian.

The rest of the paper is organized as follows. Section 2 describes the two-round EM algorithm and states the main theorem. Sections 3 to 4 present theoretical results that lead to the proof of the main theorem. Section 5 illustrates the theoretical results by some experiments on synthetic and real examples. Section 6 concludes with a discussion. In the text, we shall only state the theoretical results. The proofs can be found in the appendix.

## 2. Two-round EM with performance guarantee

### *2.1. Model and algorithm*

Let $\mathbf{P}$ be a template. It is an $n$-dimensional binary vector, i.e., $\mathbf{P} \in \Omega = \{0,1\}^n$. In the example in Fig. 1, $n = 9 \times 9 \times 18 = 1458$. Let $\mathbf{P}(s)$ be the $s$-th component of $\mathbf{P}$, $s = 1, \ldots, n$. An example $\mathbf{x}$ generated by $\mathbf{P}$ is a noisy version of $\mathbf{P}$, and we write $\mathbf{x} \sim \mathbf{P}$. Specifically, let $\mathbf{x}(s)$ be the $s$-th component of $\mathbf{x}$. Then $\mathbf{x}(s) = \mathbf{P}(s)$ with probability $1-q$, and $\mathbf{x}(s) = 1 - \mathbf{P}(s)$ with probability $q$, i.e., $q$ is the probability of switching a component of $\mathbf{P}$, and it defines the level of Bernoulli noise. We assume that $q \in (0, 1/2)$. We also assume that the components of $\mathbf{x}$ are independent given $\mathbf{P}$. We call $\mathbf{P}$ a Bernoulli template because it is binary and is subject to Bernoulli noise.

Let $\{\mathbf{P}_i, i = 1, \ldots, k\}$ be $k$ Bernoulli templates with mixture weights $\{w_i, i = 1, \ldots, k\}$. We assume that $k$ is given. Otherwise, $k$ can be determined by some model selection criteria such as BIC [4, 7]. Let $\mathbf{x}_1, \ldots, \mathbf{x}_m$ be $m$ noisy observations of these $k$ templates, where the noise level is $q$. The probability that $\mathbf{x}_j$ is generated by $\mathbf{P}_i$ is $w_i$, and we let $w_{min} = \min_{i=1,\ldots,k} w_i$. We define $\mu_i$ to be the expectation of the examples generated by $\mathbf{P}_i$, i.e., $\mu_i = E[\mathbf{x}_i]$ where $\mathbf{x}_i \sim \mathbf{P}_i$. Let $S_i$ be the set of examples coming from the template $\mathbf{P}_i$.

For two $n$-dimensional vectors $\mathbf{P}$ and $\mathbf{Q}$, let $D(\mathbf{P},\mathbf{Q})=\sum_{s=1}^{n}|\mathbf{P}(s)-\mathbf{Q}(s)|$ be the $\ell_1$ distance between $\mathbf{P}$ and $\mathbf{Q}$. Let $c_{ij}$ be the separation between $\mathbf{P}_i$ and $\mathbf{P}_j$, i.e., $D(\mathbf{P}_i,\mathbf{P}_j)=d_{ij}=nc_{ij}$.

**Definition 1.** The mixture is called $c$-separated if $\min_{ij} c_{ij}=c$.

We shall show that if the separation $c$ is sufficiently large, then the two-round EM algorithm will reliably recover $\{\mathbf{P}_i, i=1,\dots,k\}$.

We use the notation $\mathbf{T}_i$ to denote the estimated $\mathbf{P}_i$. In the two-round EM, the first round initializes $\{\mathbf{T}_i^{(0)}, i=1,\dots,l\}$ to be $l$ randomly selected training examples. The initial number of clusters, $l$, is greater than the true number $k$. Specifically, we let $l=\frac{4}{w_{min}}\ln\frac{2}{\delta w_{min}}$, where $\delta$ is the confidence parameter, i.e., with probability $1-\delta$, the algorithm will succeed in recovering the mixture components. According to the coupon collector problem, the $l$ examples cover all the $k$ clusters with high probability. We estimate the Bernoulli noise level $q_0$ so that $q_0(1-q_0)=\min_{ij} D(\mathbf{T}_i^{(0)},\mathbf{T}_j^{(0)})/2n$ based on the statistics of distances between examples derived in Prop. 3. Then we run one more iteration of EM.

After the first iteration, we prune the clusters by a starvation scheme. The pruning process consists of two stages. In the first stage, we remove all the templates $\{\mathbf{T}_i^{(1)}\}$ whose weights are below a threshold $1/4l$. In the second stage, we keep only $k$ templates that are far apart from each other through an inclusion process. Specifically, we start the inclusion process by randomly picking a template. Then in each subsequent step of the inclusion process, we add a template that is farthest away from the selected templates in terms of the minimum distance between the candidate template and the selected templates. We repeat this step until we get $k$ templates. We let $i=1,\dots,k$ to index the remaining $k$ templates.

After the pruning process, we run another iteration of EM. The estimated templates from this second round EM are already near optimal as we will show.

To be more precise, Algorithm 1 describes the two-round EM. In Step 9 the templates $\{\mathbf{T}_i^{(2)}\}$ are to be converted to binary by rounding to the nearest integer.

### 2.2. Notation

For the convenience of reference, the following summarizes the notation used in this paper:

- $n$ is the dimension of Bernoulli templates, which generate examples in $\Omega=\{0,1\}^n$.
- $m$ is the number of observations.
- $k$ is the true number of clusters.
- $q\in(0,1/2)$ is the level of noise
- $B=\frac{1}{2}(1-2q)\ln\frac{1}{(1-q)(4q+\sqrt{q})}>0$,
- $E=\min(\frac{1}{4},\frac{c(1-2q)^2/2}{c(1-2q)^2+2(1-q)(q+\sqrt{q})},\frac{3c(1-2q)/4-2q-4\sqrt{6ql/n}}{c(1-2q)+q+\sqrt{q}})$

**Algorithm 1 Two-round EM for Learning Bernoulli Templates**

```
Input: Examples $\mathbf{x}_1, \dots, \mathbf{x}_m \in \Omega$, $m \geq l$
Output: Templates $\mathbf{T}_i, i = 1, \dots, k$
[1] Initialize $\mathbf{T}_i^{(0)}$ as $l$ random training examples
[2] Initialize $w_i^{(0)} = 1/l$ and $q_0 \leq 1/2$ such that
        $q_0(1-q_0) = \frac{1}{2n} \min_{i,j} D(\mathbf{T}_i^{(0)}, \mathbf{T}_j^{(0)}).$
[3] E-Step: Compute for each $i = 1, \dots, l$
        $f_i(\mathbf{x}_j) = q_0^{D(\mathbf{x}_j, \mathbf{T}_i^{(0)})}(1-q_0)^{n-D(\mathbf{x}_j, \mathbf{T}_i^{(0)})}, j = 1, \dots, m,$
        $p_i^{(1)}(\mathbf{x}_j) = \frac{w_i^{(0)} f_i(\mathbf{x}_j)}{\sum_{i'} w_{i'}^{(0)} f_{i'}(\mathbf{x}_j)}, j = 1, \dots, m$
[4] M-Step: Update, for $i = 1, \dots, l$,
        $w_i^{(1)} = \sum_{j=1}^{m} p_i^{(1)}(\mathbf{x}_j)/m$
        $\mathbf{T}_i^{(1)} = \frac{1}{m w_i^{(1)}} \sum_{j=1}^{m} p_i^{(1)}(\mathbf{x}_j)\mathbf{x}_j$
[5] Pruning: Remove all $\mathbf{T}_i^{(1)}$ with $w_i^{(1)} < w_T = \frac{1}{4l}$
[6] Pruning: Keep only $k$ templates $\mathbf{T}_i^{(1)}$ far apart. Let $i = 1, \dots, k$ index the remaining $k$ templates.
[7] Initialize $w_i^{(1)} = 1/k$ and $q_1 = q_0$.
[8] E-Step: Compute, for $i = 1, \dots, k$,
        $f_i(\mathbf{x}_j) = q_1^{D(\mathbf{x}_j, \mathbf{T}_i^{(1)})}(1-q_1)^{n-D(\mathbf{x}_j, \mathbf{T}_i^{(1)})}, j = 1, \dots, m$
        $p_i^{(2)}(\mathbf{x}_j) = \frac{w_i^{(1)} f_i(\mathbf{x}_j)}{\sum_{i'} w_{i'}^{(1)} f_{i'}(\mathbf{x}_j)}, j = 1, \dots, m$
[9] M-Step: Update, for $i = 1, \dots, k$,
        $w_i^{(2)} = \sum_{j=1}^{m} p_i^{(2)}(\mathbf{x}_j)/m,$
        $\mathbf{T}_i^{(2)} = \frac{1}{m w_i^{(2)}} \sum_{j=1}^{m} p_i^{(2)}(\mathbf{x}_j)\mathbf{x}_j$
```

- $w_{min}$: the minimum of the mixture weights.
- $\mathbf{P}_i$ is the $i$-th Bernoulli template
- $S_i$ is the set of examples coming from the template $\mathbf{P}_i$.
- $D(\mathbf{P}, \mathbf{Q}) = \sum_{s=1}^{n} |\mathbf{P}(s) - \mathbf{Q}(s)|$ is the $\ell_1$ distance between $\mathbf{P} \in \Omega$ and $\mathbf{Q} \in \Omega$.
- $c_{ij}$ is the separation between the Bernoulli templates, $D(\mathbf{P}_i, \mathbf{P}_j) = d_{ij} = nc_{ij}$
- $c = \min_{i,j} c_{ij}$

- $l$ is the initial number of mixture components $l = \frac{4}{w_{min}} \ln \frac{2}{\delta w_{min}}$. The parameter $\delta$ is the confidence level in Theorem 1.
- $w_T = \frac{1}{4l}$ is the threshold for pruning the clusters learned by the first round.
- $C_i$ collects the templates that are initialized from examples in the $i$-th cluster $S_i$ and survive the pruning process after the first round of EM, i.e.

$$C_i = \{\mathbf{T}_{i'}^{(1)}, \mathbf{T}_{i'}^{(0)} \in S_i, w_{i'}^{(1)} \geq w_T\}$$

### 2.3. *Main result*

**Theorem 1.** *Let $m$ examples be generated from a mixture of $k$ Bernoulli templates under Bernoulli noise of level $q$ and mixing weights $w_i \geq w_{min}$ for all $i$. Let $\epsilon, \delta \in (0,1)$. If the following conditions hold:*

1. *The initial number of clusters is*
$$l = \frac{4}{w_{min}} \ln \frac{2}{\delta w_{min}}.$$
2. *The number of examples is $m \geq \max(8l, 16 \ln n, \frac{8}{w_{min}} \ln \frac{12k}{\delta})$.*
3. *The separation is*
$$c > \max\left(\frac{4}{nB} \ln \frac{5n}{\epsilon w_{min}}, \frac{\max(3(1-2q), 2)}{3(1-2q)}\left(4q + 8\sqrt{\frac{6ql}{n}}\right), \frac{\ln \frac{16l}{\min(6nq,1)}}{nB(1-2q)}\right).$$
4. *The dimension is*
$$n > \max\left(\frac{3}{\min(c, 0.5)E^2} \ln \frac{12(m+1)^2}{\delta}, \frac{6k}{\delta}\right).$$

*Then with probability at least $1-\delta$, the estimated templates after the round 2 of EM satisfy:*

$$D(\mathbf{T}_i^{(2)}, \mathbf{P}_i) \leq D(mean(S_i), \mathbf{P}_i) + \epsilon q$$

The above theorem states that with high probability, the estimated templates from the two-round EM is nearly as accurate as if we knew the memberships of the examples.

### 2.4. *Sketch of the proof*

The proof follows the steps of the two-round EM. We show that after the initialization, with high probability, the initial templates cover all the clusters and the estimated noise level $q_0$ is close to the true noise level $q$. Then after the first round, the estimated templates are likely to be close to the true templates of the same clusters. After the pruning process, we prove that it is very likely that exactly one template is kept for each cluster. Finally after the second round, the estimated templates are proved to be near optimal.

## 3. Basic facts

We shall first establish some basic facts about the Bernoulli templates perturbed by Bernoulli noise. They are concerned with the $\ell_1$ distances among templates and their examples.

**Proposition 1.** *Let $\mathbf{P}, \mathbf{Q} \in \Omega$ be Bernoulli templates with noise level $q$. We have:*

1. *If $\mathbf{x} \sim \mathbf{P}$ then*
$$E[D(\mathbf{x}, \mathbf{P})] = nq, Var[D(\mathbf{x}, \mathbf{P})] = nq(1-q)$$
2. *If $\mathbf{x} \sim \mathbf{P}$ and $\mathbf{y} \in \Omega$ then*
$$\begin{aligned} E[D(\mathbf{x}, \mathbf{y})] &= nq + D(\mathbf{P}, \mathbf{y})(1-2q) \\ Var[D(\mathbf{x}, \mathbf{y})] &= nq(1-q) \end{aligned}$$
3. *If $\mathbf{x}, \mathbf{y} \sim \mathbf{P}$ then*
$$\begin{aligned} E[D(\mathbf{x}, \mathbf{y})] &= 2nq(1-q) \\ Var[D(\mathbf{x}, \mathbf{y})] &= 2nq(1-q)(1-2q+2q^2) \end{aligned}$$
4. *If $\mathbf{x} \sim \mathbf{P}, \mathbf{y} \sim \mathbf{Q} \neq \mathbf{P}$ then*
$$\begin{aligned} E[D(\mathbf{x}, \mathbf{y})] &= 2nq(1-q) + D(\mathbf{P}, \mathbf{Q})(1-2q)^2 \\ Var[D(\mathbf{x}, \mathbf{y})] &= 2nq(1-q)(1-2q+2q^2) \end{aligned}$$

**Proposition 2.** *Let $\mathbf{P}, \mathbf{Q} \in \Omega$ be Bernoulli templates with noise level $q$. We have:*

a) *If $\mathbf{x} \sim \mathbf{P}$ and $\lambda \geq 1$ then*
$$\mathbf{P}(D(\mathbf{x}, \mathbf{P}) > \lambda nq) \leq e^{-nq(\lambda-1)^2/3}$$

b) *If $\mathbf{x} \sim \mathbf{P}$ and $\epsilon \in (0,1)$ then*
$$\mathbf{P}(|D(\mathbf{x}, \mathbf{P}) - nq| > \epsilon n\sqrt{q}) \leq 2e^{-n\epsilon^2/3}$$

c) *If $\mathbf{x} \sim \mathbf{P}, \mathbf{y} \sim \mathbf{Q}$ and*
$$\nu(\mathbf{P}, \mathbf{Q}) = 2nq(1-q) + D(\mathbf{P}, \mathbf{Q})(1-2q)^2$$
*then for any $\epsilon \in (0,1)$*
$$\mathbf{P}(|D(\mathbf{x}, \mathbf{y}) - \nu(\mathbf{P}, \mathbf{Q})| > \epsilon\nu(\mathbf{P}, \mathbf{Q})) \leq 2e^{-\nu(\mathbf{P}, \mathbf{Q})\epsilon^2/3}$$

Prop. 2 states that the $\ell_1$ distance between an example and its template is concentrated around $nq$, while the distance between two examples from two different templates is concentrated around $\nu(\mathbf{P}, \mathbf{Q})$. This leads to the following proposition.

**Proposition 3.** *Draw $m$ samples from a $c$-separated mixture of $k$ Bernoulli templates with mixing weights at least $w_{min}$. Let $\epsilon_0 > 0$. Then with probability at least* $1-m^2e^{-2n(1-q)\epsilon_0^2/3}-m^2e^{-n\min(c,0.5)\epsilon_0^2/3}-2me^{-n\epsilon_0^2/3}-2me^{-n\min(c,0.5)\epsilon_0^2/3}-ke^{-mw_{min}/8}$

*a) For any* $\mathbf{x}, \mathbf{y} \in S_i$ *we have*

$$D(\mathbf{x}, \mathbf{y}) = 2n(1-q)(q \pm \epsilon_0\sqrt{q})$$

*b) For any* $\mathbf{x} \in S_i, \mathbf{y} \in S_j$, $i \neq j$, *we have*

$$D(\mathbf{x}, \mathbf{y}) = n(2q(1-q) + c_{ij}(1-2q)^2)(1 \pm \epsilon_0)$$

*c) For any* $\mathbf{x} \in S_i$ *we have*

$$D(\mathbf{x}, \mathbf{P}_i) = n(q \pm \epsilon_0\sqrt{q})$$
$$D(\mathbf{x}, \mathbf{P}_j) = n(q + c_{ij}(1-2q))(1 \pm \epsilon_0)$$

*d) Each* $|S_i| \geq \frac{1}{2}mw_i$.

Here we employ the notation that $a = b \pm \epsilon$ means $a \in (b-\epsilon, b+\epsilon)$.

**Lemma 1.** *Let* $Z_i = \frac{1}{m}\sum_{j=1}^m B_{ij}$ *where* $B_{ij}$ *are Bernoulli random variables with* $E[B_{ij}] = q$. *Then*

$$\mathbf{P}\left(\sum_{i=1}^n Z_i - nq > \lambda\right) < \exp\left(-\frac{m\lambda^2}{3nq}\right)$$

**Proposition 4** (Average of subsets)**.** *Draw a set $S_1$ of $m$ examples randomly from template $\mathbf{P} \in \{0,1\}^n$ with noise level $q < 1/2$. Then with probability at least $1-\delta$ for any subset of size at least $t \geq n$ there is no subset of $S_1$ of size at least $t$ whose average $\mu$ has*

$$D(\mu, \mathbf{P}) \geq nq + \sqrt{3nq\left(\ln\frac{me}{t} + \frac{1}{t}\ln\frac{1}{\delta}\right)}$$

Prop. 4 states that the sample average is unlikely to deviate too far from $\mathbf{P}$.

**Proposition 5** (Weighted averages)**.** *For any finite set of points $S \subset \{0,1\}^n$ and weights $w_\mathbf{x} \in [0,1], \mathbf{x} \in S$ there exists a subset $T \subset S$ such that*

1. $|T| = \lfloor \sum_{\mathbf{x}\in S} w_\mathbf{x} \rfloor$
2. $D(\mu_T, \mathbf{P}) \geq D(\mu_w, \mathbf{P})$ *where*

$$\mu_T = \frac{1}{|T|}\sum_{\mathbf{x}\in T}\mathbf{x} \text{ and } \mu_w = \frac{\sum_{\mathbf{x}\in S} w_\mathbf{x}\mathbf{x}}{\sum_{\mathbf{x}\in S} w_\mathbf{x}}.$$

Prop. 5 states that the weighted average can be bounded by unweighted average. This result is needed because the templates are estimated as the weighted averages in both rounds of the EM algorithm and from Prop. 4 and 5 we can bound on the distance to the template.

## 4. Key steps of the proof

In this section we state the results that hold for the estimated template parameters after each EM iteration. We assume that the following technical conditions hold

C1: $nc > \frac{1}{B(1-2q)} \ln \frac{16l}{\min(6nq,1)}$
C2: $m > \max(16 \ln n, 8l)$
C3: $c > \max(1, \frac{2}{3(1-2q)})(4q + 8\sqrt{6ql/n})$

These conditions are a subset of the conditions of Theorem 1 that don't depend on $w_{min}$ and $\delta$. They will be referred to in the proofs of the statements of this section.

We also assume that $\epsilon_0 \leq E$ where condition C3 guarantees that $E > 0$. Observe that condition $C3$ imposes an upper bound on the noise level $q$ since $c < 1$. In our experiments this upper bound was between 0.2 and 0.3.

### 4.1. Initialization

This section analyzes the initial estimates for the parameters before the first round of EM.

**Proposition 6.** *With probability at least* $1 - k(l+1)e^{-lw_{min}} - ke^{lw_{min}/4}$ *we have*

1. *For each true template* $\mathbf{P}_i$*, the number of* $\mathbf{T}_j^{(0)}$ *coming from* $\mathbf{P}_i$ *is at least 2.*
2. *For each true template* $\mathbf{P}_i$*, the number of* $\mathbf{T}_j^{(0)}$ *coming from* $\mathbf{P}_i$ *is at most* $\frac{15}{8}lw_i$
3. *The noise estimate satisfies*

$$q_0(1-q_0) = (1-q)(q \pm \epsilon_0\sqrt{q}).$$

By initializing from more templates than the actual number of clusters, there is a high probability that the estimated templates cover all the clusters.

### 4.2. First round of EM

**Proposition 7.** *Suppose* $\mathbf{T}_{i'}^{(0)} \in S_i$ *and* $\mathbf{T}_{j'}^{(0)} \in S_j$*,* $i \neq j$*. In the cases when the conclusions of Proposition 3 hold, for any* $\mathbf{x} \in S_i$ *the ratio between the probabilities* $p_i$ *and* $p_j$ *is*

$$\frac{p_{i'}^{(1)}(\mathbf{x})}{p_{j'}^{(1)}(\mathbf{x})} \geq \exp(nc_{ij}B(1-2q))$$

Prop. 7 states that the first round of EM will likely give higher weights to the templates representing the correct cluster than to a wrong cluster.

**Proposition 8.** *In the cases when the conclusions of Proposition 3 hold, any non-starved estimate $\mathbf{T}_{i'}^{(1)} \in C_i$ satisfies with probability $1-1/n$*

$$D(\mathbf{T}_{i'}^{(1)}, \mathbf{P}_i) \le nq + \sqrt{6nql}$$

So the estimated template of a cluster is very likely to be close to the true template of this cluster.

### *4.3. Pruning*

We prove that with high probability the pruning step will keep exactly one template from each cluster.

**Proposition 9.** *In the cases when Propositions 3, 6 and 8 hold, the set $C_i$ obeys the following properties:*

*a) Each $C_i$ is non-empty*
*b) There exists $\tau \in \mathbb{R}$ such that for any $\mathbf{x} \in C_i$ and $\mathbf{y}, \mathbf{z} \in C_j, j \neq i$ we have $D(\mathbf{y}, \mathbf{z}) \le \tau$ and $D(\mathbf{x}, \mathbf{y}) > \tau$.*
*c) The pruning procedure finds exactly one member of each $C_i$.*

### *4.4. Second round of EM*

We permute the obtained templates $\mathbf{T}_i^{(1)}$ so that $\mathbf{T}_i^{(1)} \in S_i$.

**Proposition 10.** *Suppose $\mathbf{T}_i^{(1)} \in S_i$ and $\mathbf{T}_j^{(1)} \in S_j$, $i \neq j$. In the cases when Propositions 3, 6 and 8 hold, for any $\mathbf{x} \in S_i$ the ratio between the probabilities $p_i$ and $p_j$ is*

$$\frac{p_i^{(2)}(\mathbf{x})}{p_j^{(2)}(\mathbf{x})} \ge \exp\left(\frac{1}{4}nc_{ij}(1-2q)\ln\frac{1}{6\sqrt{q}}\right) = \exp(nc_{ij}B/2)$$

**Theorem 2.** *Suppose that $l > k$, $w_i > w_{min}$ for all $i$ and that conditions $C1-C3$ hold. Then with probability at least $1-m^2e^{-2n(1-q)\epsilon_0^2/3}-m^2e^{-n\min(c,0.5)\epsilon_0^2/3}-2me^{-n\epsilon_0^2/3}-2me^{-n\min(c,0.5)\epsilon_0^2/3}-ke^{-mw_{min}/8}-k(l+1)e^{-lw_{min}}-ke^{lw_{min}/12}-k/n$, the estimated templates after the round 2 of EM satisfy:*

$$D(\mathbf{T}_i^{(2)}, \mathbf{P}_i) \le D(mean(S_i), \mathbf{P}_i) + \frac{5}{w_{min}}e^{-ncB/4}nq$$

We are now ready to prove Theorem 1.

*Proof of Theorem 1.* From $l = \frac{4}{w_{min}}\ln\frac{2}{\delta w_{min}}$, we get $ke^{-lw_{min}/4} = k\delta w_{min}/2 \le \delta/2$. Also

$$\begin{aligned} k(l+1)e^{-lw_{min}} &< 2kle^{-lw_{min}/12}e^{-11lw_{min}/12} \\ &\le \frac{\delta}{2}2l\frac{\delta^{11}w_{min}^{11}}{2^{11}} = lw_{min}\delta^{11}w_{min}^{10}\delta/2^{10} \end{aligned}$$

But

$$lw_{min} = 12 \ln \frac{2}{\delta w_{min}} \leq \frac{24}{\delta w_{min}}$$

so

$$k(l+1)e^{-lw_{min}} < 24\delta^{10} w_{min}^9 \delta / 2^{10} < \delta/12$$

Take $\epsilon_0 = E > 0$ (because of C3). From the dimension condition

$$n > \frac{3}{\min(c, 0.5)E^2} \ln \frac{12(m+1)^2}{\delta}$$

we get $(m+1)^2 e^{-n \min(c,0.5)\epsilon_0^2/3} \leq \delta/12$, so

$$m^2 e^{-2n(1-q)\epsilon_0^2/3} + 2me^{-n\epsilon_0^2/3} + m^2 e^{-n \min(c,0.5)\epsilon_0^2/3} + 2me^{-n \min(c,0.5)\epsilon_0^2/3}$$
$$\leq 2(m+1)^2 e^{-n \min(c,0.5)\epsilon_0^2/3} \leq \delta/6.$$

From the dimension condition $n > 6k/\delta$ we get $k/n < \delta/6$.

From the condition on the number of examples, we get $ke^{-mw_{min}/8} < \delta/12$.

From Theorem 2, putting all of the above inequalities together and taking $nc > \frac{4}{B} \ln \frac{5n}{\epsilon w_{min}}$, we obtain that Theorem 1 holds with probability at least $1 - \delta$. □

## 5. Experiments

This section illustrates the theoretical results obtained in the previous sections by a simulation study as well as experiments on synthetic image sketches and real images.

### *5.1. Simulation study*

In this section we conduct experiments showing that indeed, the true templates are found with high probability when the conditions of Theorem 1 hold.

We will work with a mixture of two templates, $\mathbf{P}_1 = \mathbf{0}$ and $\mathbf{P}_2 = (1, 1, \ldots, 1, 0, 0, \ldots, 0)$ where the number of 1's is $\lfloor cn \rfloor$, to obtain a desired separation $c \in [0, 1]$ in dimension $n$. We experiment with standard EM for 2, 10 and 20 iterations. The standard EM starts from $k$ clusters, instead of $l$ clusters followed by pruning as in the two-step EM. For the standard EM we also assumed the noise level $q$ is a known parameter. All results are obtained from 100 runs.

Figs. 3 and 4 show the domains where the two-step EM and the standard EM find the templates $\mathbf{P}_1, \mathbf{P}_2$ with 90% probability, thus $\delta = 0.1$.

In the two plots of Fig. 3, the horizontal axis is the minimum weight $w_{\min}$, and the vertical axis is the separation $c$. The domain for each algorithm is the region above and to the right of the corresponding curve. Two version of the two step EM algorithm were evaluated: the two-step EM, and 10-step version that does 9 EM steps after the pruning step. Five version of the original EM were evaluated, with 2 or 10 iterations, and 1, 5 or 10 random initializations

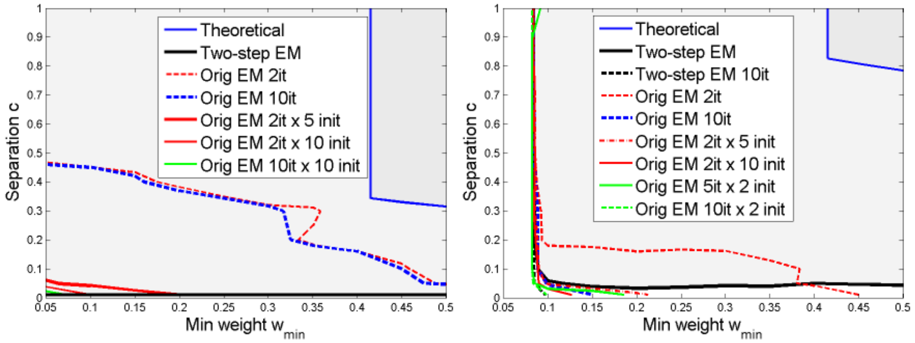

FIG 3. *Domains where the two-round EM and the standard EM find the $k = 2$ binary templates correctly 90% of the time when $m = 300$. The first plot is for $q = .01$, with $n = 2,000$, and the second plot is for $q = .1$ with $n = 10,000$. Also shown is the domain theoretically guaranteed by Theorem 1. Each domain is above and to the right of the corresponding curve.*

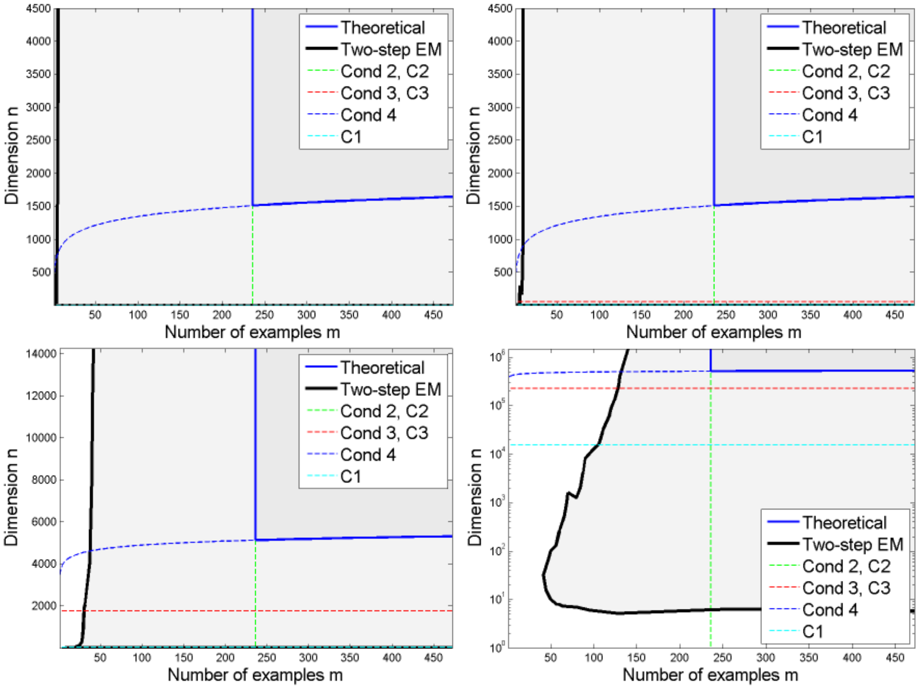

FIG 4. *Theoretical and practical domains of validity of the two-step EM algorithm for four noise levels. From left to right are noise levels: $q = 0.0001, q = 0.01$ (top) and $q = 0.1, q = 0.2$ (bottom). In these examples $c = 1, k = 2, w_{min} = 0.5, \delta = \epsilon = 0.1$. Each domain is above and to the right of the corresponding curve.*

(and selecting from the 5 or 10 obtained results the largest likelihood one as the final result). The first plot is obtained at the noise level $q = .01$, while the second plot is for the noise level $q = .1$. We take the number of observations $m = 300$. For the first plot the dimension is $n = 2,000$ and for the second plot, $n = 10,000$. One can see that for low noise, the two-step EM works better than the original EM. Also displayed is the domain where the conditions of our theorem are satisfied.

In the four plots from Fig. 4 the horizontal axis is the number $m$ of observations and the vertical axis is the dimension $n$. The four plots show the domain where the two-step EM algorithm finds the templates $\mathbf{P}_1, \mathbf{P}_2$ with 90% probability for the levels of noise $q \in \{0.0001, 0.01, 0.1, 0.2\}$. The curves corresponding to conditions 2–4 of Theorem 1 and the technical conditions C1–C3 are also displayed, as well as the domain where all conditions of our theorem are satisfied.

From the experiments we observe that the domain where the templates are found with high probability is larger than the domain where the conditions of Theorem 1 hold. The largest discrepancy is in the dimensionality conditions, where the gap between theory and experiments is considerable. This gap could be substantially decreased if tighter bounds could be obtained for Prop. 4 and consequently for Prop. 8 and Theorem 2.

### *5.2. Experiments on synthetic image sketches*

In this experiment we work with a mixture of two Bernoulli templates, shown in the bottom row of Fig. 5, in a space of dimension $n = 9 \times 9 \times 18 = 1458$. By perturbing the entries with Bernoulli noise of level $q$ we obtain images such as those shown in the top row of Fig. 5.

Fig. 6 shows the success rate of finding the two templates exactly using the two-round EM algorithm vs. the number of training examples. The experiments are run for two levels of noise $q \in \{.1, .2\}$ and two mixture weights $w_{min} \in \{.2, .4\}$.

Also shown is the bound $1 - \delta > 1 - 12ke^{-mw_{min}/8}$ from condition 2 of Theorem 1.

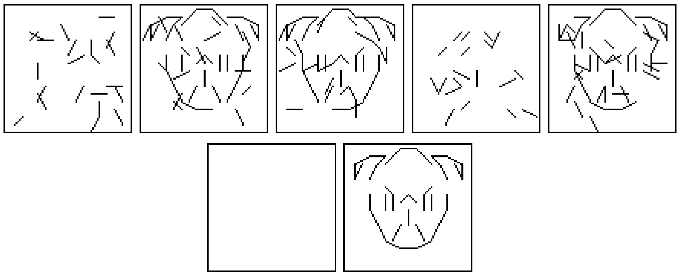

FIG 5. *Top row: Examples of training images. Bottom row: the Bernoulli templates used to generate the training images.*

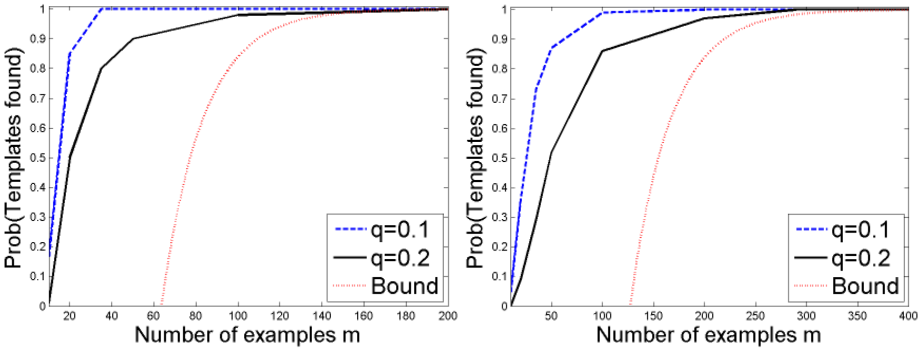

Fig 6. *Success rates vs. number of training examples for learning from a mixture of two templates with the two-round EM algorithm for two levels of noise* $q \in \{0.1, 0.2\}$ *and two mixture weights* $w_{min} = 0.4$ *(left) and* $w_{min} = 0.2$ *(right).*

The separation between the two templates is quite small $c = .02$, because the two templates share a lot of zero components. So the separation conditions fail in this case. Since we are not in the conditions of the Theorem 1, the bound on the training examples is not expected to hold. We may achieve a better bound if we reduce the dimension $n$ while increasing $c$ by selecting those features that differentiate the templates. In any case, we see that in the given scenarios the two templates can be recovered with 100% certainty with the two-round EM given sufficiently many examples. So Theorem 1 might hold under milder assumptions than ours.

### *5.3. Experiments on real images*

We also performed experiments on real images. Each image is first convolved with Gabor filters tuned to 16 orientations. Then the image domain is partitioned into equal sized squared cells (the size ranges from $5 \times 5$ pixels to $7 \times 7$ pixels). Within each cell, at each orientation, we pool the maximum of the absolute values of the filter responses. If the maximum is above a threshold, we declare that there is a sketch within this cell at this orientation. Thus each cell produces a binary vector of 16 components. We then concatenate the binary vectors of all the cells into a large binary vector. So each image is transformed into a binary vector.

*Evaluation metrics.* To evaluate the clustering quality, we introduce two metrics: conditional purity and conditional entropy. Given the underlying ground-truth category labels $X$ (which are unknown to the algorithm) and the obtained cluster labels $Y$, the conditional purity is defined as the mean of the maximum category probabilities for $(X, Y)$,

$$\text{Purity}(X|Y) = \sum_{y \in Y} p(y) \max_{x \in X} p(x|y)$$

TABLE 1

*Comparison of the two-step EM algorithm and the original EM for clustering motorcycles, bicycles and cars. Shown are the mean±std of the conditional purity, conditional entropy and log-likelihood. #Round is the number of steps of an EM algorithm, and #Init the number of random initializations used to select the best result (in terms of the log-likelihood). $N$ is the number of runs (out of 100 total runs) that recover the clusters perfectly*

| Method | #Round<br>(×#Init) | Cond. Purity | Cond. Entropy | Log-Likelihood | $N$ |
|---|---|---|---|---|---|
| Tow-round EM | 2 (×1) | 0.9402 ± 0.1124 | 0.1098 ± 0.1862 | -110625.6 ± 7914.1 | 61 |
| | 10 (×1) | 0.9822 ± 0.0653 | 0.0351 ± 0.0937 | -108460.3 ± 6491.8 | 76 |
| Original EM | 2 (×1) | 0.8511 ± 0.1500 | 0.2464 ± 0.2193 | -115852.1 ± 11079.3 | 27 |
| | 10 (×1) | 0.9004 ± 0.1464 | 0.1555 ± 0.2000 | -113476.2 ± 10889.9 | 43 |
| | 20 (×1) | 0.8722 ± 0.1572 | 0.1917 ± 0.2144 | -115083.1 ± 10828.8 | 40 |
| | 100 (×1) | 0.9051 ± 0.1460 | 0.1447 ± 0.2006 | -113205.1 ± 10809.3 | 51 |
| | 2 (×5) | 0.9911 ± 0.0295 | 0.0260 ± 0.0599 | -106268.2 ± 5448.4 | 77 |
| | 10 (×5) | 0.9987 ± 0.0053 | 0.0050 ± 0.0198 | -106067.0 ± 5122.5 | 94 |
| | 20 (×5) | 0.9956 ± 0.0336 | 0.0088 ± 0.0493 | -106249.3 ± 5540.3 | 94 |
| | 100 (×5) | 0.9996 ± 0.0031 | 0.0017 ± 0.0117 | -106045.7 ± 5106.5 | 98 |
| | 2 (×10) | 0.9991 ± 0.0054 | 0.0030 ± 0.0179 | -107549.1 ± 5366.8 | 97 |
| | 10 (×10) | 1.0000 ± 0.0000 | 0.0000 ± 0.0000 | -107534.8 ± 5377.5 | 100 |
| | 20 (×10) | 1.0000 ± 0.0000 | 0.0000 ± 0.0000 | -107534.8 ± 5377.5 | 100 |
| | 100 (×10) | 0.9998 ± 0.0022 | 0.0008 ± 0.0083 | -107541.0 ± 5390.3 | 99 |

and the conditional entropy is defined as,

$$\mathcal{H}(X|Y) = -\sum_{y\in Y} p(y) \sum_{x\in X} p(x|y)\log p(x|y)$$

where both $p(y)$ and $p(x|y)$ are estimated on the training set, and we would expect higher purity and lower entropy for a better clustering algorithm.

We then use the two-round EM algorithm to cluster the images and learn a binary template for each cluster. We compare with the original EM algorithm running for different numbers of iterations $(2, 10, 20, 100$ in the experiments) and starting with desired number $k$ of clusters (while the two-round EM starts with $l > k$ clusters and prunes them). A more robust EM could be obtained by starting with many random initialization and choosing the clustering result that has the largest log-likelihood. Such robust versions with different number of initializations are also evaluated in Tables 1, 2 and 3. A ten-round version of the two-round EM (with eight additional EM iterations after the pruning step) is also evaluated. The methods are evaluated in terms of conditional purity and conditional entropy. From the experiments one could see that the two-round EM algorithm can only be outperformed with a five or ten random initializations of the standard EM algorithm. All the results are obtained based on 100 runs.

Fig. 7 and 8 show the results of two experiments (vehicles and animal faces). Table 1 and 2 show the performance comparisons. Table 3 shows the performance all data combined. In the learned templates, the existence of a sketch at each cell is represented by a bar at the center of this cell and at the orientation of the sketch. In each experiment, there are 15 images in each cluster, and the two-round EM is able to separate the clusters perfectly. For the real images, the templates are denser than those in Fig. 5 because the numbers of cells are larger.

Table 2
*Performance comparison of our two-round EM algorithm and the original EM algorithm for clustering cats, wolves and deers*

| Method | #Round<br>(×#Init) | Cond. Purity | Cond. Entropy | Log-Likelihood | $N$ |
|---|---|---|---|---|---|
| Tow-round EM | 2 (×1) | 0.8918 ± 0.1268 | 0.2450 ± 0.2230 | -398033.9 ± 14285.5 | 17 |
| | 10 (×1) | 0.9222 ± 0.1169 | 0.1705 ± 0.2016 | -396428.5 ± 13583.5 | 32 |
| Original EM | 2 (×1) | 0.6853 ± 0.1472 | 0.6300 ± 0.2429 | -408941.0 ± 16116.3 | 2 |
| | 10 (×1) | 0.7342 ± 0.1386 | 0.5151 ± 0.2341 | -406125.6 ± 13973.7 | 2 |
| | 20 (×1) | 0.7489 ± 0.1484 | 0.4876 ± 0.2534 | -405384.8 ± 15374.4 | 5 |
| | 100 (×1) | 0.7493 ± 0.1426 | 0.4788 ± 0.2497 | -405643.4 ± 16290.6 | 6 |
| | 2 (×5) | 0.8587 ± 0.1039 | 0.3459 ± 0.1985 | -562506.3 ± 45380.3 | 6 |
| | 10 (×5) | 0.9276 ± 0.0881 | 0.1895 ± 0.1795 | -556137.0 ± 45076.8 | 19 |
| | 20 (×5) | 0.9027 ± 0.0985 | 0.2373 ± 0.1861 | -558410.6 ± 45627.2 | 11 |
| | 100 (×5) | 0.9287 ± 0.0835 | 0.1834 ± 0.1656 | -555971.3 ± 44003.9 | 20 |
| | 2 (×10) | 0.9242 ± 0.0731 | 0.2114 ± 0.1688 | -395855.2 ± 11914.2 | 15 |
| | 10 (×10) | 0.9618 ± 0.0423 | 0.1189 ± 0.1021 | -394102.3 ± 11968.6 | 24 |
| | 20 (×10) | 0.9578 ± 0.0443 | 0.1344 ± 0.1220 | -394252.5 ± 11701.0 | 25 |
| | 100 (×10) | 0.9651 ± 0.0459 | 0.1065 ± 0.1136 | -394137.2 ± 12224.4 | 31 |

Table 3
*Performance comparison of our two-round EM algorithm and the original EM algorithm for clustering cats, wolves, deers, motorcycles, bicycles and cars*

| Method | #Round<br>(×#Init) | Cond. Purity | Cond. Entropy | Log-Likelihood | $N$ |
|---|---|---|---|---|---|
| Tow-round EM | 2 (×1) | 0.8823 ± 0.0982 | 0.2526 ± 0.1891 | -401125.5 ± 28775.3 | 0 |
| | 10 (×1) | 0.9030 ± 0.0911 | 0.1841 ± 0.1506 | -398672.2 ± 28364.9 | 6 |
| Original EM | 2 (×1) | 0.7389 ± 0.0995 | 0.5152 ± 0.1775 | -416130.8 ± 32791.7 | 0 |
| | 10 (×1) | 0.7744 ± 0.1160 | 0.4042 ± 0.2001 | -412488.2 ± 32696.2 | 2 |
| | 20 (×1) | 0.7961 ± 0.1129 | 0.3669 ± 0.1862 | -409443.2 ± 33581.6 | 2 |
| | 100 (×1) | 0.7883 ± 0.1265 | 0.3862 ± 0.2196 | -410602.3 ± 34139.1 | 0 |
| | 2 (×5) | 0.8468 ± 0.0769 | 0.3212 ± 0.1364 | -402958.6 ± 28555.4 | 1 |
| | 10 (×5) | 0.9082 ± 0.0857 | 0.1793 ± 0.1351 | -396123.3 ± 30142.8 | 9 |
| | 20 (×5) | 0.9158 ± 0.0790 | 0.1686 ± 0.1263 | -395430.3 ± 27609.8 | 8 |
| | 100 (×5) | 0.9022 ± 0.0854 | 0.1843 ± 0.1290 | -396627.6 ± 29355.7 | 5 |
| | 2 (×10) | 0.8836 ± 0.0746 | 0.2669 ± 0.1376 | -398549.9 ± 28767.4 | 1 |
| | 10 (×10) | 0.9483 ± 0.0602 | 0.1163 ± 0.0937 | -392213.1 ± 27651.5 | 10 |
| | 20 (×10) | 0.9504 ± 0.0633 | 0.1072 ± 0.0985 | -392517.4 ± 29030.8 | 16 |
| | 100 (×10) | 0.9574 ± 0.0566 | 0.0992 ± 0.0882 | -391457.3 ± 27788.0 | 10 |

Currently we use a very simple sketch detector by thresholding the Gabor filter responses at different orientations. We will design more sophisticated features and associated detectors in future work.

## 6. Discussion

This paper obtains theoretical guarantees on the performance of a two-round EM algorithm for learning mixture of Bernoulli templates, by generalizing the theory of [1]. Unlike the theoretical results for supervised learning, results on unsupervised learning such as clustering are relatively scarce. The results obtained in this paper can be useful for understanding the behavior of EM-type algorithms for unsupervised learning.

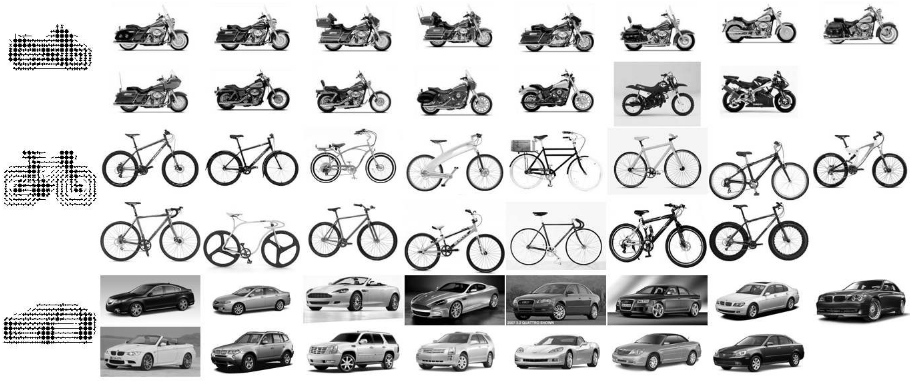

Fig 7. *Clustering motorcycles, bicycles and cars by the two-round EM algorithm. In each row, the first plot displays the learned template and the rest of the plots show some of the examples in the corresponding cluster. There are 15 images in each cluster.*

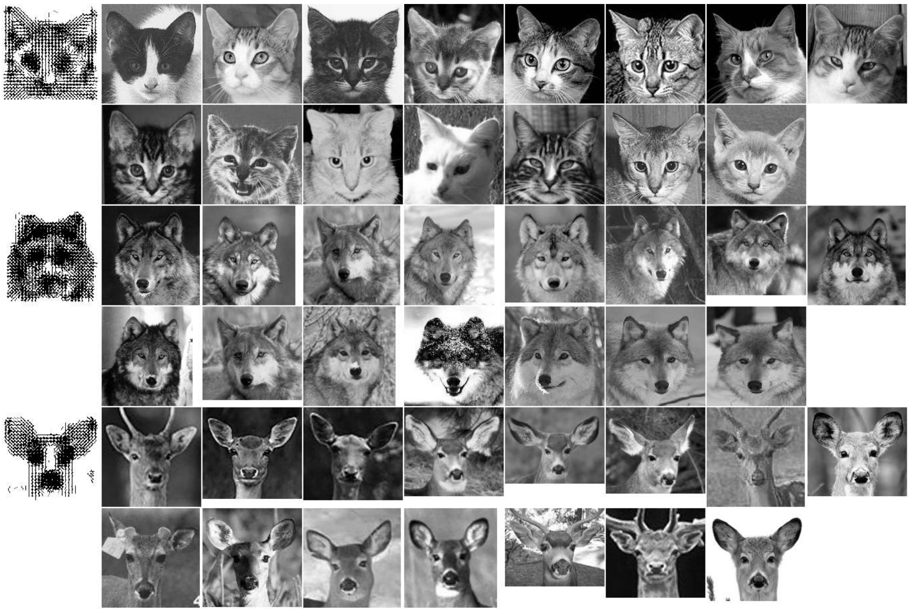

Fig 8. *Clustering cats, wolves and deers by the two-round EM algorithm. In each row, the first plot displays the learned template and the rest of the plots show some of the examples in the corresponding cluster. There are 15 images in each cluster.*

In our future work, we shall improve the theoretical results by relaxing the conditions on the separation between the templates as well as the sample size. We shall also generalize Bernoulli templates to more general statistical models for images, such as templates with dependent switching of the binary components, as well as other non-Gaussian models such as exponential family models.

## Acknowledgments

The authors wish to acknowledge support from DARPA MSEE grant FA 8650-11-1-7149 and NSF grant DMS 1007889. The authors thank Maria Pavlovskaia for her help with the experiment on synthetic sketches. Also thanks to Jianwen Xie for assistance with the animal face experiments, and to Prof. Song-Chun Zhu for valuable suggestions.

## Appendix: Proofs

*Proof of Prop. 1.* 1. We have

$$E[D(\mathbf{x},\mathbf{P})] = E\left[\sum_{k=0}^{n} B_k\right] = \sum_{k=0}^{n} E[B_k] = nq$$

and

$$E[D(\mathbf{x},\mathbf{P})^2] = E\left[\left(\sum_{k=0}^{n} B_k\right)^2\right] = E\left[\sum_{i=0}^{n} B_i^2 + \sum_{i\neq j} B_i B_j\right]$$
$$= \sum_{i=0}^{n} E[B_i] + \sum_{i\neq j} E[B_i B_j] = nq + n(n-1)q^2$$
$$Var(D(\mathbf{x},\mathbf{P})) = E[D(\mathbf{x},\mathbf{P})^2] - E[D(\mathbf{x},\mathbf{P})]^2$$
$$= n(n-1)q^2 + nq - n^2q^2 = nq(1-q)$$

2. Let $d = D(\mathbf{P},\mathbf{y})$. Without loss of generality, let $\mathbf{P} = (\mathbf{A},\mathbf{B}), \mathbf{y} = (\mathbf{A}, 1-\mathbf{B})$ where $\mathbf{B} \in \{0,1\}^d$ and $\mathbf{x} = (\mathbf{u},\mathbf{z}), \mathbf{u} \sim \mathbf{A}, \mathbf{z} \sim \mathbf{B}$. Observe that if two random variables are independent then $Var(A+B) = Var(A) + Var(B)$. Then

$$\begin{aligned} E[D(\mathbf{x},\mathbf{y})] &= E[D(\mathbf{u},\mathbf{A}) + D(\mathbf{z}, 1-\mathbf{B})] \\ &= (n-d)q + (d - E[D(\mathbf{z},\mathbf{B})]) = (n-d)q + d - dq \\ Var(D(\mathbf{x},\mathbf{y})) &= Var[D(\mathbf{u},\mathbf{A}) + d - D(\mathbf{z},\mathbf{B})] \\ &= Var[D(\mathbf{u},\mathbf{A})] + Var[d - D(\mathbf{z},\mathbf{B})] \\ &= (n-d)q(1-q) + dq(1-q) = nq(1-q) \end{aligned}$$

3. In the case when $\mathbf{x},\mathbf{y} \sim \mathbf{P}$ we have

$$E_{\mathbf{x},\mathbf{y}}[D(\mathbf{x},\mathbf{y})] = E_{\mathbf{x}}[E_{\mathbf{y}}[D(\mathbf{x},\mathbf{y})]] = E_{\mathbf{x}}[nq + D(\mathbf{x},\mathbf{P})(1-2q)]$$
$$= nq + nq(1-2q) = 2nq(1-q)$$
$$Var_{\mathbf{x},\mathbf{y}}(D(\mathbf{x},\mathbf{y})) = E_{\mathbf{x},\mathbf{y}}[D(\mathbf{x},\mathbf{y})^2] - (E_{\mathbf{x},\mathbf{y}}[D(\mathbf{x},\mathbf{y})])^2$$
$$= E_{\mathbf{x}}(E_{\mathbf{y}}[D(\mathbf{x},\mathbf{y})^2]) - E_{\mathbf{x}}(E_{\mathbf{y}}^2[D(\mathbf{x},\mathbf{y})])$$
$$+ E_{\mathbf{x}}(E_{\mathbf{y}}^2[D(\mathbf{x},\mathbf{y})]) - (E_{\mathbf{x}}[E_{\mathbf{y}}(D(\mathbf{x},\mathbf{y}))])^2$$
$$= E_{\mathbf{x}}(Var_{\mathbf{y}}[D(\mathbf{x},\mathbf{y})]) + Var_{\mathbf{x}}[E_{\mathbf{y}}(D(\mathbf{x},\mathbf{y}))]$$
$$= E_{\mathbf{x}}(nq(1-q)) + Var_{\mathbf{x}}[nq + D(\mathbf{x},\mathbf{P})(1-2q)]$$
$$= nq(1-q) + nq(1-q)(1-2q)^2$$

4. In the case when $\mathbf{x} \sim \mathbf{P}, \mathbf{y} \sim \mathbf{Q}$ we have

$$\begin{aligned}
E_{\mathbf{x},\mathbf{y}}[D(\mathbf{x},\mathbf{y})] &= E_{\mathbf{x}}[E_{\mathbf{y}}[D(\mathbf{x},\mathbf{y})]] = E_{\mathbf{x}}[nq + D(\mathbf{x},\mathbf{Q})(1-2q)] \\
&= nq + (nq + D(\mathbf{P},\mathbf{Q})(1-2q))(1-2q) \\
&= 2nq(1-q) + D(\mathbf{P},\mathbf{Q})(1-2q)^2 \\
Var_{\mathbf{x},\mathbf{y}}(D(\mathbf{x},\mathbf{y})) &= E_{\mathbf{x}}(Var_{\mathbf{y}}[D(\mathbf{x},\mathbf{y})]) + Var_{\mathbf{x}}[E_{\mathbf{y}}(D(\mathbf{x},\mathbf{y}))] \\
&= E_{\mathbf{x}}(nq(1-q)) + Var_{\mathbf{x}}[nq + D(\mathbf{x},\mathbf{Q})(1-2q)] \\
&= nq(1-q) + nq(1-q)(1-2q)^2. \qquad \square
\end{aligned}$$

*Proof of Prop. 2.* Statements a), b) follow directly from the Chernoff inequality.

c) Let $C$ be indices of the $n-d$ common elements of $\mathbf{P}$ and $\mathbf{Q}$. Let $B_i$ be the Bernoulli event that the $i$-th element of $\mathbf{x}$ and $\mathbf{y}$ are different. Then $E(B_i) = 2q(1-q)$ if $i \in C$ and $E(B_i) = q^2 + (1-q)^2$ if $i \notin C$. Observe that $D(\mathbf{x},\mathbf{y}) = \sum_{i=1}^n B_i$. Thus by the Chernoff inequality, since $\nu = E[D(\mathbf{x},\mathbf{y})] = 2nq(1-q) + d(1-2q)^2$ we get

$$\mathbf{P}(|D(\mathbf{x},\mathbf{y}) - \nu| > \epsilon\nu) \le 2e^{-\nu\epsilon^2/3}. \qquad \square$$

*Proof of Prop. 3.* a) From point c) of Prop. 2 with $\mathbf{P} = \mathbf{Q}$, we have $\nu = \nu(\mathbf{P},\mathbf{P}) = 2nq(1-q)$ so for any two points $\mathbf{x},\mathbf{y} \in S_i$ we have $\mathbf{P}(|D(\mathbf{x},\mathbf{y}) - \nu| > \nu\epsilon_0/\sqrt{q}) \le 2e^{-\nu\epsilon_0^2/3q}$. Thus for all $m(m-1)/2$ combinations of two points we have

$$\mathbf{P}(|D(\mathbf{x},\mathbf{y}) - \nu| > \nu\epsilon_0/\sqrt{q}) \le m(m-1)e^{-\nu\epsilon_0^2/3q} < m^2 e^{-2n(1-q)\epsilon_0^2/3}$$

b) Similar to the proof of a), with $\nu = \nu(\mathbf{P},\mathbf{Q}) = 2nq(1-q) + d(\mathbf{P},\mathbf{Q})(1-2q)^2 = 2nq(1-q) + nc_{ij}(1-2q)^2 \ge n\min(c, 0.5)$. We obtain

$$\mathbf{P}(|D(\mathbf{x},\mathbf{y}) - \nu| > \nu\epsilon_0) < m^2 e^{-n\min(c,0.5)\epsilon_0^2/3}$$

c) From point b) of Prop. 2 we have $\mathbf{P}(|D(\mathbf{x},\mathbf{P}_i) - nq| > \epsilon_0 n\sqrt{q}) \le 2e^{-n\epsilon_0^2/3}$ so for all $m$ points we have

$$\mathbf{P}(|D(\mathbf{x},\mathbf{P}_i) - nq| > \epsilon_0 n\sqrt{q}) \le 2me^{-n\epsilon_0^2/3}$$

Similarly, we have

$$\begin{aligned}
\mathbf{P}(|D(\mathbf{x},\mathbf{P}_j) - &n(q + c_{ij}(1-2q))| > \epsilon_0 n(q + c_{ij}(1-2q))) \\
&\le 2e^{-n(q+c_{ij}(1-2q))\epsilon_0^2/3} \le 2e^{-n\min(c,0.5)\epsilon_0^2/3}
\end{aligned}$$

so for all $m$ points we have

$$\begin{aligned}
\mathbf{P}(|D(\mathbf{x},\mathbf{P}_j) - &n(q + c_{ij}(1-2q))| > \epsilon_0 n(q + c_{ij}(1-2q))) \\
&\le 2me^{-n\min(c,0.5)\epsilon_0^2/3}
\end{aligned}$$

d) Let $B_j$ be Bernoulli event that sample $j$ is drawn from template $\mathbf{P}_i$. Then $E[B_j] = w_i$ and from the Chernoff bound

$$\mathbf{P}\left(|S_i| < \frac{1}{2}mw_i\right) = \mathbf{P}\left(\frac{\sum_{j=1}^m B_j}{m} < w_i\left(1 - \frac{1}{2}\right)\right)$$
$$< e^{-mw_i(1/2)^2/2} < e^{-mw_{min}/8}. \qquad \square$$

*Proof of Lemma 1.* The mean of $mn$ Bernoullis $B_{ij}$ with $E[B_{ij}] = q$ (the coordinates of the $Z_i$) satisfies

$$\mathbf{P}\left(\frac{\sum B_{ij}}{mn} - q > \epsilon q\right) < e^{-mnq\epsilon^2/3}$$

So

$$\mathbf{P}\left(\sum_{i=1}^n Z_i - nq > \epsilon nq\right) \leq e^{-mnq\epsilon^2/3}$$

and we take $\epsilon = \lambda/nq$. $\square$

*Proof of Prop. 4.* First, it is sufficient to prove it for subsets of size exactly $t$, otherwise we increase $t$. Without loss of generality, we can assume $\mathbf{P} = \mathbf{0}$. From Lemma 1 we have

$$\mathbf{P}(D(\mu, \mathbf{P}) - nq > \lambda) \leq e^{-t\lambda^2/3nq}$$

The number of $t$-point subsets of $S_1$ is $\binom{m}{t} < (me/t)^t$, thus

$$\mathbf{P}(\exists \text{ subset of } t \text{ points s.t. } D(\mu, \mathbf{P}) - nq > \lambda) \leq \left(\frac{me}{t}\right)^k e^{-t\lambda^2/3nq}$$

Solving for $(\frac{me}{t})^k e^{-t\lambda^2/3nq} = \delta$ we get

$$\lambda = \sqrt{\frac{3nq}{t}\left(t \ln \frac{me}{t} + \ln \frac{1}{\delta}\right)}$$

therefore

$$\mathbf{P}\left[\exists \text{ subset of } t \text{ points s.t. } D(\mu, \mathbf{P}) - nq > \sqrt{\frac{3nq}{t}\left(t \ln \frac{me}{t} + \ln \frac{1}{\delta}\right)}\right] \leq \delta. \quad \square$$

*Proof of Prop 5.* Sort the points $\mathbf{x} \in S$ by $D(\mathbf{x}, \mathbf{P}) = \sum_{i=1}^n |\mathbf{x}_i - \mathbf{P}_i|$ and take $T$ as the ones with $|T| = \lfloor \sum_{\mathbf{x} \in S} w_{\mathbf{x}} \rfloor$ largest values. Then

$$\sum_{\mathbf{x} \in T} \sum_{i=1}^n |\mathbf{x}_i - \mathbf{P}_i| \geq \sum_{\mathbf{x} \in S} w_{\mathbf{x}} \sum_{i=1}^n |\mathbf{x}_i - \mathbf{P}_i|$$

so

$$D(\mu_T, \mathbf{P}) = \sum_{i=1}^n \frac{\sum_{\mathbf{x} \in T} |\mathbf{x}_i - \mathbf{P}_i|}{|T|}$$

$$\ge \sum_{i=1}^{n} \frac{\sum_{\mathbf{x}\in S} w_{\mathbf{x}}|\mathbf{x}_i - \mathbf{P}_i|}{|T|}$$
$$\ge \sum_{i=1}^{n} \frac{\sum_{\mathbf{x}\in S} w_{\mathbf{x}}|\mathbf{x}_i - \mathbf{P}_i|}{\sum_{\mathbf{x}\in S} w_{\mathbf{x}}} = D(\mu_w, \mathbf{P}). \qquad \square$$

*Proof of Prop. 6.* Let $B_i$ be the Bernoulli event that a random sample from the mixture comes from the $i$-th true template $\mathbf{P}_i$. Then $E[B_i] = w_i$. Having $l$ random samples $B_{ij}$ from the Bernoulli event $B_i$, then

$$\mathbf{P}\left(\sum_{j=1}^{l} B_{ij} \le 1\right) = (1-w_i)^l + l w_i (1-w_i)^{l-1}$$
$$\le (1+l)(1-w_{min})^l \le (l+1)e^{-l w_{min}}$$

so $\mathbf{P}(\sum_{j=1}^{l} B_{ij} \ge 2) \ge 1-(l+1)e^{-l w_{min}}$. Thus $\mathbf{P}(\mathbf{P}_i$ is represented twice$) \ge 1-(l+1)e^{-l w_{min}}$, so $\mathbf{P}(\mathbf{P}_i$ is represented twice, $\forall i = \overline{1,k}) \ge (1-(l+1)e^{-l w_{min}})^k \ge 1 - k(l+1)e^{-l w_{min}}$.

2. From Chernoff bound we have $\mathbf{P}(\sum_{j=1}^{l} B_{ij} > 15/8 l w_i) < e^{-l w_i (7/8)^2/3} < e^{-l w_i/4}$, which implies the results.

3. As there exist $\mathbf{T}'_i, \mathbf{T}'_j$ representing the same cluster, then $2nq_0(1-q_0) \le D(\mathbf{T}'_i, \mathbf{T}'_j) \le 2n(1-q)(q+\epsilon_0\sqrt{q})$ (from Prop. 3, a). Also from Prop. 3, if the minimum is attained for two centers $\mathbf{T}'_i, \mathbf{T}'_j$ representing the same cluster, we are done. Otherwise

$$2nq_0(1-q_0) = (2nq(1-q) + nc_{ij}(1-2q)^2)(1 \pm \epsilon_0)$$
$$\ge 2nq(1-q)(1-\epsilon_0) \ge 2n(1-q)(q-\epsilon_0\sqrt{q})$$

so both parts of the inequality are proved. $\square$

*Proof of Prop 7.* We have

$$\frac{p^{(1)}_{i'}(\mathbf{x})}{p^{(1)}_{j'}(\mathbf{x})} = \frac{q_0^{D(\mathbf{x},\mathbf{T}^{(0)}_{i'})}(1-q_0)^{n-D(\mathbf{x},\mathbf{T}^{(0)}_{i'})}}{q_0^{D(\mathbf{x},\mathbf{T}^{(0)}_{j'})}(1-q_0)^{n-D(\mathbf{x},\mathbf{T}^{(0)}_{j'})}} = a^{D(\mathbf{x},\mathbf{T}^{(0)}_{j'})-D(\mathbf{x},\mathbf{T}^{(0)}_{i'})},$$

with $a = \frac{1-q_0}{q_0} > 1$. But from Prop. 3

$$\begin{aligned} D(\mathbf{x},\mathbf{T}^{(0)}_{j'}) - D(\mathbf{x},\mathbf{T}^{(0)}_{i'}) &> (2nq(1-q) + nc_{ij}(1-2q)^2)(1-\epsilon_0) \\ &\quad - 2n(1-q)(q+\epsilon_0\sqrt{q}) \\ &= -2n(q+\sqrt{q})(1-q)\epsilon_0 + nc_{ij}(1-2q)^2(1-\epsilon_0) \\ &> nc_{ij}(1-2q)^2/2 \end{aligned}$$

since we have the following condition

$$c(1-2q)^2\left(\frac{1}{2}-\epsilon_0\right) \ge 2\epsilon_0(1-q)(q+\sqrt{q})$$

obtained from $\epsilon_0 \leq E$. We also have since $\epsilon_0 < 1/4$

$$\begin{aligned} a = \frac{(1-q_0)^2}{q_0(1-q_0)} &\geq \frac{1/4}{(1-q)(q+\epsilon_0\sqrt{q})} \\ &\geq \frac{1}{4(1-q)(q+1/4\sqrt{q})} = \frac{1}{4(1-q)(4q+\sqrt{q})} \end{aligned}$$

So

$$\begin{aligned} \frac{p_{i'}^{(1)}(\mathbf{x})}{p_{j'}^{(1)}(\mathbf{x})} &\geq \exp\left(\frac{n}{2}c_{ij}(1-2q)^2 \ln \frac{1}{4(1-q)(4q+\sqrt{q})}\right) \\ &= \exp(nc_{ij}B(1-2q)). \end{aligned} \qquad \square$$

*Proof of Prop. 8.* Without loss of generality we can assume $\mathbf{P}_i = 0$.

$$\begin{aligned} D(\mathbf{T}_{i'}^{(1)}, \mathbf{P}_i) &= \frac{\sum_{k=1}^n \sum_{\mathbf{x}} p_{i'}^{(1)}(\mathbf{x})\mathbf{x}_k}{\sum_{\mathbf{x}} p_{i'}^{(1)}(\mathbf{x})} \\ &\leq \frac{\sum_{k=1}^n \sum_{\mathbf{x}\in S_i} p_{i'}^{(1)}(\mathbf{x})\mathbf{x}_k}{\sum_{\mathbf{x}} p_{i'}^{(1)}(\mathbf{x})} + \frac{\sum_{k=1}^n \sum_{\mathbf{x}\notin S_i} p_{i'}^{(1)}(\mathbf{x})\mathbf{x}_k}{\sum_{\mathbf{x}} p_{i'}^{(1)}(\mathbf{x})} \\ &\leq \frac{\sum_{k=1}^n \sum_{\mathbf{x}\in S_i} p_{i'}^{(1)}(\mathbf{x})\mathbf{x}_k}{\sum_{\mathbf{x}\in S_i} p_{i'}^{(1)}(\mathbf{x})} + \frac{\sum_{j\neq i} \sum_{\mathbf{x}\in S_j} p_{i'}^{(1)}(\mathbf{x})D(\mathbf{x},0)}{\sum_{\mathbf{x}} p_{i'}^{(1)}(\mathbf{x})} \end{aligned}$$

From Prop 7, for any $\mathbf{x} \in S_j, j \neq i$ we have $p_{i'}^{(1)}(\mathbf{x}) \leq e^{-nc_{ij}B(1-2q)} \leq e^{-ncB(1-2q)}$. Then

$$\begin{aligned} \sum_{\mathbf{x}\in S_i} p_{i'}^{(1)}(\mathbf{x}) &\geq \sum_{\mathbf{x}} p_{i'}^{(1)}(\mathbf{x}) - \sum_{j\neq i}\sum_{\mathbf{x}\in S_j} p_{i'}^{(1)}(\mathbf{x}) \\ &\geq mw_T - me^{-ncB(1-2q)} \geq mw_T/4 + 1 \end{aligned}$$

from $w_T = 1/4l$ and conditions $m \geq 8l$ (C2) and $ncB(1-2q) \geq \ln(16l)$ (C1).

From Prop. 5 there exists $T \subset S_i$ with $|T| = \lfloor mw_T/4 + 1\rfloor$ such that $D(\mu_T, 0) \geq D(\mu_w, 0)$. From Prop 4, with probability $1 - 1/n$

$$\frac{\sum_{j=1}^n \sum_{\mathbf{x}\in S_i} p_{i'}^{(1)}(\mathbf{x})\mathbf{x}_j}{\sum_{\mathbf{x}\in S_i} p_{i'}^{(1)}(\mathbf{x})} \leq D(\mu_T, 0) \leq nq + \sqrt{3nq\left(\ln\frac{4|S_i|e}{mw_T} + \frac{4}{mw_T}\ln n\right)} \tag{A.1}$$

Then since $1/w_T = 4l$ we have

$$\frac{\sum_{j=1}^n \sum_{\mathbf{x}\in S_i} p_{i'}^{(1)}(\mathbf{x})\mathbf{x}_j}{\sum_{\mathbf{x}\in S_i} p_{i'}^{(1)}(\mathbf{x})} \leq nq + \sqrt{3nq\left(\ln 16el + \frac{16l}{m}\ln n\right)} \leq nq + \sqrt{6nql} \tag{A.2}$$

from condition $m > 16\ln n$ (C2) and $\ln 16el < l$ (which holds for $l \geq 9$).

For the second term, from Prop. 3 we have, for $\mathbf{x} \in S_j$

$$D(\mathbf{x}, \mathbf{P}_i) \leq (nq + nc_{ij}(1-2q))(1+\epsilon_0)$$

where since $\epsilon_0 \leq 0.5$ we have

$$\begin{aligned} p_{i'} D(\mathbf{x}, \mathbf{P}_i) &\leq e^{-nc_{ij}B(1-2q)}(nq + nc_{ij}(1-2q))(1+\epsilon_0) \\ &\leq e^{-nc_{ij}B(1-2q)/2} \leq e^{-ncB(1-2q)/2} \end{aligned}$$

so

$$\begin{aligned} \frac{\sum_{j\neq i}\sum_{\mathbf{x}\in S_j} p_{i'}^{(1)}(\mathbf{x}) D(\mathbf{x}, \mathbf{P}_i)}{\sum_{\mathbf{x}} p_{i'}^{(1)}(\mathbf{x})} &\leq \frac{1}{mw_T} \sum_{j\neq i} \sum_{\mathbf{x}\in S_j} p_{i'}^{(1)}(\mathbf{x}) D(\mathbf{x}, \mathbf{P}_i) \\ &\leq \frac{1}{w_T} e^{-ncB(1-2q)/2} < \sqrt{6nql} \end{aligned} \tag{A.3}$$

using condition $ncB(1-2q) \geq \ln(8l/3nq)$ (C1). Putting together (A.2) and (A.3) we get the result. □

*Proof of Prop. 9.* a). From Proposition 3 and 6 we have that $|S_i| > mw_i/2$ and at most $15lw_i/8$ initial centers are from $S_i$.

Let $i'$ be such that $\mathbf{T}_{i'}^{(0)} \in S_i$ and $\mathbf{x} \in S_i$. For any $j$ such that $\mathbf{T}_j^{(0)} \notin S_i$ we have from Prop 7 $p_{i'}^{(1)}(\mathbf{x})/p_j^{(1)}(\mathbf{x}) \geq e^{nc_{ij}B(1-2q)} \geq e^{ncB(1-2q)}$. Then $p_j^{(1)}(\mathbf{x}) \leq e^{-ncB(1-2q)}$ and thus $\sum_{k, \mathbf{T}_k^{(1)} \in S_i} p_k^{(1)}(\mathbf{x}) \geq 1 - le^{-ncB(1-2q)}$. But then

$$\begin{aligned} \sum_{k, \mathbf{T}_k^{(1)} \in S_i} w_k^{(1)} &= \frac{\sum_{\mathbf{x}\in S} \sum_{k, \mathbf{T}_k^{(1)}\in S_i} p_k^{(1)}(\mathbf{x})}{m} \\ &\geq \frac{|S_i|(1 - le^{-ncB(1-2q)})}{m} \geq \frac{w_i}{2}(1 - le^{-ncB(1-2q)}) \end{aligned}$$

But $|\{j, \mathbf{T}_j^{(1)} \in S_i\}| \leq 15lw_i/8$ so there is a $j, \mathbf{T}_j^{(1)} \in S_i$ such that

$$\begin{aligned} w_j^{(1)} &\geq \frac{w_i(1 - le^{-ncB})/2}{15lw_i/8} = \frac{1 - le^{-ncB(1-2q)}}{15l/4} \\ &\geq \frac{1}{4l} = w_T \end{aligned}$$

using condition $ncB(1-2q) \geq \ln(16l)$ (C1), thus $C_i$ is not empty.

b) Pick any $\mathbf{T}_{i'}^{(1)} \in C_i$ and $\mathbf{T}_{j'}^{(1)}, \mathbf{T}_{j''}^{(1)} \in C_j$ for $i \neq j$. Then from Proposition 8 we have

$$D(\mathbf{T}_{j'}^{(1)}, \mathbf{T}_{j''}^{(1)}) \leq 2nq + 4\sqrt{6nql}$$

while using Proposition 8 and the triangle inequality we get

$$\begin{aligned} D(\mathbf{T}_{i'}^{(1)}, \mathbf{T}_{j'}^{(1)}) &\geq D(\mathbf{P}_i, \mathbf{P}_j) - 2nq - 4\sqrt{6nql} \\ &\geq nc - 2nq - 4\sqrt{6nql} > 2nq + 4\sqrt{6nql} \end{aligned}$$

from condition $c \geq 4q + 8\sqrt{6ql/n}$ (C3), so we can take $\tau = \frac{1}{2}nc$.

c) There are $k$ true clusters, exactly as many as selected templates. If two selected templates were from the same cluster, there should be a cluster that has no selected templates. But the two templates from the same cluster are at distance at most $\tau$ while the distance of a template from the unselected cluster has distance more than $\tau$, we get a contradiction. □

*Proof of Prop. 10.* Using the triangle inequality, Prop. 3 and Prop. 8 we have

$$D(\mathbf{x}, \mathbf{T}_i^{(1)}) \le D(\mathbf{x}, \mathbf{P}_i) + D(\mathbf{T}_i^{(1)}, \mathbf{P}_i) \le n(q + \epsilon_0\sqrt{q}) + nq + 2\sqrt{6nql}$$

and

$$\begin{aligned} D(\mathbf{x}, \mathbf{T}_j^{(1)}) &\ge D(\mathbf{x}, \mathbf{P}_j) - D(\mathbf{T}_j^{(1)}, \mathbf{P}_j) \\ &\ge n(q + c_{ij}(1-2q))(1-\epsilon_0) - nq - 2\sqrt{6nql}, \end{aligned}$$

so

$$\frac{p_i^{(2)}(\mathbf{x})}{p_j^{(2)}(\mathbf{x})} = \frac{q_0^{D(\mathbf{x},\mathbf{T}_i^{(1)})}(1-q_0)^{n-D(\mathbf{x},\mathbf{T}_i^{(1)})}}{q_0^{D(\mathbf{x},\mathbf{T}_j^{(1)})}(1-q_0)^{n-D(\mathbf{x},\mathbf{T}_j^{(1)})}} = a^{D(\mathbf{x},\mathbf{T}_j^{(1)}) - D(\mathbf{x},\mathbf{T}_i^{(1)})},$$

where $a = \frac{1-q_0}{q_0} > 1$, and therefore

$$\begin{aligned} \frac{p_i^{(2)}(\mathbf{x})}{p_j^{(2)}(\mathbf{x})} &\ge \exp([n(q + c_{ij}(1-2q))(1-\epsilon_0) - n(q+\epsilon_0\sqrt{q}) - -2nq - 4\sqrt{6nql}] \ln a) \\ &= \exp\left(n\left[c_{ij}(1-2q)(1-\epsilon_0) - 2q - \epsilon_0(q+\sqrt{q}) - 4\sqrt{\frac{6ql}{n}}\right]\ln a\right) \\ &\ge \exp\left(nc_{ij}\frac{1}{4}(1-2q)\ln\frac{1}{6\sqrt{q}}\right) \end{aligned}$$

using the condition

$$c(1-2q)\left(\frac{3}{4} - \epsilon_0\right) \ge 2q + \epsilon_0(q+\sqrt{q}) + 4\sqrt{\frac{6ql}{n}}$$

obtained from $\epsilon_0 \le t$. □

*Proof of Theorem 2.* First we compute the probability that the theorem holds.

Proposition 3 holds with probability at least $1 - m^2e^{-2n(1-q)\epsilon_0^2/3} - m^2e^{-n\min(c,0.5)\epsilon_0^2/3} - 2me^{-n\epsilon_0^2/3} - 2me^{-n\min(c,0.5)\epsilon_0^2/3} - ke^{-mw_{min}/8}$. Proposition 6 holds with probability at least $1 - k(l+1)e^{-lw_{min}} - ke^{lw_{min}/4}$. Proposition 8 holds with probability at least $1 - 1/n$ for each of the $k$ clusters. All other propositions hold if these three propositions hold.

Thus with probability $1 - m^2e^{-2n(1-q)\epsilon_0^2/3} - m^2e^{-n\min(c,0.5)\epsilon_0^2/3} - 2me^{-n\epsilon_0^2/3} - 2me^{-n\min(c,0.5)\epsilon_0^2/3} - ke^{-mw_{min}/8} - k(l+1)e^{-lw_{min}} - ke^{lw_{min}/4} - k/n$ all propositions hold for all clusters.

Now we prove the distance inequality. Similar to the proof of Proposition 7 we have

$$\begin{aligned} D(\mathbf{T}_i^{(2)}, \mathbf{P}_i) &= D\left(\frac{\sum_{\mathbf{x}} p_i^{(2)}(\mathbf{x})\mathbf{x}}{\sum_{\mathbf{x}} p_i^{(2)}(\mathbf{x})}, \mathbf{P}_i\right) = \frac{\sum_{\mathbf{x}} p_i^{(2)}(\mathbf{x}) D(\mathbf{x}, \mathbf{P}_i)}{\sum_{\mathbf{x}} p_i^{(2)}(\mathbf{x})} \\ &\le \frac{\sum_{\mathbf{x}\in S_i} p_i^{(2)}(\mathbf{x}) D(\mathbf{x}, \mathbf{P}_i)}{\sum_{\mathbf{x}\in S_i} p_i^{(2)}(\mathbf{x})} + \frac{\sum_{j\neq i}\sum_{\mathbf{x}\in S_j} p_i^{(2)}(\mathbf{x}) D(\mathbf{x}, \mathbf{P}_i)}{\sum_{\mathbf{x}\in S_i} p_i^{(2)}(\mathbf{x})} \end{aligned}$$

From Proposition 10 we have for $\mathbf{x} \in S_i$, $p_j^{(2)}(\mathbf{x}) \le p_i^{(2)}(\mathbf{x}) e^{-ncB/2} \le e^{-ncB/2}$ so

$$p_i^{(2)}(\mathbf{x}) = 1 - \sum_{j\neq i} p_j^{(2)}(\mathbf{x}) \ge 1 - ke^{-ncB/2}$$

So the first term is bounded as:

$$\begin{aligned} \frac{\sum_{\mathbf{x}\in S_i} p_i^{(2)}(\mathbf{x}) D(\mathbf{x}, \mathbf{P}_i)}{\sum_{\mathbf{x}\in S_i} p_i^{(2)}(\mathbf{x})} &\le \frac{\sum_{\mathbf{x}\in S_i} (1 - ke^{-ncB/2}) D(\mathbf{x}, \mathbf{P}_i)}{|S_i|(1 - ke^{-ncB/2})} \\ &+ \frac{\sum_{\mathbf{x}\in S_i} (p_i^{(2)}(\mathbf{x}) - (1 - ke^{-ncB/2})) D(\mathbf{x}, \mathbf{P}_i)}{\sum_{\mathbf{x}\in S_i} p_i^{(2)}(\mathbf{x})} \\ &\le D(\mathrm{mean}(S_i), \mathbf{P}_i) + \frac{\sum_{\mathbf{x}\in S_i} ke^{-ncB/2} D(\mathbf{x}, \mathbf{P}_i)}{|S_i|(1 - ke^{-ncB/2})} \\ &\le D(\mathrm{mean}(S_i), \mathbf{P}_i) + \frac{|S_i| ke^{-ncB/2} n(q + \epsilon\sqrt{q})}{|S_i|(1 - ke^{-ncB/2})} \\ &\le D(\mathrm{mean}(S_i), \mathbf{P}_i) + 2ke^{-ncB/2} nq \end{aligned}$$

when $\epsilon < \sqrt{q}(1 - 2ke^{-ncB/2})$.

The second term is bounded as:

$$\begin{aligned} &\frac{\sum_{j\neq i}\sum_{\mathbf{x}\in S_j} p_i^{(2)}(\mathbf{x}) D(\mathbf{x}, \mathbf{P}_i)}{\sum_{\mathbf{x}\in S_i} p_i^{(2)}(\mathbf{x})} \\ &\le \frac{mne^{-ncB/2}}{|S_i|(1 - ke^{-ncB/2})} \le \frac{2nqe^{-ncB/4}}{w_i(1 - ke^{-ncB/2})} \le \frac{3}{w_i} nqe^{-ncB/4} \end{aligned}$$

when $e^{-ncB/8} < q$ and $ke^{-ncB/2} < 1/3$.

From the inequality

$$ke^{-ncB/4} \le 1 \le \frac{1}{w_{min}}$$

we get the result. □

## References

[1] Dasgupta, S. and Shulman, L. J., A Two-round variant of EM for Gaussian mixtures. *Proceedings of 16th Conference on Uncertainty in Artificial Intelligence (UAI-2000)*, 152–159, 2000.

[2] Daugman, J. G., Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression. *IEEE Trans. on Acoustics, Speech and Signal Processing*, **36**, 1169–1179, 1988.

[3] Dempster, A. P., Laird, N. M., and Rubin, D. B., Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, B*, **39**, 1–38, 1977. MR0501537

[4] Fraley, C. and Raftery, A. E., Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**, 611–631, 2002. MR1951635

[5] Goodman, L. A., Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, **61**, 215–231, 1974. MR0370936

[6] Huo, X. and Donoho, D. L., Applications of beamlets to detection and extraction of lines, curves and objects in very noisy images. *Nonlinear Signal and Image Processing*, 2001.

[7] Schwarz, G. E., Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464, 1978. MR0468014

[8] Si, Z., Gong, H., Zhu, S. C., and Wu, Y. N., Learning active basis models by EM-type algorithms. *Statistical Science*, **25**, 458–475, 2010. MR2807764

[9] Vapnik, V. N., *The Nature of Statistical Learning Theory*. Springer, 2000. MR1719582

[10] Zhu, S. C. and Mumford, D. B., A stochastic grammar of images. *Foundations and Trends in Computer Graphics and Vision*, **2**, 259–362, 2006.