Learning Mixtures of Bernoulli Templates by Two-Round EM with Performance Guarantee

Adrian Barbu, Tianfu Wu and Ying Nian Wu

Provable EM

Dasgupta and Shulman, 2000

A two-round variant of EM for mixture of isotropic Gaussian distributions

Say m data points in $\mathbb{R}^n$ are generated from a c-separated mixture of k Gaussian distributions, and denote by S_i the points from the i-th Gaussian.

We will start with $l = O(k \ln k)$ clusters

Provable EM Algorithm

Initialize $\tilde{\mu}_i,\ i = 1, \cdots, l$ as random data points and let $\tilde{\omega}_i = \frac{1}{l}$,

$\tilde{\sigma}^2 = \frac{1}{2n} \min_{j \neq i} ||\tilde{\mu}_i - \tilde{\mu}_j||$

1. One EM step
2. Pruning Step
3. One EM Step

The results: $||\tilde{\mu}_i - \mu_i|| \leq ||mean(S_i) - \mu_i|| + e^{-\Omega(c^2 n)}$, which means we couldn't do much better than EM even if we knew the point labels

What if the mixture is NOT of Gaussian distributions?

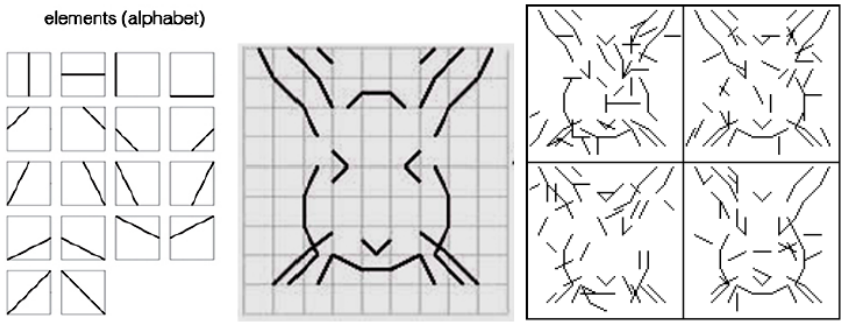
Provable Two-Round EM

for mixture of Bernoulli templates

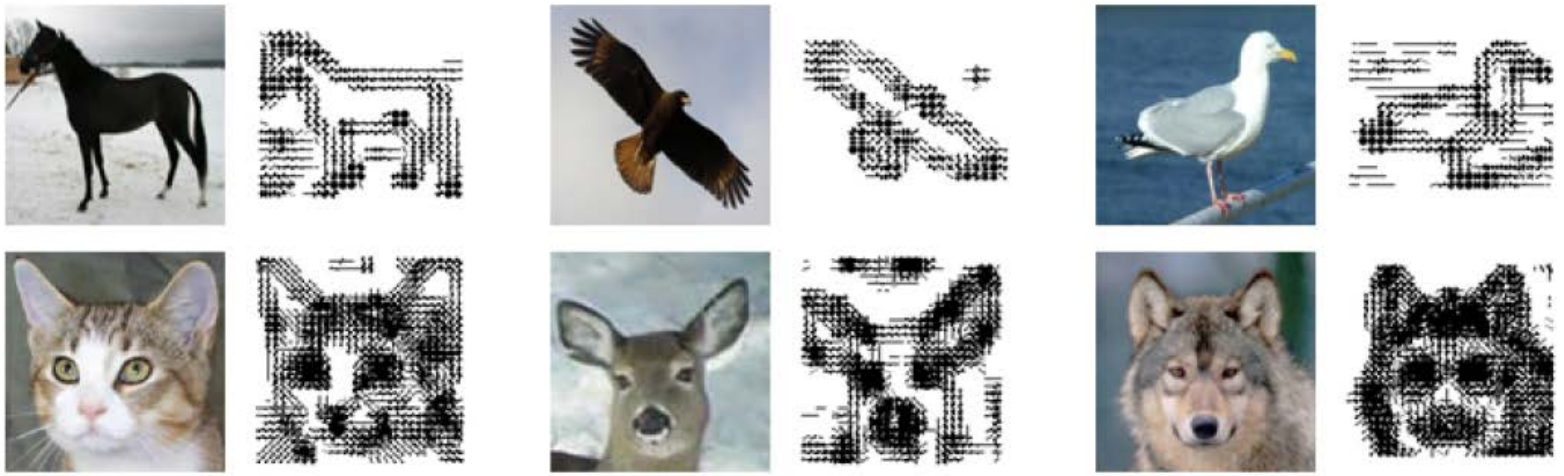
A Bernoulli template P: a n-dimensional binary vector, i.e., $P \in \Omega = \{0,1\}^n$

An example $\mathbf{x}$ generated by P: a Bernoulli noisy version of P, $\mathbf{x} \sim P$,

$$x(s) = \begin{cases} P(s), & \text{with prob. } 1-q \\ 1-P(s), & \text{with prob. } q \end{cases}$$
$$q \in (0, 1/2)$$



$$n = 9 \times 9 \times 18 = 1458$$



Model

i. $\{(P_i, \omega_i), i = 1, \cdots, k\}$: k Bernoulli templates with mixture weights ω_i.
(Assume k is given)

ii. $\omega_{\min} = \min_{i=1,\cdots,k} \omega_i$

ii. $\{\mathbf{x}_1, \cdots, \mathbf{x}_m\}$: m noisy observations of the k templates with the noise level q

iii. $\mu_i = E[\mathbf{x}_i]$ where $\mathbf{x}_i \sim P_i$

iv. S_i : the set of examples coming from P_i

v. $D(P_i, P_j) = \sum_{s=1}^{n} |P_i(s) - P_j(s)| \triangleq d_{ij}$: ℓ_1 distance

vi. $c_{ij} = \frac{1}{n} d_{ij}$: the separation between P_i and P_j

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

The challenge from *Gaussian mixtures* to the *mixtures of Bernoulli distributions*:

- The sample space is no longer Euclidean, and some results for Gaussian distributions cannot be translated directly into those for the Bernoulli models.

Algorithm

Denote by $\mathbf{T}_i$ the estimated P_i.

Start with $l = \frac{4}{\omega_{\min}} \ln \frac{2}{\delta \omega_{\min}}$ clusters where δ is the confidence parameter.

Two-Round EM for Learning Bernoulli Templates

Initialize $\{\mathbf{T}_i^{(0)}, i = 1, \cdots, l\}$ as random training examples ($l > k$), $\omega_i^{(0)} = \frac{1}{l}$,

Estimate the noise level q_0 such that $q_0(1-q_0) = \min_{i,j} D(\mathbf{T}_i^{(0)}, \mathbf{T}_j^{(0)})/2n$

1. One EM step

$$f_i(\mathbf{x}_j) = q_0^{D(\mathbf{x}_j, \mathbf{T}_i^{(0)})}(1-q_0)^{n - D(\mathbf{x}_j, \mathbf{T}_i^{(0)})}, j = 1, ..., m,$$

$$p_i^{(1)}(\mathbf{x}_j) = \frac{w_i^{(0)} f_i(\mathbf{x}_j)}{\sum_{i'} w_{i'}^{(0)} f_{i'}(\mathbf{x}_j)}, j = 1, ..., m$$

$$w_i^{(1)} = \sum_{j=1}^{m} p_i^{(1)}(\mathbf{x}_j)/m$$

$$\mathbf{T}_i^{(1)} = \frac{1}{m w_i^{(1)}} \sum_{j=1}^{m} p_i^{(1)}(\mathbf{x}_j)\mathbf{x}_j$$

2. Pruning Step

Remove all $\mathbf{T}_i^{(1)}$ with $w_i^{(1)} < w_T = \frac{1}{4l}$

Keep only k templates $\mathbf{T}_i^{(1)}$ far apart.

3. One EM Step

$$f_i(\mathbf{x}_j) = q_1^{D(\mathbf{x}_j, \mathbf{T}_i^{(1)})}(1-q_1)^{n - D(\mathbf{x}_j, \mathbf{T}_i^{(1)})}, j = 1, ..., m$$

$$p_i^{(2)}(\mathbf{x}_j) = \frac{w_i^{(1)} f_i(\mathbf{x}_j)}{\sum_{i'} w_{i'}^{(1)} f_{i'}(\mathbf{x}_j)}, j = 1, ..., m$$

$$w_i^{(2)} = \sum_{j=1}^{m} p_i^{(2)}(\mathbf{x}_j)/m,$$

$$\mathbf{T}_i^{(2)} = \frac{1}{m w_i^{(2)}} \sum_{j=1}^{m} p_i^{(2)}(\mathbf{x}_j)\mathbf{x}_j$$

Main Theorem

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 conditions $C1 - C3$ hold and in addition 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* $m \geq \frac{8}{w_{min}} \ln \frac{12k}{\delta}$.
3. *The separation is* $c > \frac{4}{nB} \ln \frac{5n}{\epsilon w_{min}}$.
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:

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

C1: $nc > \frac{1}{B(1-2q)} \ln \frac{16l}{\min(6nq, 1)}$

C2: $m > \max(16 \ln n, 8l)$

C3: $c > \max(1, \frac{2}{3(1-2q)})(4q + 8\sqrt{6ql/n})$

Bernoulli templates are sufficiently different from each other, and n and m are sufficiently large.

Sketch of Proof

The proof follows the steps of the two-round EM.

1. Initialization:

   We prove that the initial templates cover all the clusters and the estimated noise level q_0 is close to the true noise level q .

2. One EM step

   We prove that the estimated templates are likely to be close to the true templates of the same clusters.

3. Pruning Step

   We prove that it is very likely that exactly one template is kept for each cluster.

4. One EM Step

   The estimated templates are proved to be near optimal

Experimental Results I: *simulation study*

Settings: a mixture of two templates $P_0 = \mathbf{0}$ and $P_1 = (1, \cdots, 1, 0, \cdots, 0)$ where the number of 1's is $\lfloor cn \rfloor$;

For the original EM we also assumed the noise level q is a known parameter.

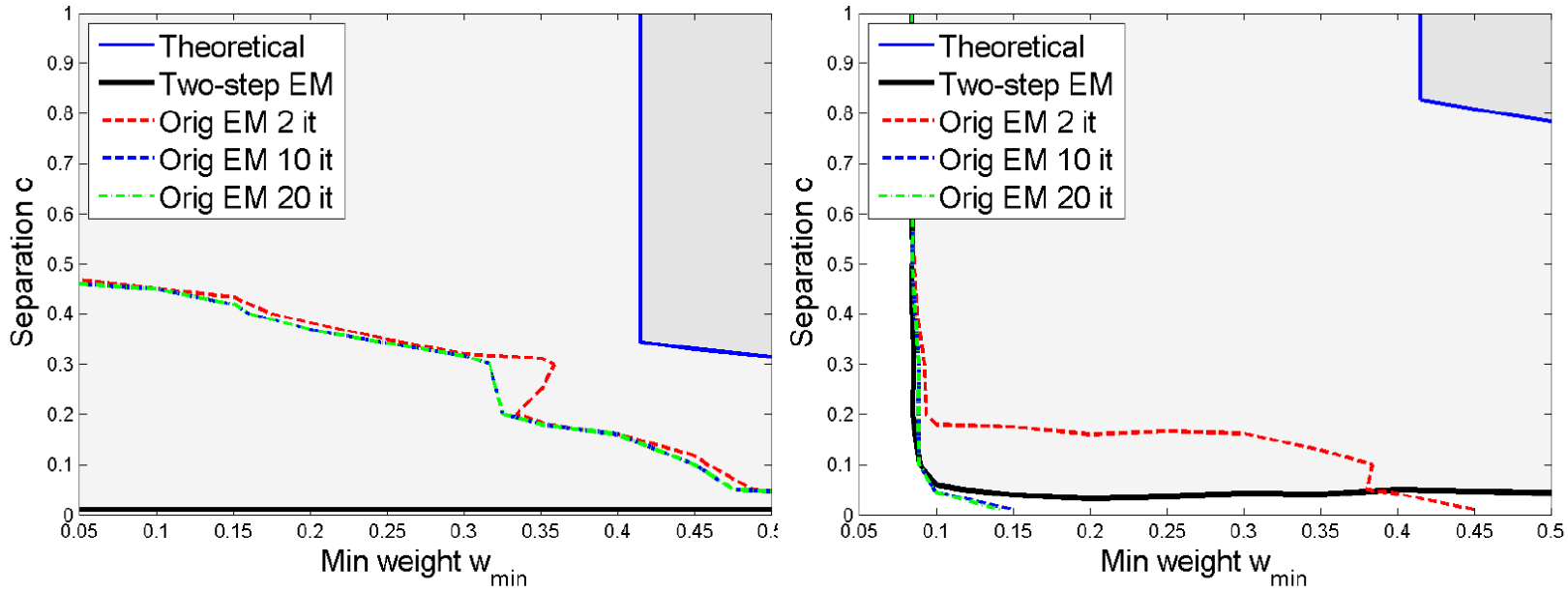


FIG 3. *Domains where the two-round EM and the standard EM find the* $k = 2$ *binary templates correctly* 90% *of the time when* $m = 300$. *The first plot is for* $q = .01$, *with* $n = 2,000$, *and the second plot is for* $q = .1$ *with* $n = 10,000$. *Also shown is the domain theoretically guaranteed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

Experimental Results I: *simulation study (Cont.)*

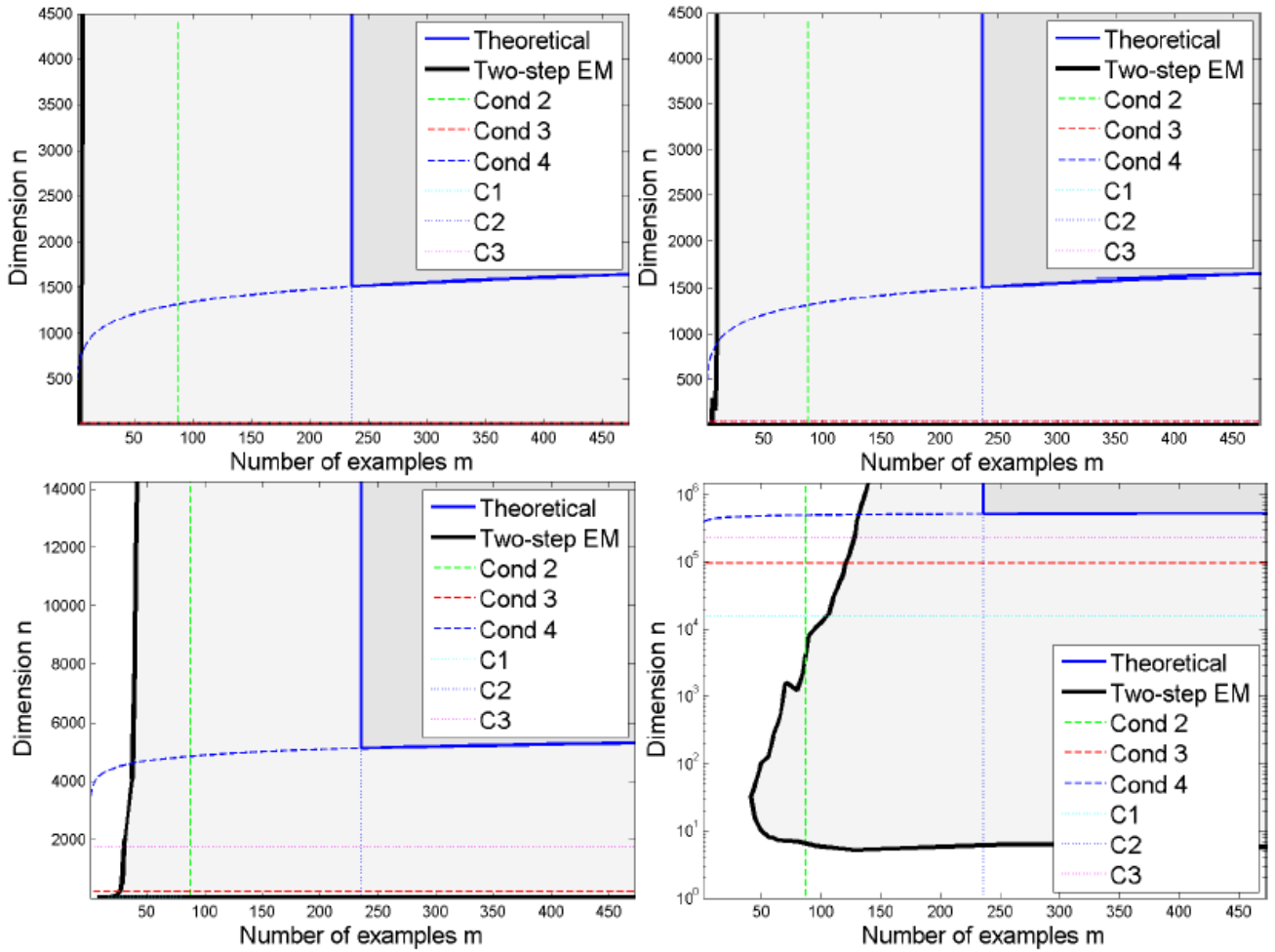
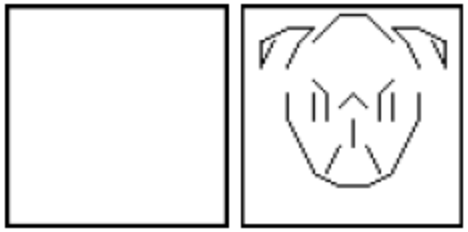


FIG 4. *Theoretical and practical domains of validity of the two-step EM algorithm for four noise levels. From left to right are noise levels:* $q = 0.0001, q = 0.01$ *(top) and* $q = 0.1, q = 0.2$ *(bottom). In these examples* $c = 1, k = 2, w_{min} = 0.5, \delta = \epsilon = 0.1$. *Each domain is above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

Experimental Results II: *synthetic image sketches.*

Two templates:



$$n = 9 \times 9 \times 18 = 1458$$

Examples:

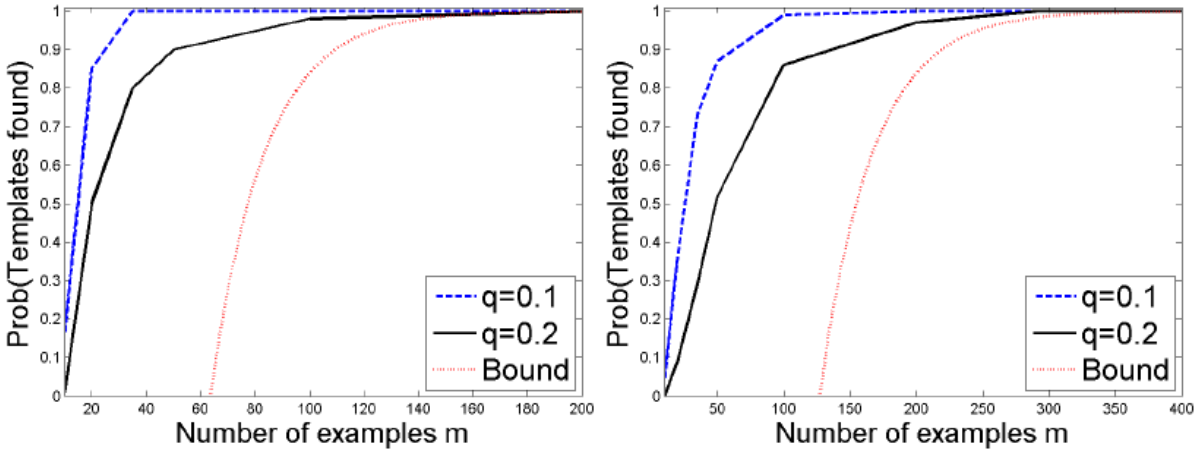


FIG 6. *Success rates vs. number of training examples for learning from a mixture of two templates with the two-round EM algorithm for two levels of noise* $q \in \{0.1, 0.2\}$ *and two mixture weights* $w_{min} = 0.4$ *(left) and* $w_{min} = 0.2$ *(right).*

Experimental Results III: *real images*

Evaluation metrics: conditional purity and conditional entropy.

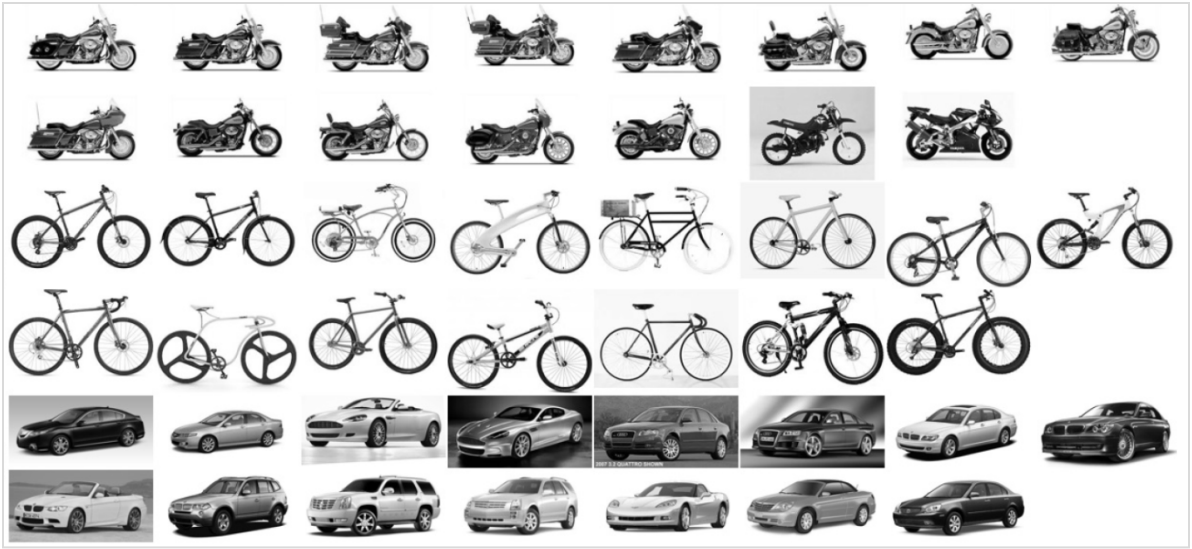
Given the underlying ground-truth category labels X (which is unknown to the algorithm) and the obtained cluster labels Y

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

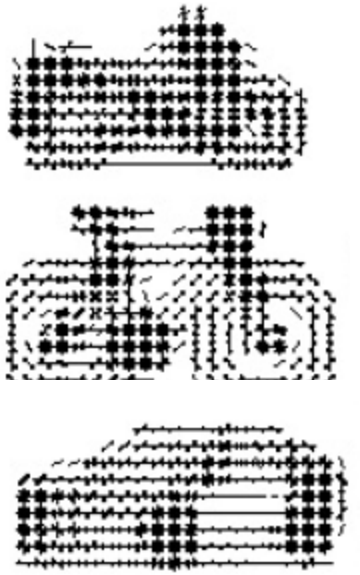
$$\mathcal{H}(X|Y) = \sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log\left(\frac{1}{p(x|y)}\right)$$

Experimental Results III: *real images*

Motorcycles, bicycles and cars



Recovered templates



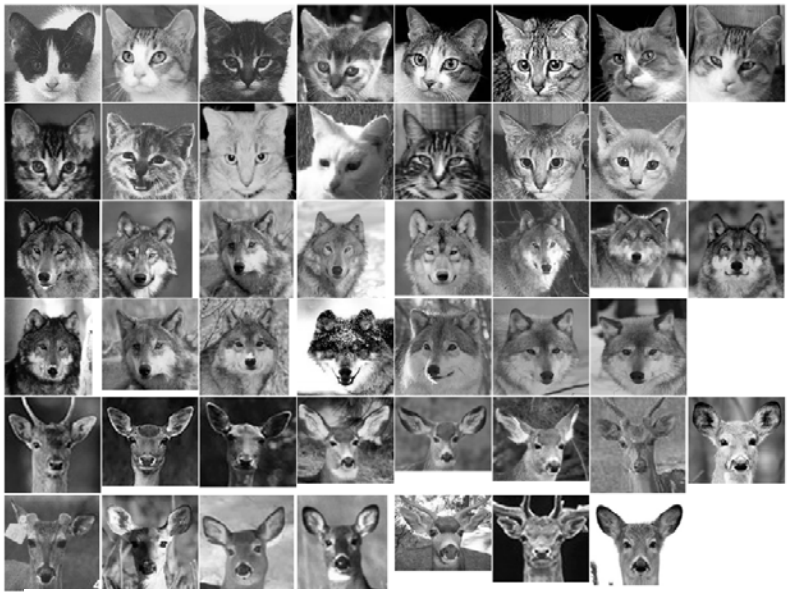
		Cond. Purity	Cond. Entropy	$\text{Num}_{\text{perfect}}$ /100
Two-round EM		**0.944**± 0.108	**0.106**± 0.183	**61**
Original EM	maxIter=2	0.851±0.157	0.247± 0.231	25
	maxIter=10	0.886±0.155	0.170±0.211	47
	maxIter=20	0.886±0.155	0.170±0.211	47
	maxIter=100	0.886±0.155	0.170±0.211	47

TABLE 1

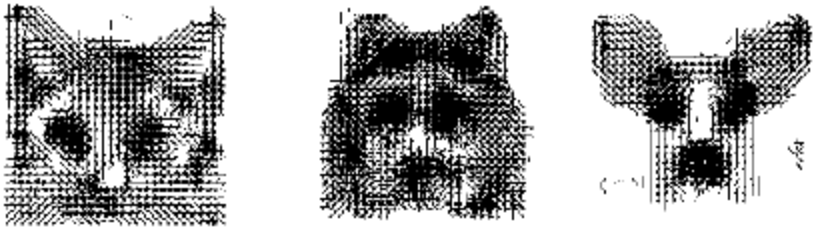
Performance comparison between our two-round EM algorithm and the original EM algorithm for clustering motorcycles, bicycles and cars. In the table, $\text{Num}_{perfect}$ means the number of runs (out of the total 100 runs) that recover the underlying clusters perfectly.

Experimental Results III: *real images*

cats, wolves and deers



Recovered templates



		Cond. Purity	Cond. Entropy	$\text{Num}_{\text{perfect}}$ /100
Two-round EM		**0.846**± 0.139	**0.319**± 0.234	**12**
Original EM	maxIter=2	0.704±0.139	0.581± 0.245	2
	maxIter=10	0.744±0.152	0.494±0.271	8
	maxIter=20	0.744±0.152	0.494±0.271	8
	maxIter=100	0.744±0.152	0.494±0.271	8

TABLE 2

Performance comparison between our two-round EM algorithm and the original EM algorithm for clustering cats, wolves and deers. In the table, $\text{Num}_{\text{perfect}}$ means the number of runs (out of the total 100 runs) which recover the underlying clusters perfectly.

Experimental Results III: *real images*

car, motorcycle and bicycle + cats, wolves and deers

		Cond. Purity	Cond. Entropy	$\text{Num}_{\text{perfect}}$ /100
Two-round EM		**0.851**± 0.095	**0.323**± 0.190	**5**
Original EM	maxIter=2	0.751±0.121	0.509± 0.223	0
	maxIter=10	0.803±0.129	0.360±0.229	3
	maxIter=20	0.803±0.129	0.360±0.229	3
	maxIter=100	0.803±0.129	0.360±0.229	3

TABLE 3

Performance comparison between our two-round EM algorithm and the original EM algorithm for clustering cats, wolves, deers, motorcycles, bicycles and cars. In the table, $\text{Num}_{\text{perfect}}$ means the number of runs (out of the total 100 runs) which recover the underlying clusters perfectly.

Summary

Provable EM gives us advice how to design better EM the right way in general:

1. Start with more cluster centers than k
2. Do one or more EM steps
3. Prune weak cluster centers and keep the best separated k clusters
4. Do one or more EM steps