Estimation of Non-Normalized Mixture Models (AISTATS 2019)

Takeru Matsuda (The University of Tokyo, RIKEN Center for Brain Science)
Aapo Hyvärinen (University College London, University of Helsinki)

Problem Setting

non-normalized model = statistical model with an intractable normalization constant

- Markov random field, Boltzmann machine, overcomplete ICA, ...

$$p(x \mid \theta) = \frac{1}{Z(\theta)} \widetilde{p}(x \mid \theta), \quad Z(\theta) = \int \widetilde{p}(x \mid \theta) \mathrm{d}x$$

computationally intractable

non-normalized mixture model
= finite mixture of non-normalized models

$$p(x \mid \theta, \pi) = \sum_{k=1}^{K} \pi_k \cdot p(x \mid \theta_k)$$

$$p(x \mid \theta_k) = \frac{1}{Z(\theta_k)} \widetilde{p}(x \mid \theta_k), \quad \pi_k > 0, \quad \sum_{k=1}^{K} \pi_k = 1$$

$$x_1, \cdots, x_N \sim p(x \mid \theta, \pi)$$

We develop a general method for estimating $\theta = (\theta_1, \cdots, \theta_K)$ and $\pi = (\pi_1, \cdots, \pi_K)$ without computing $Z(\theta_k)$

- extension of noise contrastive estimation (Gutmann and Hyvärinen, 2012)
- can even be used on deep image representations

Proposed Method

Reparametrization: $(\theta, \pi) \to (\theta, c)$

$$p(x \mid \theta, c) = \sum_{k=1}^{K} p(x \mid \theta_k, c_k)$$

$$\log p(x \mid \theta_k, c_k) = \log \widetilde{p}(x \mid \theta_k) + c_k, \quad c_k = \log \pi_k - \log Z(\theta_k)$$

Noise generation

We generate artificial noise $y_1, \cdots, y_M \sim n(y)$

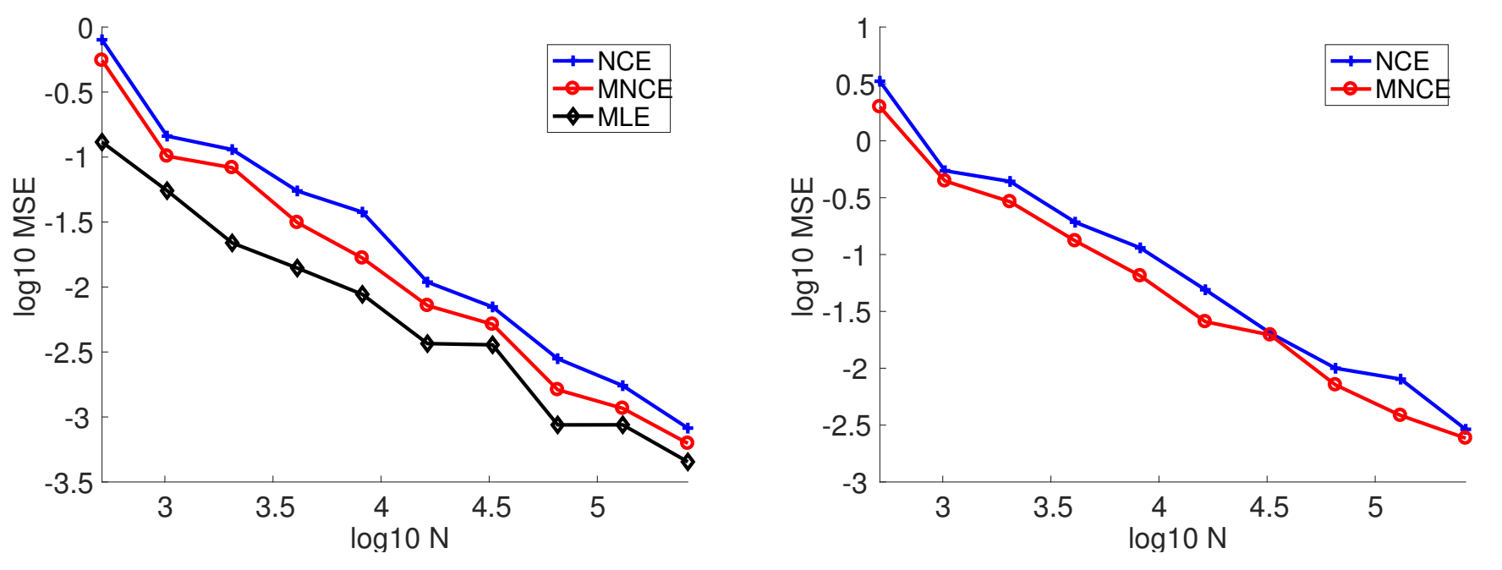
- should be difficult to discriminate from data (cf. GAN)
- e.g., Gaussian with same mean and covariance as data

idea: estimate θ and c by discriminating between data and noise

$$(\hat{\theta}, \hat{c}) = \arg\max_{\theta, c} \sum_{t=1}^{N} \log \frac{N p(x_t \mid \theta, c)}{N p(x_t \mid \theta, c) + M n(x_t)} + \sum_{t=1}^{M} \log \frac{M n(y_t)}{N p(y_t \mid \theta, c) + M n(y_t)}$$

Gaussian mixture

$$p(x \mid \theta_k, c_k) = \exp(\theta_{k1} x^2 + \theta_{k2} x + c_k)$$



This estimator has consistency under mild regularity conditions (Theorem 1)

We can improve estimation accuracy by using multiple noise distributions:

$$y_1^{(l)}, \cdots, y_{M_l}^{(l)} \sim n_l(y) \ (l = 1, \cdots, L)$$

$$(\hat{\theta}, \hat{c}) = \arg\max_{\theta, c} \sum_{t=1}^{N} \log \frac{N p(x_t \mid \theta, c)}{N p(x_t \mid \theta, c) + \sum_{l=1}^{L} M_l n_l(x_t)} + \sum_{l=1}^{L} \sum_{t=1}^{M_l} \log \frac{M_l n_l(y_t^{(l)})}{N p(y_t^{(l)} \mid \theta, c) + \sum_{l=1}^{L} M_l n_l(x_t)}$$

This estimator is equivalent to the original one with a mixture noise distribution (Theorem 2)

Clustering with Deep Representation

x: data (e.g., image), z: label (e.g., "dog")

Classification with neural network (softmax)

$$p(z = l \mid x) = \frac{\exp(\sum_{i=1}^{d} w_{li} f_i(x))}{\sum_{j=1}^{L} \exp(\sum_{i=1}^{d} w_{ji} f_i(x))}$$

$f(x) = (f_1(x), \cdots, f_d(x))$: feature vector (activation of last hidden layer)

→

Non-normalized exponential family

$$p(x \mid z = l) = \underset{\substack{\uparrow \\ \text{unknown}}}{h(x)} \exp\left(\sum_{i=1}^{d} w_{li} f_i(x) - \underset{\substack{\uparrow \\ \text{unknown}}}{A(w_l)} \right)$$

We propose a probabilistically principled method for transferring the deep representation f to clustering of unlabeled data $x_1, \cdots, x_N$.

$$p(x \mid \theta, c) = h(x) \sum_{k=1}^{K} \exp\left(\sum_{i=1}^{d} \theta_{ki} f_i(x) + c_k \right) \quad \text{: target data}$$

$$n_l(x) = h(x) \exp\left(\sum_{i=1}^{d} w_{li} f_i(x) - A(w_l) \right) \quad \text{: original training data}$$

$$(l = 1, \cdots, L)$$

$\to (\hat{\theta}, \hat{c}) \to p(z_t = k \mid x_t; \hat{\theta}, \hat{c}) \to$ clustering

Note: the unknown function h cancels out

Image clustering

data: 12,500 dog images & 12,500 cat images

deep representation: inception-v3 ($d = 2048$, $L = 149$, $K = 2$)

clustering results (GMM = Gaussian Mixture Model; diagonal/isotropic covariance)

proposed	dog	cat
cluster 1	12400	145
cluster 2	100	12355

GMM1	dog	cat
cluster 1	12490	325
cluster 2	10	12175

GMM2	dog	cat
cluster 1	12490	792
cluster 2	10	11708

The proposed method has better classification accuracy

estimated and actual numbers of misclassifications

	estimate	actual
proposed	169.98	245
GMM1	0.66	335
GMM2	5.58	802

estimate

$= \sum_x \min(p(z = 1 \mid x), p(z = 2 \mid x))$

The proposed method quantifies clustering uncertainty more accurately

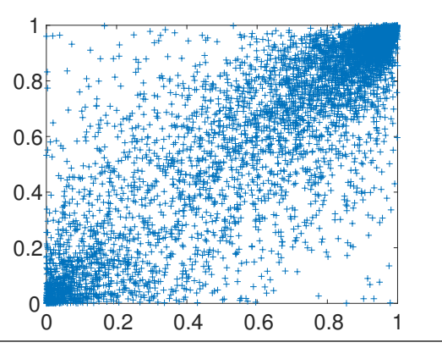
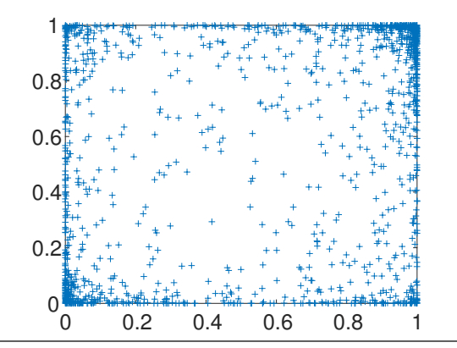
Brain state clustering

data: 306-ch magnetoencephalography (MEG) from CamCAN repository

deep representation: obtained by nonlinear ICA with Time Contrastive Learning (Hyvärinen and Morioka, 2016)

scatter plots of $p(z_{t-1} = 1 \mid x_{t-1})$ and $p(z_t = 1 \mid x_t)$ for $K = 2$

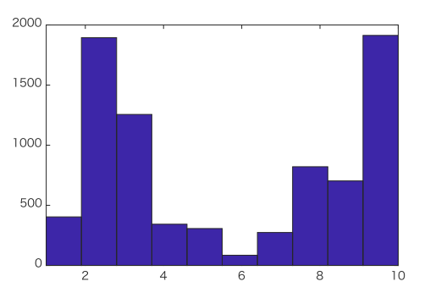
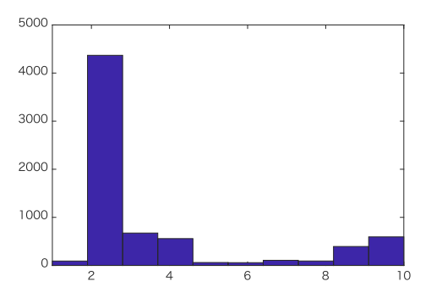
→ proposed method extracts stable brain states

proposed	GMM
	

state histogram for $K = 10$

→ Resting MEG has more temporal variability of brain states

- consistent with previous findings in neuroscience

resting	task
	
1.996 bits	1.351 bits