Supplementary Material
Estimation of Non-Normalized Mixture Models

Proof of Lemma 1 (Section 3.3)

Here, we consider the non-normalized mixture model

$$p(x \mid \theta, \pi) = \sum_{k=1}^{K} \pi_k \cdot p(x \mid \theta_k), \tag{34}$$

where the K components belong to the same parametric family:

$$p(x \mid \theta_k) = \frac{1}{Z(\theta_k)} \widetilde{p}(x \mid \theta_k). \tag{35}$$

We also use the parametrization with (θ, c) defined by

$$p(x \mid \theta, c) = \sum_{k=1}^{K} p(x \mid \theta_k, c_k), \tag{36}$$

where

$$\log p(x \mid \theta_k, c_k) = \log \widetilde{p}(x \mid \theta_k) + c_k. \tag{37}$$

Two parametrizations are connected by the transformation $c_k = \log \pi_k - \log Z(\theta_k)$.

From Theorem 1 of Gutmann and Hyvärinen (2012) with assumption (d),

$$(\theta, c) \in \arg\max_{\theta, c} J_{\mathrm{NCE}}(\theta, c) \tag{38}$$

if and only if

$$p(x \mid \theta, c) = p(x \mid \theta^*, c^*), \tag{39}$$

which is rewritten as

$$\sum_{k=1}^{K} \pi_k p(x \mid \theta_k) = \sum_{k=1}^{K} \pi_k^* p(x \mid \theta_k^*). \tag{40}$$

Then, from assumptions (a)-(c), it leads to

$$\begin{aligned} &\{\pi_1 p(x \mid \theta_1), \cdots, \pi_K p(x \mid \theta_K)\} \\ =&\{\pi_1^* p(x \mid \theta_1^*), \cdots, \pi_K^* p(x \mid \theta_K^*)\}. \end{aligned} \tag{41}$$

Therefore, there exists $\sigma \in S_n$ such that

$$\pi_k p(x \mid \theta_k) = \pi_{\sigma(k)}^* p(x \mid \theta_{\sigma(k)}^*) \tag{42}$$

for $k = 1, \cdots, K$, which is equivalent to

$$\pi_k = \pi_{\sigma(k)}^*, \quad \theta_k = \theta_{\sigma(k)}^* \tag{43}$$

for $k = 1, \cdots, K$ by using assumption (a). Thus, we obtain (16).

Proof of Theorem 2 (Section 4.2)

Since $n(y_t)$ does not depend on θ and c, we can rewrite (3) and (4) as

$$(\hat{\theta}_{\mathrm{NCE}}, \hat{c}_{\mathrm{NCE}}) = \arg\max_{\theta,c} \widetilde{J}_{\mathrm{NCE}}(\theta, c), \tag{44}$$

where

$$\begin{aligned}\widetilde{J}_{\mathrm{NCE}}(\theta, c) = &\sum_{t=1}^{N} \log \frac{Np(x_t \mid \theta, c)}{Np(x_t \mid \theta, c) + Mn(x_t)} \\ &+ \sum_{t=1}^{M} \log \frac{1}{Np(y_t \mid \theta, c) + Mn(y_t)}\end{aligned} \tag{45}$$

Similarly, since $n_l(y_t^{(l)})$ does not depend on θ and c, we can rewrite (12) and (13) as

$$(\hat{\theta}_{\mathrm{MNCE}}, \hat{c}_{\mathrm{MNCE}}) = \arg\max_{\theta,c} \widetilde{J}_{\mathrm{MNCE}}(\theta, c), \tag{46}$$

where

$$\begin{aligned}&\widetilde{J}_{\mathrm{MNCE}}(\theta, c) \\ =&\sum_{t=1}^{N} \log \frac{Np(x_t \mid \theta, c)}{Np(x_t \mid \theta, c) + M_1 n_1(x_t) + \cdots + M_L n_L(x_t)} \\ +&\sum_{l=1}^{L} \sum_{t=1}^{M_l} \log \frac{1}{Np(y_t^{(l)} \mid \theta, c) + M_1 n_1(y_t^{(l)}) + \cdots + M_L n_L(y_t^{(l)})}.\end{aligned} \tag{47}$$

Now, from (14), we obtain $\widetilde{J}_{\mathrm{NCE}}(\theta, c) = \widetilde{J}_{\mathrm{MNCE}}(\theta, c)$. Therefore,

$$(\hat{\theta}_{\mathrm{MNCE}}, \hat{c}_{\mathrm{MNCE}}) = (\hat{\theta}_{\mathrm{NCE}}, \hat{c}_{\mathrm{NCE}}). \tag{48}$$

Additional figure for Section 6

Figure 1 shows the histogram of the logit score of the posterior probability in the first cluster $\log p(z = 1 \mid x; \hat{\theta}, \hat{c}) - \log(1 - p(z = 1 \mid x; \hat{\theta}, \hat{c}))$. Compared to the proposed method, the Gaussian mixture models assign extremely large or small logit scores.

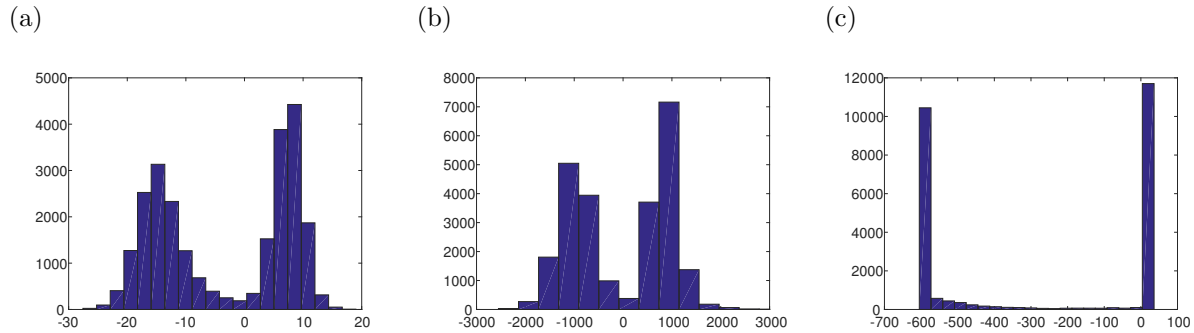

Figure 1: Histogram of the logit score of posterior probability. Please note different horizontal ranges in the three plots. (a) The proposed method. (b) Gaussian mixture model with diagonal covariance matrices. (c) Gaussian mixture model with isotropic covariance matrices.