Open-Set Recognition with Gaussian Mixture
Variational Autoencoders: Supplementary Material

Alexander Cao
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208
a-cao@u.northwestern.edu

Yuan Luo
Department of Preventive Medicine
Northwestern University
Chicago, IL 60611
yuan.luo@northwestern.edu

Diego Klabjan
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208
d-klabjan@northwestern.edu

1 Neural network assumptions

We call a neural network \( f_\tau \) an \( n \)-headed neural network if

1. \( f_\tau : \mathbb{R}^m \to \prod_{i=1}^n \mathbb{R}^s \), i.e. it maps \( b \) to \((a_1, a_2, ..., a_n)\) with \( a_i \in \mathbb{R}^s \),

2. for each \( i, 1 \leq i \leq n \), we have \( a_i = f_{i, \ell_i} \circ f_{i, \ell_i-1} \circ ... \circ f_{i,t+1} \circ f_t \circ ... \circ f_1(b) \) for an integer \( t \) not depending on \( i \), \( \ell_i \geq t + 1 \), and each \( f_j, f_j^i \) is a typical neural network single layer parameterized by a matrix and a bias vector, and it includes an activation function. Vector \( \tau \) corresponds to all these parameters.

In GMVAE, neural networks corresponding to \( q_{\phi_z}, q_{\phi_w} \) are 2-headed neural networks (mean and covariance) with \( \phi_z, \phi_w \) denoting all of the respective parameters. Probability \( p_\beta \) is a 1 or 2-headed network with parameters \( \theta \), and \( p_\beta \) for \( \beta = (\beta_{K_1}, \beta_{K_2}, ..., \beta_{K_C}) \) consists of a \( (2 \sum_{c=1}^C K_c) \)-headed neural network.

**Assumption 1.** In each network \( q_{\phi_z}, q_{\phi_w}, p_\theta, \) and \( p_\beta \), the last layer in each head \( f_{i,\ell_i} \) has an identity activation function.

**Assumption 2.** Neural network \( p_{\beta'} \) for \( \beta' = (\beta_{K_1}, ..., \beta_{K_{c+1}}, ..., \beta_{K_C}) \) consists of \( p_\beta \) with simply two additional heads, while all other network architectures are the same.

**Lemma 1.** Under Assumption 1 for an \( n \)-headed network, we have that given any \( \overline{a} = (\overline{a}_1, ..., \overline{a}_n) \), there exists \( \tau = \tau(\overline{a}) \) such that \( f_\tau(b) = \overline{a} \) for every \( b \).

**Proof.** Let \( \overline{a} \) be given. We define \( \tau \) to consist of 0 matrices and biases for each layer except \( f_{i,\ell_i} \). In \( f_{i,\ell_i} \), the matrix is 0 but the bias is \( \overline{a}_i \). Since \( f_{i,\ell_i} \) has the identity activation, it follows \( f_\tau(b) = \overline{a} \) for every \( b \).

Preprint. Under review.
2  Proof of Proposition 1

Proposition 1. Let us assume that \( x \in \mathcal{X} \) is distributed as \( x \sim p_{\text{data}} = \mathcal{B}(\mu, \sigma^2, C = 1) \), and Assumption 1 holds. Then the optimal GMVAE loss is constant with respect to \( K \). In fact, we have that \( \min -\mathbb{E}_X[L(K)] = -\mathbb{E}_X[\log p_{\text{data}}] \) for every \( K \geq 1 \) and a globally optimal solution reads

\[
\begin{align*}
\mu(x; \phi^*_w) &= \mu_{z=1,b}(w; \beta^*) = \mu_z \\
\sigma^2(x; \phi^*_w) &= \sigma^2_{z=1,b}(w; \beta^*) = \sigma^2_z \\
\mu(x, y; \phi^*_w) &= 0 \\
\sigma^2(x, y; \phi^*_w) &= 1 \\
\mu(z; \theta^*) &= \mu_z
\end{align*}
\]

for any constant vectors \( \mu_z, \sigma_z \).

Proof. Note that \( (\phi^*_w, \phi^*_w, \beta^*, \theta^*) \) exist due to Assumption 1 and Lemma 1. First, we show that \( (\theta^*, \beta^*) \) given in (1) maximize the log likelihood \( \mathbb{E}_X[\log p_{\theta, \beta}(x|y = 1)] \) and results in \( p_{\theta^*, \beta^*}(x|y = 1) = p_{\text{data}} \). We have

\[
KL(p_{\text{data}}||p_{\theta, \beta}(x|y = 1)) = \mathbb{E}_X[\log p_{\text{data}}] - \mathbb{E}_X[\log p_{\theta, \beta}(x|y = 1)]
\]

and thus maximizing \( \mathbb{E}_X[\log p_{\theta, \beta}(x|y = 1)] \) is equivalent to minimizing \( KL(p_{\text{data}}||p_{\theta, \beta}(x|y = 1)) \). The global minimum of \( KL(p_{\text{data}}||p_{\theta, \beta}(x|y = 1)) \) is clearly when \( p_{\text{data}} = p_{\theta, \beta}(x|y = 1) \). This is indeed the case for \( (\theta^*, \beta^*) \), since

\[
p_{\theta^*, \beta^*}(x|y = 1) = \int_{w, z, v} p_{\theta^*, \beta^*}(x, v, w, z|y = 1)dwdzdv
\]

because of GMVAE’s generative model factorization and (1). Now we have

\[
\mathbb{E}_X[\log p_{\text{data}}] = \mathbb{E}_X[\log p_{\theta^*, \beta^*}(x|y = 1)]
\]

\[
= \mathbb{E}_X\left[ \mathbb{E}_{q_{\theta^*, \beta^*}(v, w, z|x, y = 1)}\left[ \log \frac{p_{\theta^*, \beta^*}(x, z, w, v|y = 1)}{q_{\theta^*, \beta^*}(v, w, z|x, y = 1)} \right] \right]
\]

\[
+ \mathbb{E}_X\left[ \mathbb{E}_{q_{\theta^*, \beta^*}(v, w, z|x, y = 1)}\left[ \log \frac{q_{\theta^*, \beta^*}(v, w, z|x, y = 1)}{p_{\theta^*, \beta^*}(x, z, w, v|y = 1)} \right] \right] = \mathbb{E}_X[L(K; \phi^*_w, \phi^*_w, \beta^*, \theta^*)] + \mathbb{E}_X[\text{VG}(\phi^*_z, \phi^*_w, \beta^*, \theta^*)]
\]

where \( \text{VG}(\phi^*_z, \phi^*_w, \beta^*, \theta^*) \) corresponds to (3). We next show that \( \text{VG}(\phi^*_z, \phi^*_w, \beta^*, \theta^*) = 0 \). This together with the facts that maximized \( \mathbb{E}_X[L(K; \phi^*_z, \phi^*_w, \beta, \theta)] \) corresponds with minimized \( \mathbb{E}_X[\text{VG}(\phi^*_z, \phi^*_w, \beta, \theta)] \), and \( \text{VG}(\phi^*_z, \phi^*_w, \beta, \theta) \geq 0 \) (it is a KL divergence), shows optimality.

From (1) we have that \( p_{\theta^*}(x|z) = p_{\text{data}}(x) \) for all \( x \) and \( z \) and thus with (2) we have

\[
p_{\theta^*, \beta^*}(z, w, v|x, y = 1) = \frac{p_{\theta^*}(x|z, w, v, y = 1)p_{\beta^*}(z, w, v|y = 1)}{p_{\theta^*, \beta^*}(x|y = 1)} = \frac{p_{\theta^*}(x|z)p_{\beta^*}(z, w, v|y = 1)}{p_{\text{data}}(x)} = \frac{p_{\beta^*}(z, w, v|y = 1)}{p_{\beta^*}(z, w|y = 1)}.
\]

The reconstruction term \( p_{\theta}(x|z, w, v, y = 1) = p_{\theta}(x|z) \) for every \( \theta \) because in GMVAE, data reconstruction depends only on \( z \) and is independent of \( w \) and \( v \) (see §3.1 of the paper).

Also from Bayes’ and GMVAE’s generative model factorization, we have the following simplification

\[
p_{\beta^*}(v|z, w, y = 1) = \frac{p_{\beta^*}(z, w, y = 1)v(y = 1)p(w)}{p_{\beta^*}(z, w|y = 1)}
\]
which is a valid choice by Assumption 2, and have

\[
p_{\beta^*}(z|w, y = 1, v) p(v|y = 1) p(w) = p_{\beta^*}(z|w, y = 1) p(v|y = 1) p(w) = \sum_{v'} p_{\beta^*}(z|w, y = 1, v') p(v'|y = 1) = p(v|y = 1) \tag{6}
\]

where (1) is only used in the last line. Substituting (5) into \( VG(\phi^*_w, \phi^*_w, \beta^*, \theta^*) \) we obtain \( VG(\phi^*_w, \phi^*_w, \beta^*, \theta^*) \)

\[
= \mathbb{E}_{q_{\phi^*_w}(v,w,z|x,y=1)} \log \frac{q_{\phi^*_w}(v,w,z|x,y=1)}{p_{\beta^*}(v,w,z|x,y=1)} = \mathbb{E}_{q_{\phi^*_w}(v,w,z|x,y=1)} \log \frac{q_{\phi^*_w}(v,w,z|x,y=1)}{p_{\beta^*}(z|w, y = 1, v) p(v|y = 1)} = \mathbb{E}_{q_{\phi^*_w}(w|x,y=1)} q_{\phi^*_w}(w|x,y=1) \log p_{\beta^*}(z|w, y = 1, v) = 0 \tag{7}
\]

due to (1) and (7). To complete the proof, simply note that negating (4) yields

\[-\mathbb{E}_X [\mathcal{L}(K; \phi^*_w, \phi^*_w, \beta^*, \theta^*)] = -\mathbb{E}_X [\log p_{\text{data}}]. \]

\section{Proof of Proposition 2}

\textbf{Lemma 2.} For every \( \delta > 0 \) and \( \mu \), there exists \( \sigma^2 \) such that if \( f(z) \) is the pdf of a d-dimensional Normal random vector with mean \( \mu \) and diagonal covariance \( \sigma^2 \) then

\[ f(z) \leq \delta \text{ for every } z. \]

\textbf{Proof.} Let \( u = \left( \frac{1}{\delta} (2\pi)^{-d/2} \right)^{1/d} \) and \( \sigma = (u, \ldots, u) \). We have

\[
f(z) = \prod_i \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma_i^2} (z_i - \mu_i)^2 \right\} \leq \prod_i \frac{1}{\sigma_i \sqrt{2\pi}} = \delta. \]

\textbf{Proposition 2.} Let us assume \( C = 1 \), Assumptions 1 and 2 hold, and that \( p(v|y = 1) \) is uniform in the appropriate dimension. We have

\[
\min \{-\mathbb{E}_X[\mathcal{L}(K; \phi_w, \phi_w, \beta, \theta)]\} - \min \{-\mathbb{E}_X[\mathcal{L}(K + 1; \phi_w, \phi_w, \beta, \theta)]\} \geq \epsilon_K
\]

where \( -\log 2 \leq \log(K/(K + 1)) \leq \epsilon_K \) for all \( K \).

\textbf{Proof.} We show that for every solution \((\phi^*_w, \phi^*_w, \beta^*, \theta')\) to minimize \( \mathbb{E}_X[-\mathcal{L}(K; \phi_w, \phi_w, \beta, \theta)] \), there exists a corresponding solution \((\phi^*_w, \phi^*_w, \beta^*, \theta^*)\) such that

\[
-\mathbb{E}_X[\mathcal{L}(K; \phi^*_w, \phi^*_w, \beta^*, \theta')] = -\mathbb{E}_X[\mathcal{L}(K + 1; \phi^*_w, \phi^*_w, \beta^*, \theta')] + \epsilon_K.
\]

Let us assume that \((\phi^*_w, \phi^*_w, \beta^*, \theta')\) minimizes \( -\mathbb{E}_X[\mathcal{L}(K; \phi_w, \phi_w, \beta, \theta)] \). Then we can choose

\[
\phi^*_w = \phi'_w
\]

\[
\phi^*_w = \phi'_w
\]

\[
\theta^* = \theta'
\]

\[ \tag{8} \]

which is a valid choice by Assumption 2, and have \( \beta^* \) such that

\[
p_{\beta^*}(z|w, y = 1, v) = p_{\beta^*}(z|w, y = 1, v) \text{ for all } v \leq K \tag{9} \]
Inserting (9) and (10) into (6) and combined with uniform priors, we get that for all \( w, \beta \) and 2 and Lemmas 1 and 2. In essence, we choose \( \beta^* \) such that the first \( K \) subcluster generative distributions are the same as the case \( \beta^* \) but we take the \((K + 1)\)-th subcluster generative distribution to map all points \( w \) to the same Normal distribution with large enough covariance.

Inserting (9) and (10) into (6) and combined with uniform priors, we get that

\[
p_{\beta^*}(v = K + 1|z, w, y = 1) = \frac{p_{\beta^*}(v = K + 1|w, y = 1)}{\sum_{j=1}^{K} p_{\beta^*}(v = j|w, y = 1)}
\]

and

\[
p_{\beta^*}(v = k|z, w, y = 1) = \frac{p_{\beta^*}(v = k|w, y = 1)}{\sum_{j=1}^{K} p_{\beta^*}(v = j|w, y = 1)} \leq \frac{p_{\beta^*}(v = k|z, w, y = 1)}{p_{\beta^*}(v = K + 1|w, y = 1)} = \delta A(z, w, v = k).
\]

for all \( k \leq K \). The absolute difference between the two posteriors for \( k \leq K \) in (12) is bounded by a factor of \( \delta \) as follows:

\[
\left| p_{\beta^*}(v = k|z, w, y = 1) - p_{\beta^*}(v = k|z, w, y = 1) \right| \leq \frac{\delta}{\sum_{j=1}^{K} p_{\beta^*}(v = j)} \left( \frac{\sum_{j=1}^{K} p_{\beta^*}(v = j)}{\sum_{j=1}^{K} p_{\beta^*}(v = j)} \right)^2 \leq \delta A(z, w, v = k).
\]

Now we calculate \( \epsilon_K \) given by

\[
\mathbb{E}_X[-\mathcal{L}(K; \phi_z^*, \phi_w^*, \beta^*, \theta^*)] - \mathbb{E}_X[-\mathcal{L}(K + 1; \phi_z^*, \phi_w^*, \beta^*, \theta^*)] = \epsilon_K
\]

Because of (8), \( \epsilon_K \) simplifies to

\[
\epsilon_K = -\mathbb{E}_X \left[ \mathbb{E}_{q_{z|w}}(w|x, y=1)q_{z|x}(z|x) \sum_{j=1}^{K} p_{\beta^*}(v = j|z, w, y = 1) \log p_{\beta^*}(z|w, y = 1, v = j) \right] + \mathbb{E}_X \left[ \mathbb{E}_{q_{z|w}}(w|x, y=1)q_{z|x}(z|x) \sum_{j=1}^{K+1} p_{\beta^*}(v = j|z, w, y = 1) \log p_{\beta^*}(z|w, y = 1, v = j) \right] + \mathbb{E}_X \left[ \mathbb{E}_{q_{z|w}}(w|x, y=1)q_{z|x}(z|x) KL(p_{\beta^*}(v|z, w, y = 1) || p_K(v|y = 1)) \right] - \mathbb{E}_X \left[ \mathbb{E}_{q_{z|w}}(w|x, y=1) q_{z|x}(z|x) KL(p_{\beta^*}(v|z, w, y = 1) || p_{K+1}(v|y = 1)) \right] = \epsilon^{(1)}_K + \epsilon^{(2)}_K
\]

\[
\epsilon_K = \epsilon^{(1)}_K + \epsilon^{(2)}_K
\]
where $p_K(v|y=1)$ indicates that $v$ is $K$-dimensional, and $\epsilon_k^{(1)}$ are the first two terms while $\epsilon_k^{(2)}$ are the last two terms.

We first analyze $\epsilon_k^{(1)}$. For brevity, we combine the expectations and simply write $E[\cdot]$. Together with (9), (11), and (13), we get

$$
|\epsilon_k^{(1)}| = -E \left[ \sum_{j=1}^{K} p_{\beta'}(v = j|z, w, y = 1) \log p_{\beta'}(z|w, y = 1, v = j) \right]
+ E \left[ \sum_{j=1}^{K} p_{\beta'}(v = j|z, w, y = 1) \log p_{\beta'}(z|w, y = 1, v = j) \right]
+ E \left[ p_{\beta'}(v = K + 1|z, w, y = 1) \log p_{\beta'}(z|w, y = 1, v = K + 1) \right]
= E \left[ \sum_{j=1}^{K} \log p_{\beta'}(z|w, y = 1, v = j) \left( p_{\beta'}(v = j|z, w, y = 1) - p_{\beta'}(v = j|z, w, y = 1) \right) \right]
+ E \left[ p_{\beta'}(z|w, y = 1, v = K + 1) \log p_{\beta'}(z|w, y = 1, v = K + 1) \right]
\sum_{j=1}^{K} \log p_{\beta'}(z|w, y = 1, v = j) + p_{\beta'}(z|w, y = 1, v = K + 1) \leq \delta \cdot E \left[ \sum_{j=1}^{K} \log p_{\beta'}(z|w, y = 1, v = j) \right] A(z, w, v = j)
+ |\delta(\log \delta)| E \left[ \frac{1}{\sum_{j=1}^{K} p_{\beta'}(z|w, y = 1, v = j)} \right] = o(1),
$$

(14)

where the last inequality follows from $|x \log x|$ being increasing for $x \leq 1/e$ and in $o(1)$ we consider $\delta \to 0$.

Next we study $\epsilon_k^{(2)}$. For shorthand, let us define

$$
\log \left( (K + 1)p_{\beta'}(v = K + 1|z, w, y = 1) \right)
= \log \left( \frac{(K + 1)p_{\beta'}(z|w, y = 1, v = K + 1)}{\sum_{j=1}^{K} p_{\beta'}(z|w, y = 1, v = j) + p_{\beta'}(z|w, y = 1, v = K + 1)} \right)
= \log p_{\beta'}(z|w, y = 1, v = K + 1) + B(z, w)
$$

and note that

$$
|B(z, w)| = \left| \log \left( \frac{(K + 1)}{\sum_{j=1}^{K} p_{\beta'}(z|w, y = 1, v = j) + p_{\beta'}(z|w, y = 1, v = K + 1)} \right) \right|
\leq \max \left\{ \left| \log \left( \frac{(K + 1)}{\sum_{j=1}^{K} p_{\beta'}(z|w, y = 1, v = j)} \right) \right|, \left| \log \left( \frac{(K + 1)}{\sum_{j=1}^{K} p_{\beta'}(z|w, y = 1, v = j) + 1/e} \right) \right| \right\}
= C(z, w).
$$

We have

$$
\epsilon_k^{(2)}
= E \left[ \sum_{j=1}^{K} p_{\beta'}(v = j|z, w, y = 1) \log (Kp_{\beta'}(v = j|z, w, y = 1)) \right]
$$
\begin{align*}
&- \mathbb{E} \left[ \sum_{j=1}^{K} p_{\beta^*}(v = j \mid z, w, y = 1) \log \left( (K + 1)(p_{\beta^*}(v = j \mid z, w, y = 1)) \right) \right] \\
&- \mathbb{E} \left[ p_{\beta^*}(v = K + 1 \mid z, w, y = 1) \log ((K + 1)p_{\beta^*}(v = K + 1 \mid z, w, y = 1)) \right] \\
&= \mathbb{E} \left[ \sum_{j=1}^{K} (\log K)p_{\beta^*}(v = j \mid z, w, y = 1) - (\log (K + 1))p_{\beta^*}(v = j \mid z, w, y = 1) \right] \\
&+ \mathbb{E} \left[ \sum_{j=1}^{K} p_{\beta^*}(v = j \mid z, w, y = 1) \log p_{\beta^*}(v = j \mid z, w, y = 1) - p_{\beta^*}(v = j \mid z, w, y = 1) \log p_{\beta^*}(v = j \mid z, w, y = 1) \right] \\
&- \mathbb{E} \left[ p_{\beta^*}(v = K + 1 \mid z, w, y = 1) \log ((K + 1)p_{\beta^*}(v = K + 1 \mid z, w, y = 1)) \right] \\
&\geq \log(K) - (\log(K + 1))\mathbb{E} \left[ \sum_{j=1}^{K} p_{\beta^*}(v = j \mid z, w, y = 1) \right] \\
&+ \mathbb{E} \left[ \sum_{j=1}^{K} (p_{\beta^*}(v = j \mid z, w, y = 1) - p_{\beta^*}(v = j \mid z, w, y = 1)) \log(p_{\beta^*}(v = j \mid z, w, y = 1)) \right] \tag{15} \\
&- \mathbb{E} \left[ \frac{p_{\beta^*}(z \mid w, y = 1, v = K + 1) \log p_{\beta^*}(z \mid w, y = 1, v = K + 1)}{\sum_{j=1}^{K} p_{\beta^*}(z \mid w, y = 1, v = j) + p_{\beta^*}(z \mid w, y = 1, v = K + 1)} \right] B(z, w) \\
&\geq \log(K) - \log(K + 1) \\
&- \delta \cdot \mathbb{E} \left[ \sum_{j=1}^{K} A(z, w, v = j) \left| \log(p_{\beta^*}(v = j \mid z, w, y = 1)) \right| \right] \tag{16} \\
&- \delta (\log \delta) \mathbb{E} \left[ \frac{1}{\sum_{j=1}^{K} p_{\beta^*}(z \mid w, y = 1, v = j)} \right] \tag{17} \\
&- \delta \cdot \mathbb{E} \left[ \frac{1}{\sum_{j=1}^{K} p_{\beta^*}(z \mid w, y = 1, v = j)} C(z, w) \right] \\
&= \log \left( \frac{K}{K + 1} \right) + o(1). \\
\end{align*}

In (15) we use (12), in (16) we rely on (13), and in (17) we use (14) again.

To summarize, we have $\epsilon_K \geq -|\epsilon^{(1)}_K| + \epsilon^{(2)}_K \geq -o(1) + o(1) + \log \frac{K}{K + 1} = \log \frac{K}{K + 1} + o(1)$. Thus $\epsilon_K \geq \log \frac{K}{K + 1}$. \qed