Mixtures of posterior and prior variational autoencoders for representation learning and cluster analysis in latent space

Mashfiqul Huq Chowdhury[0000−0003−3438−6645], Yuichi Hirose[0000−0001−5565−3314], Stephen Marsland[0000−0002−9568−848X], and Yuan Yao[0009−0002−2381−0033]

Victoria University of Wellington, Wellington, New Zealand

1 Estimation of $q(y_k)$: MVAE Model

Finding $q(y_k)$ involves a constraint optimization problem. Under the constraint, $q(y_k) \geq 0$ and $\sum\limits_{k=1}^{K} q(y_k) = 1$, we need to solve the following Lagrangian function:

$$\frac{1}{L}\sum_{k=1}^{K} q(y_k) \sum_{l=1}^{L} \log\left[\frac{p_\theta(\mathbf{x}|\mathbf{z}_k^{(l)})p(y_k)p(\mathbf{z}_k^{(l)}|y_k)}{q_\phi(\mathbf{z}_k^{(l)}|y_k,\mathbf{x})q(y_k)}\right] + \lambda\left[1 - \sum_{k=1}^{K} q(y_k)\right], \tag{1}$$

where λ is a Lagrange multiplier. Now, differentiating (1) w.r.t. $q(y_k)$ and setting equal to 0 gives:

$$\sum_{l=1}^{L} \log\left[\frac{p_\theta(\mathbf{x}|\mathbf{z}_k^{(l)})p(y_k)p(\mathbf{z}_k^{(l)}|y_k)}{q_\phi(\mathbf{z}_k^{(l)}|y_k,\mathbf{x})q(y_k)}\right] = \lambda L + L. \tag{2}$$

Exponentiating both sides gives

$$\prod_{l=1}^{L} \left[\frac{p_\theta(\mathbf{x}|\mathbf{z}_k^{(l)})p(y_k)p(\mathbf{z}_k^{(l)}|y_k)}{q_\phi(\mathbf{z}_k^{(l)}|y_k,\mathbf{x})q(y_k)}\right] = \exp(L + \lambda L). \tag{3}$$

Multiplying by $q(y_k)$ on both sides and summing over $k = 1$ to K with the constraint $\sum\limits_{k=1}^{K} q(y_k) = 1$ gives:

$$\sum_{k=1}^{K} \prod_{l=1}^{L} \left[\frac{p_\theta(\mathbf{x}|\mathbf{z}_k^{(l)})p(y_k)p(\mathbf{z}_k^{(l)}|y_k)}{q_\phi(\mathbf{z}_k^{(l)}|y_k,\mathbf{x})}\right] = \exp(L + \lambda L). \tag{4}$$

By substituting (4) into (3), we obtain an analytical form of $q(y_k)$:

$$\hat{q}(y_k) = \frac{\prod\limits_{l=1}^{L} \left[\frac{p_\theta(\mathbf{x}|\mathbf{z}_k^{(l)})p(y_k)p(\mathbf{z}_k^{(l)}|y_k)}{q_\phi(\mathbf{z}_k^{(l)}|y_k,\mathbf{x})}\right]}{\sum\limits_{j=1}^{K} \prod\limits_{l=1}^{L} \left[\frac{p_\theta(\mathbf{x}|\mathbf{z}_j^{(l)})p(y_j)p(\mathbf{z}_j^{(l)}|y_j)}{q_\phi(\mathbf{z}_j^{(l)}|y_j,\mathbf{x})}\right]}. \tag{5}$$

2 Estimation of $q(y_k)$: MVAE(EM-V1) Model)

Finding $q(y_k)$ involves a constrained optimization problem. Under the constraints: $q(y_k) \geq 0$ and $\sum\limits_{k=1}^{K} q(y_k) = 1$, we are going to solve the following Lagrangian function:

$$\frac{1}{L}\sum_{k=1}^{K} q(y_k) \sum_{l=1}^{L} \log\left[\frac{p(\mathbf{z}_k^{(l)}|y_k)\ p(y_k)}{q(y_k)}\right] \quad + \quad \lambda\Bigg[1 \quad - \quad \sum_{k=1}^{K} q(y_k)\Bigg], \tag{6}$$

where λ is a Lagrange multiplier. Now, differentiating (6) w.r.t. $q(y_k)$ and setting the derivative equal to 0 gives:

$$\frac{1}{L}\sum_{l=1}^{L} \log\left[\frac{p(\mathbf{z}_k^{(l)}|y_k)\ p(y_k)}{q(y_k)}\right] = 1 + \lambda. \tag{7}$$

Multiplying both sides by L and exponentiating both sides gives:

$$\prod_{l=1}^{L} \frac{p(\mathbf{z}_k^{(l)}|y_k)\ p(y_k)}{q(y_k)} = \exp(L + \lambda L). \tag{8}$$

Multiplying by $q(y_k)$ on both sides and summing over $k = 1$ to K in (8), then using the constraint $\sum\limits_{k=1}^{K} q(y_k) = 1$, we get:

$$\sum_{k=1}^{K} \prod_{l=1}^{L} p(\mathbf{z}_k^{(l)}|y_k)\ p(y_k) = \exp(L + \lambda L). \tag{9}$$

Substituting (9) into (8) and solving for $q(y_k)$, we obtain an analytical form of $q(y_k)$:

$$\hat{q}(y_k) = \frac{\prod\limits_{l=1}^{L} p(y_k)\ p(\mathbf{z}_k^{(l)}|y_k)}{\sum\limits_{j=1}^{K} \prod\limits_{l=1}^{L} p(y_j)\ p(\mathbf{z}_j^{(l)}|y_j)}. \tag{10}$$

3 Estimation of $q(y_k)$: MVAE(EM-V2) Model)

We need to solve the following Lagrangian function with the Lagrange multiplier λ:

$$\frac{1}{L}\sum_{k=1}^{K} q(y_k) \sum_{l=1}^{L} \log\left[\frac{q_{\boldsymbol{\phi}}(\mathbf{z}_k^{(l)}|\mathbf{x}, y_k)\ q(y_k)}{p(\mathbf{z}_k^{(l)}|y_k)\ p(y_k)}\right] + \lambda\Bigg[1 - \sum_{k=1}^{K} q(y_k)\Bigg]. \tag{11}$$

Differentiating (11) with respect to $q(y_k)$ and setting the derivative equal to 0 gives:

$$\frac{1}{L}\sum_{l=1}^{L}\log\left[\frac{p(\mathbf{z}_k^{(l)}|y_k)\ p(y_k)}{q_{\boldsymbol{\phi}}(\mathbf{z}_k^{(l)}|\mathbf{x},\ y_k)\ q(y_k)}\right]=1-\lambda. \tag{12}$$

Multiplying both sides by L and exponentiating both sides gives:

$$\prod_{l=1}^{L}\left[\frac{p(\mathbf{z}_k^{(l)}|y_k)\ p(y_k)}{q_{\boldsymbol{\phi}}(\mathbf{z}_k^{(l)}|\mathbf{x},y_k)\ q(y_k)}\right]=\exp(L-\lambda L). \tag{13}$$

Multiplying $q(y_k)$ on both sides and summing over $k=1$ to K with the constraint $\sum_{k=1}^{K}q(y_k)=1$ to get:

$$\sum_{k=1}^{K}\prod_{l=1}^{L}\left[\frac{p(\mathbf{z}_k^{(l)}|y_k)\ p(y_k)}{q_{\boldsymbol{\phi}}(\mathbf{z}_k^{(l)}|\mathbf{x},\ y_k)}\right]=\exp(L-\lambda L). \tag{14}$$

Substituting (14) into (13), we obtain a closed-form expression of the $q(y_k)$:

$$\hat{q}(y_k)=\frac{\prod_{l=1}^{L}\left[\frac{p(y_k)\ p(\mathbf{z}_k^{(l)}|y_k)}{q_{\boldsymbol{\phi}}(\mathbf{z}_k^{(l)}|\mathbf{x},\ y_k)}\right]}{\sum_{j=1}^{K}\prod_{l=1}^{L}\left[\frac{p(y_j)\ p(\mathbf{z}_j^{(l)}|y_j)}{q_{\boldsymbol{\phi}}(\mathbf{z}_j^{(l)}|\mathbf{x},\ y_j)}\right]}. \tag{15}$$

4 Clustering Procedures

We train each evaluated model-VAE [3] , VADE [2], k-DVAE [1], and MVAE-with five different random initializations. The experimental setup is detailed in the main article.

4.1 Comparator Algorithms

We train the VAE, VADE, and k-DVAE algorithms, then use the trained model to extract the latent representations from the test set of each dataset. Next, we apply the Gaussian mixture model (GMM) to the learned latent embeddings for clustering and report the highest unsupervised clustering accuracy achieved by each model across datasets, based on results from five experiments.

4.2 Proposed Algorithms

We train our proposed model using both soft and hard optimization approaches. The trained model is then applied to the test set of each benchmark dataset, generating K latent representations for each observation. We select the latent representation based on the MAP estimate of the clustering assignment probabilities using Equations 5, 10, 15. Next, we apply the GMM over the learned latent embeddings for clustering and report the highest unsupervised clustering accuracy achieved by the proposed model across datasets, based on results from five experiments.

The next section reports the optimal latent dimension (M) of each evaluated model.

5 Latent Dimensions

5.1 VAE

We trained the VAE model [3] and then applied the GMM clustering method over the latent embeddings of the test dataset. We performed hyper-parameter tuning for the latent dimension (M) of the VAE model. We achieved the optimal clustering performance with the latent dimension: $M = 5$ for the Fashion-MNIST, Digits, HAR, and USPS datasets; and $M = 10$ for the MNIST, STL-10, and Reuters datasets.

5.2 VADE Model

The VADE algorithm [2] uses the following equation to calculate clustering assignment probabilities:

$$\hat{q}(y_k) = \frac{p(y_k)\ p(\mathbf{z}|y_k)}{\sum\limits_{j=1}^{K} p(y_j)\ p(\mathbf{z}|y_j)}. \tag{16}$$

Clustering assignment is predicted by using:

$$\hat{y} = \arg\max_{k} \hat{q}(y_k). \tag{17}$$

We applied both the posterior assignment estimate in (16) and the GMM for clustering, achieving the best performance using GMM. The optimal performance is obtained with $M = 5$ for the Fashion MNIST, Digits, HAR, and STL-10 datasets and with $M = 10$ for the MNIST, USPS, and Reuters datasets.

5.3 *k*-DVAE Model

We obtained the best clustering performance with $M = 5$ for the HAR, USPS, Reuters, Fashion-MNIST, and Digits datasets and $M = 10$ for the MNIST and STL-10 datasets.

5.4 MVAE Model

The MVAE model can be trained using soft and hard optimization approaches. We achieved the best clustering performance with $M = 5$ for the HAR, USPS, and STL-10 datasets; with $M = 7$ for the Fashion-MNIST dataset; and with $M = 10$ for the MNIST, Digits, and Reuters datasets under the soft optimization method. For the hard optimization approach, we obtained good clustering performance using $M = 5$ for HAR, Reuters, and STL-10 datasets; $M = 7$ for the Fashion-MNIST dataset; and $M = 10$ for the MNIST, Digits, and USPS datasets.

5.5 MVAE (EM-V1) Model

The proposed MVAE(EM-V1) model is trained with both soft and hard optimization approaches. We achieved the best clustering accuracy with the latent dimension: $M = 5$ for the Fashion-MNIST, STL-10, USPS, and Reuters datasets; $M = 7$ for the HAR dataset; and $M = 8$ for the Digits, and MNIST datasets under the soft optimization approach.

Again for the hard optimization algorithm, we obtained the best clustering performance using latent dimension: $M = 5$ for the STL-10, Reuters, and USPS datasets; $M = 6$ for the Digits and Fashion-MNIST datasets; $M = 7$ for the HAR dataset; and $M = 10$ for the MNIST dataset.

5.6 MVAE (EM-V2) Model

The proposed MVAE(EM-V2) model is trained using both soft and hard optimization approaches. We achieved the highest unsupervised clustering accuracy with the latent dimension: $M = 5$ for the USPS dataset; and $M = 6$ Fashion-MNIST dataset; $M = 7$ for the MNIST and Reuters datasets; and $M = 10$ for the Digits, HAR, and STL-10 datasets under the soft optimization method.

Again for the hard optimization algorithm, we obtained the best clustering accuracy with the latent dimension: $M = 5$ for the STL-10, Fashion-MNIST, Reuters, and USPS datasets; and $M = 10$ for the MNIST, Digits, and HAR datasets.

5.7 β-MVAE(EM-V2) Model

The β-MVAE(EM-V2) model is also trained using both soft and hard optimization approaches. We achieved the highest unsupervised clustering accuracy with the latent dimension: $M = 5$ for the USPS and Reuters datasets; and $M = 10$ for the Digits, MNIST, STL-10, Fashion-MNIST, and HAR datasets under the soft optimization method.

We obtained the highest accuracy for the hard optimization algorithm with the latent dimension: $M = 5$ for the Digits, HAR, Fashion-MNIST, Reuters, and USPS datasets; and $M = 10$ for the MNIST and STL-10 datasets.

5.8 Scheduled β-MVAE(EM-V2) Model

We achieved the highest unsupervised clustering accuracy with the latent dimension: $M = 5$ for the USPS and Reuters datasets; $M = 7$ for the Digits dataset; and $M = 10$ for the MNIST, STL-10, Fashion-MNIST, and HAR datasets under the soft optimization method of this model.

Again, for the hard optimization algorithm, we obtained the best clustering accuracy using latent dimension: $M = 5$ for the Reuters and USPS datasets; $M = 6$ for the Digits and HAR datasets; $M = 7$ for the Fashion-MNIST and STL-10 datasets; and $M = 10$ for the MNIST dataset.

Table 1: Latent dimensions of the evaluated and proposed models.

Datasets	**Models**														
	VAE	**VADE**	***k*-DVAE**	**MVAE**		**MVAE(EM-V1)**		**MVAE(EM-V2)**		**β-MVAE(EM-V2)**		**Scheduled β-MVAE(EM-V2)**	
				Soft	Hard	Soft	Hard	Soft	Hard	Soft	Hard	Soft	Hard
Digits	5	5	5	10	10	8	6	10	10	10	5	7	6
MNIST	10	10	10	10	10	8	10	7	10	10	10	10	10
USPS	5	10	5	5	10	5	5	5	5	5	5	5	5
STL-10	10	5	10	5	5	5	5	10	5	10	10	10	7
Fashion-MNIST	5	5	5	7	7	5	6	6	5	10	5	10	7
Reuters	10	10	5	10	5	5	5	7	5	5	5	5	5
HAR	5	5	5	5	5	7	7	10	10	10	5	10	6

6 Computational Time of the Benchmark and Proposed Models

Table 2: Execution time (seconds) for 10 training sessions of benchmark and proposed algorithms

Datasets	Benchmark Models						Proposed Models						
							MVAE		MVAE(EM-V1)		MVAE(EM-V2)		No. of parameters
	VAE	**No. of parameters**	**VADE**	**No. of parameters**	***k*-DVAE**	**No. of parameters**	Soft	Hard	Soft	Hard	Soft	Hard	
Digits	2.57	172,628	3.60	172,838	3.83	218,888	6.28	6.87	5.01	4.69	4.98	4.65	219,098
MNIST	60.63	3,350,804	98.08	3,351,014	115.37	3,710,984	222.34	224.09	132.34	122.35	128.51	120.16	3,711,194
USPS	10.56	2,822,276	12.05	2,822,486	13.54	3,182,456	26.61	26.81	15.71	14.91	15.38	14.47	3,182,666
STL-10	9.32	4,616,068	10.56	4,616,278	9.27	4,976,248	24.53	24.44	10.94	10.56	10.85	9.88	4,976,458
Fashion-MNIST	73.81	3,350,804	81.92	3,351,014	114.32	3,710,984	228.33	227.93	133.94	121.09	131.19	120.36	3,711,194
Reuters	9.56	4,568,020	12.63	4,568,104	10.06	4,688,080	16.81	17.08	13.56	11.50	12.84	11.44	4,688,164
HAR	7.65	3,127,581	10.41	3,127,707	9.50	3,327,681	15.60	16.12	10.73	9.52	10.46	8.66	3,327,807

7 Impact of Regularization on the Mixtures VAE Model

Here, we present the impact of the regularization coefficient (β) on the proposed regularized mixtures VAE or the β-MVAE(EM (V-2)) model, across seven benchmark datasets. We train the model on the training set of these datasets. Then, we perform clustering on the latent embeddings of the test set of each dataset. The best clustering performance for each $\beta \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ in terms of unsupervised cluster accuracy is reported after five runs.

Table 3: Impact of β on cluster analysis performance measures (%) for the Digits test dataset.

β	**Measures (%)**											
	Soft Optimization						**Hard Optimization**					
	GMM			Posterior Probability			GMM			Posterior Probability		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
0.1	80.37	76.45	64.36	29.26	43.59	13.08	**81.48**	77.93	65.23	44.07	45.34	24.71
0.2	80.37	73.17	62.90	37.04	47.75	18.90	80.00	72.77	61.79	58.52	60.72	40.59
0.3	**84.44**	78.15	69.92	54.44	66.66	44.86	74.81	68.82	56.78	45.18	48.65	24.55
0.4	80.00	71.73	62.44	62.22	66.13	49.85	71.85	64.79	51.74	52.22	54.93	35.75
0.5	73.70	66.91	54.06	70.37	67.29	54.03	68.15	63.48	48.95	55.56	56.97	38.57

Table 4: Impact of β on cluster analysis performance measures (%) for the MNIST test dataset.

β	**Measures (%)**											
	Soft Optimization						**Hard Optimization**					
	GMM			Posterior Probability			GMM			Posterior Probability		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
0.1	96.20	90.77	91.81	61.34	70.93	56.19	94.07	87.88	87.28	35.60	56.87	27.53
0.2	**96.42**	91.20	92.26	50.25	69.96	42.29	**95.49**	89.82	90.35	64.69	72.74	52.52
0.3	95.48	89.57	90.32	31.31	51.77	22.23	94.90	88.85	89.03	54.97	63.10	34.81
0.4	95.60	89.38	90.52	31.21	52.01	22.45	95.12	89.08	89.54	55.24	67.64	42.70
0.5	93.08	86.70	85.32	39.64	60.63	29.94	93.37	87.02	86.19	55.78	68.96	44.73

Table 5: Impact of β on cluster analysis performance measures (%) for the USPS test dataset.

β	**Measures (%)**											
	Soft Optimization						**Hard Optimization**					
	GMM			Posterior Probability			GMM			Posterior Probability		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
0.1	82.66	72.27	68.57	58.25	53.17	43.64	**87.84**	77.61	77.73	38.46	42.91	23.62
0.2	77.38	71.23	65.24	53.41	56.40	41.81	83.41	73.85	71.82	42.95	44.29	28.42
0.3	**85.80**	75.90	75.67	39.21	45.84	23.96	79.27	69.43	62.55	52.37	55.67	40.26
0.4	73.74	70.33	65.58	54.96	64.87	50.68	76.68	72.31	68.62	56.20	63.42	51.60
0.5	75.83	71.88	62.03	46.69	55.29	30.72	75.49	70.48	58.48	54.41	60.64	46.70

Table 6: Impact of β on cluster analysis performance measures (%) for the STL-10 test dataset.

β	**Measures (%)**											
	Soft Optimization						**Hard Optimization**					
	GMM			Posterior Probability			GMM			Posterior Probability		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
0.1	**95.06**	89.77	89.41	29.44	53.37	22.23	**93.50**	87.46	86.09	35.74	53.27	25.56
0.2	93.30	87.55	85.87	20.00	44.32	18.42	92.26	86.12	84.04	50.20	63.82	37.03
0.3	93.84	87.67	87.01	20.00	44.58	18.45	91.32	84.77	82.21	36.40	59.57	30.89
0.4	92.82	86.87	84.95	29.82	52.76	22.11	89.04	83.39	78.86	49.38	68.54	42.46
0.5	92.56	85.47	84.36	20.00	44.41	18.42	92.32	84.92	83.81	58.16	70.70	51.95

Table 7: Impact of β on cluster analysis performance measures (%) for the Fashion MNIST test dataset.

β	**Measures (%)**											
	Soft Optimization						**Hard Optimization**					
	GMM			Posterior Probability			GMM			Posterior Probability		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
0.1	**67.15**	65.74	54.44	47.69	53.16	33.19	64.15	64.16	51.13	53.11	59.44	41.61
0.2	63.92	64.84	49.62	32.71	45.33	16.29	**67.57**	64.95	52.67	33.66	46.29	22.71
0.3	62.11	62.95	48.07	54.21	57.85	37.35	62.61	64.78	49.07	38.34	51.44	28.87
0.4	62.08	64.15	48.32	48.43	52.95	35.01	62.43	62.23	47.81	36.20	51.72	31.06
0.5	61.58	63.68	48.07	58.44	60.25	44.87	64.35	65.62	51.79	48.14	56.50	36.03

Table 8: Impact of β on cluster analysis performance measures (%) for the Reuters test dataset.

β	**Measures (%)**											
	Soft Optimization						**Hard Optimization**					
	GMM			Posterior Probability			GMM			Posterior Probability		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
0.1	75.90	48.49	54.34	55.90	22.37	20.49	78.40	56.99	57.64	51.80	14.13	6.70
0.2	76.40	49.47	56.14	46.90	22.66	16.14	74.40	41.35	43.64	64.60	27.56	28.50
0.3	77.80	48.52	54.38	52.70	27.71	21.91	70.40	44.95	50.42	62.20	32.84	38.38
0.4	73.40	48.57	52.43	53.20	33.17	24.69	**78.50**	49.48	56.44	58.10	32.07	23.41
0.5	**82.70**	60.44	67.51	60.90	37.07	26.03	71.90	38.57	42.88	49.90	24.10	10.17

Table 9: Impact of β on cluster analysis performance measures (%) for the HAR test dataset.

β	**Measures (%)**											
	Soft Optimization						**Hard Optimization**					
	GMM			Posterior Probability			GMM			Posterior Probability		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
0.1	77.09	70.50	62.35	35.05	55.41	33.24	**80.49**	75.42	66.97	37.12	52.09	32.04
0.2	73.74	68.76	57.90	35.05	55.41	33.24	66.68	63.61	51.78	36.51	53.02	32.44
0.3	73.23	74.89	64.91	49.34	64.97	43.10	65.56	62.81	50.80	36.65	52.63	32.06
0.4	68.92	70.57	57.64	60.77	70.65	51.78	62.10	68.18	55.18	53.75	68.39	48.59
0.5	**79.64**	70.76	63.06	52.56	63.48	41.52	59.79	67.15	53.81	54.39	70.08	49.72

References

1. Caciularu, A., and Goldberger, J. An entangled mixture of variational autoencoders approach to deep clustering. *Neurocomputing 529* (2023), 182–189.
2. Jiang, Z., Zheng, Y., Tan, H., Tang, B., and Zhou, H. Variational deep embedding: A generative approach to clustering. *CoRR, abs/1611.05148* (2016).
3. Kingma, D. P., and Welling, M. Auto-encoding variational Bayes. *International Conference on Learning Representations* (2014).