The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

VAEを用いた受注の年間推移に基づくクラスタリングの評価

Evaluation of Clustering Based on Annual Trends of Receiving Orders with VAE

三井 康行
Yasuyuki Mitsui

アスクル株式会社
Askul Corporation

For E-Commerce(EC) business, highly accurate demand-forecasting is important. However, the trends of receiving orders of items vary by the attributes of them, so it is necessary to forecast by the suitable models for each trends. The purpose of this paper is to cluster the items by the suitable representation for trends of receiving orders for realizing highly accurate demand-forecasting. Using real data in EC business, we evaluate the clustering based on the annual trends of receiving orders. We employ dimensional reduction by Variational Autoencoder(VAE) and clustering in latent space by Gaussian Mixture Model(GMM). And we introduce degree of forecasting difficulty as an new index. By our experiments, we confirm that the result of clustering is valid with degree of forecasting difficulty.

1. はじめに

近年，深層学習を始めとする機械学習手法の発展により，精度評価のための実験を目的としたデータセットではなく，現実の業務等において収集された大規模データを高精度に分析することが可能になりつつある．これにより，従来人手で行ってきた作業を自動化することによる業務効率の改善が進められている．しかし，様々な事業体が持つデータは日々蓄積されて大規模化しているものの，どのような種類のデータであっても，正解データがラベル付けされているものはごく少数である．そのため，教師なし学習，あるいは半教師あり学習により大規模データを分析し，新たな価値や指標に対する仮説を立案し，検証していく必要がある．

我々は企業向けおよび消費者向けの EC 事業者として，多種の商品を取り扱っており，その受注情報が日々蓄積されている．過去の受注数量情報および関連情報から，各商品が未来に受注されるであろう数量を予測するタスク，いわゆる需要予測は，在庫の最適化や欠品率の低減を実現するために，EC が隆盛する以前から小売業者にとって必須の課題であり，従来から盛んに研究されてきた．しかし，商品需要傾向は商品の属性によって大きく異なる．年間を通じて常に一定量の需要がある商品であれば，移動平均や指数平滑化を用いたシンプルな手法でも比較的高精度に予測が可能である．一方，年間を通じて安定した需要がなく，特定のトリガーによって需要がスパイクする（突発的な需要がある）商品も一定数存在し，これら商品は同一の手法では予測できない．また，需要傾向によって商品を明確に分類することは困難でありながら重要であり，適切な分類軸で商品を分類した上で,それぞれの分類ごとに需要を予測する必要がある．さらに，それぞれの商品がどのような属性を持つかを判断し，分類する作業を人手によって実施することは，数万～数十万の商品を扱う EC では現実的でない．

そこで，本稿では，商品の年間受注推移データに基づいたクラスタリングを行い，受注傾向に応じた商品の分類を試みる．また，予測困難度という概念を導入し，クラスタ毎に需要予測の困難性を数値化する．本タスクは，正解となるラベルが存在しない教師なし学習であるため，クラスタリング手法として，正解ラベルを必要としない Variational Autoencoder（VAE）[Kingma 14a]により学習したモデルのエンコーダ部を用いて次元圧縮された特徴量ベクトルに対し，GMM を用いたクラスタリングをするという手法を採用する．実際の受注データを用い，本手法により受注傾向による分類が適切に行われているかを確認する．

連絡先：三井 康行，アスクル株式会社，東京都江東区豊洲 3-2-3，
yasuyuki.mitsui@askul.com

2. 関連研究

深層学習を用いた教師なしクラスタリング手法として，いくつかの手法が提案されている[Aljalbout 18]．特に，VAE に代表される生成モデル(generative model)を用いたクラスタリング手法は，比較的シンプルなネットワークでも高精度な生成器および識別器を学習でき，観測データの特徴をよく表現した潜在空間内で低次元のクラスタリングができる手法として，近年注目されている[Dilokthanakul 16]．VAE は，変分ベイズ法に対して，ニューラルネットワークの一種である自己符号器（Autoencoder，AE）を適用して，潜在空間の高い表現性と確率的な生成モデルの学習を可能にした手法である．主に画像データに対して適用され，画像の分類や類似の特徴を持つ別の画像を生成するタスクで活用されている．また，テキストの分類やマルチモーダルなデータの生成や分類にも利用され始めている[鈴木 16]．

3. 提案手法

提案手法によるクラスタリングの方法および予測困難度の定義について説明する．

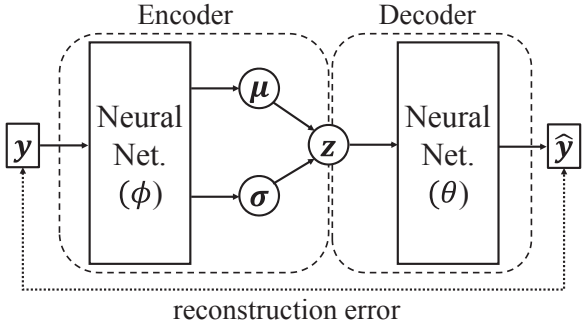

図 1 VAE の概略図

3.1 VAE による次元圧縮

VAE の概略図を図 1 に示す. VAE では, 式(1)に示す目的関数 L を用いて, 変分パラメータ$\boldsymbol{\phi}$と生成パラメータ$\boldsymbol{\theta}$を最適化する.

$$\boldsymbol{L}(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{y}^{(i)})$$

$$= -\boldsymbol{D_{KL}}\left(\boldsymbol{q_\phi}(\boldsymbol{z}|\boldsymbol{y}^{(i)})\|\boldsymbol{p_\theta}(\boldsymbol{z})\right) + \boldsymbol{E}_{\boldsymbol{q_\phi}(\boldsymbol{z}|\boldsymbol{y}^{(i)})}\left[\log \boldsymbol{p_\theta}(\boldsymbol{y}^{(i)}|\boldsymbol{z})\right]$$

$$\simeq \frac{\boldsymbol{1}}{\boldsymbol{2}}\sum_{\boldsymbol{j=1}}^{\boldsymbol{J}}\left(\boldsymbol{1}+\log\left(\left(\boldsymbol{\sigma}_{\boldsymbol{j}}^{(i)}\right)^{\boldsymbol{2}}\right)-\left(\boldsymbol{\mu}_{\boldsymbol{j}}^{(i)}\right)^{\boldsymbol{2}}-\left(\boldsymbol{\sigma}_{\boldsymbol{j}}^{(i)}\right)^{\boldsymbol{2}}\right) + \frac{\boldsymbol{1}}{\boldsymbol{L}}\sum_{\boldsymbol{l=1}}^{\boldsymbol{L}}\log \boldsymbol{p_\theta}(\boldsymbol{y}^{(i)}|\boldsymbol{z}^{(i,l)})$$

$$\text{where}\quad \boldsymbol{z}^{(i,l)} = \boldsymbol{\mu}^{(i)} + \boldsymbol{\sigma}^{(i)} \odot \boldsymbol{\epsilon}^{(l)}, \boldsymbol{\epsilon}^{(l)} \sim \boldsymbol{\mathcal{N}}(\boldsymbol{0},\boldsymbol{I}) \qquad (1)$$

ここで, J は潜在空間の次元数, L はモンテカルロサンプリングのサンプル数(通常は$L = 1$[Kingma 14a])である. 式(1)の第 1 項は Kullback -Leibler 距離による正則化項であり, 第 2 項はデコーダにより再構築されるであろうと推測される観測ベクトル$\hat{\boldsymbol{y}}$と, 元の観測ベクトル$\boldsymbol{y}$との差を確率的に表現した reconstruction error と呼ばれる負値の項である.

3.2 クラスタリング

3.1 の方法で次元圧縮されたベクトルに対し, GMM を用いてクラスタリングを行う. 本稿のタスクにおけるクラスタリングは, 正解となるラベルが存在しない, 純粋な教師なしクラスタリング問題である. そのため, VAE の潜在空間におけるクラスタ分割数, すなわち, GMM の混合数を事前に決定する必要がある. クラスタ分割数は, ベイズ情報量基準(*BIC*)および式(2)に示す GMM 平均ベクトルと各クラスタに属する商品の潜在空間ベクトルとの距離の自乗の総和(*SSD*)に基づいて決定する.

$$\boldsymbol{SSD} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{M}}\left\|\boldsymbol{z}_{\boldsymbol{k}}^{(i)} - \boldsymbol{\mu}_{\boldsymbol{k}}\right\|^{\boldsymbol{2}} \qquad (2)$$

ここで, K はクラスタ分割数(Gaussian の混合数), M は各クラスタに属するサンプル数, μ_kは k 番目の Gaussian の平均ベクトルである.

3.3 予測困難度

本稿では, 予測困難度を, 学習データ群を特徴付ける特定の傾向をモデル化することが困難であり, その困難性によって予測の精度が低くなる程度を示す数値であると定義する. 需要予測の文脈では, 過去の受注実績データから未来の需要を予測することが困難である商品群が存在し, これら商品群については予測困難度が高くなり, 逆に予測困難度が低いほど高い需要予測精度を期待できるということになる.

予測困難度に類する概念として, 学習したモデルあるいはテストデータにおける予測結果の信頼度が挙げられ, 学習結果の信頼性を測る指標として一般的に用いられている. にもかかわらず, 本稿で新たに予測困難度を定義した意図は, 観測されたデータによっては, 学習に用いるサンプル数やモデルの性能によらず予測が困難であるものが存在し, これらが持つ予測困難性の程度を定量評価できる枠組みを創ることにある. つまり, 信頼度が予測結果の信頼性を示す指標であるのに対し, 困難度は観測データの性質上, 予測することが困難である程度を示す指標であり, 困難度が高い場合は自ずと信頼度が低くなる.

我々は, VAE における reconstruction error を予測困難度として用いることを提案する. reconstruction error は, 前述の通り式(1)の第 2 項に相当するが, VAE の学習が収束した後であれば, より直接的に算出が可能である. すなわち, VAE で学習されたエンコーダにより入力ベクトル$\boldsymbol{y}_i$を潜在空間に写像し, 写像されたベクトル$\boldsymbol{z}_i$に対して, 同じく VAE で学習されたデコーダにより復号された出力ベクトル$\hat{\boldsymbol{y}}_i$と, 元の入力ベクトルとの差分によって表される. 本稿ではこの値を便宜上, direct-calculated reconstruction error(*DRE*)と呼び, 特定のサンプル群(ここではクラスタ)ごとに式(3)で示した式により正値で表現する.

$$\boldsymbol{DRE} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{M}}\left(\frac{\boldsymbol{1}}{\boldsymbol{n_i}}\sum_{\boldsymbol{j=1}}^{\boldsymbol{D}}\left(\boldsymbol{y}_{\boldsymbol{j}}^{(i)} - \hat{\boldsymbol{y}}_{\boldsymbol{j}}^{(i)}\right)^{\boldsymbol{2}}\right)$$

$$\text{where}\quad \boldsymbol{n_i} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{D}}\boldsymbol{s_j}, \boldsymbol{s_i} = \begin{cases}\boldsymbol{1} & (\boldsymbol{y_i} > \boldsymbol{0})\\ \boldsymbol{0} & (\boldsymbol{y_i} = \boldsymbol{0})\end{cases} \qquad (3)$$

ここで, D は観測ベクトルの次元数, $y_j^{(i)}$および$\hat{y}_j^{(i)}$は, i 番目の観測ベクトルにおける j 番目の要素, および対応する再構築ベクトルにおけるj 番目の要素であり, 後述する実験においてはそれぞれの商品における日ごとの受注量に相当する.

4. 実験

提案手法を用いて, 実際の商品受注データを用いたクラスタリングおよび予測困難度の評価実験を行った. 実験内容について説明する.

4.1 実験データ

実験に用いるデータとして, ある年の 1 年間を対象期間とし, 国内の配送センターから出荷された商品の年間受注推移データを用いた. 当該配送センターは, 主に企業向けの商品を出荷しており, 扱った受注データは企業向けの商品が多くを占めている.

実験用データのサンプル数(商品数)は約 40,000, それぞれについて日ごとの受注実績を 365 次元の特徴量ベクトルとして保持している. 商品によって受注量が大きく異なるため, 各サンプルにおける受注量として, 受注量の最大値で除算することによって正規化した値を用いている. なお, 学習用データは対象期間に受注実績のあった商品のみを対象としており, 当該配送センターに在庫しているが対象期間に受注の無かった商品については対象外としている. また, 受注量だけでなく, 商品によって受注頻度も大きく異なり, 毎日受注する商品もあれば, 1 年を通じて数日しか受注のない商品も存在する.

4.2 VAE のパラメータ

3.1 に記した手法を用いて次元圧縮を行う. VAE としてエンコーダ, デコーダにそれぞれ 3 層からなるネットワークを用い, 隠れ層の次元数は 256, 128, 64(デコーダはこの逆順), 潜在空間の次元数は 2 とした. 各隠れ層の活性化関数は ReLU, 潜在空間の活性化関数はシグモイドとし, 最適化手法として Adam[Kingma 14b]を用いた.

4.3 クラスタリング

続いて, 3.2 に記した方法により, 4.2 で生成した VAE モデルのエンコーダによって潜在空間に写像したベクトルに対してクラ

スタリングを行う. 3.2 で記した通り, クラスタ分割数を決定するために, クラスタ分割数を2から50まで変化させて *BIC* および *SSD* が取る値の変化を調べた. 横軸にクラスタ分割数, 縦軸に *BIC* および *SSD* を取って変化をプロットした図を, 図 2 に示す. 図 2 により, 妥当なクラスタ分割数は 10 であると判断した. GMM によるクラスタリングでは, k-means 法を用いて初期値を決定し, その後 EM アルゴリズムにより収束させた.

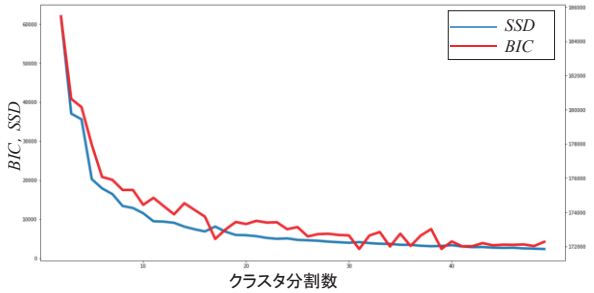

図 2: クラスタ分割数と *BIC* と *SSD* の関係

4.4 実験結果

図 3 に, 学習に使用したサンプルを 2 次元の潜在空間に写像した結果の分布図を示す. また, 受注量および受注頻度が高いサンプルを例にとって, 学習された VAE のエンコーダ／デコーダによって再構築されたデータおよび観測データを図 4 に示す. 緑線が観測データ, 赤線が再構築されたデータである. 図 4 を見ると, 再構築されたデータが元の観測データの形状(傾向)をよく再現していることが分かり, 潜在空間への写像が正しく行われているといえる. なお, 図 4 で示したサンプルが属するクラスタは 2 であり, *DRE* 値は 0.0027 である.

図 5 に, 提案手法を用いて潜在空間に写像された学習用サンプルをクラスタリングした結果を示す. 図中の黒丸点は, 各クラスタを表す Gaussian の平均ベクトルを示している.

図 6 に, 各クラスタに属するサンプルの観測データ例を示す. 観測データ例として, 各クラスタを表す Gaussian の平均ベクトルとの距離が小さい順にそれぞれ 5 サンプルを抽出している. 図 6 を見ると, 各クラスタごとに観測データの傾向が異なっており, 正しくクラスタリングできていることが分かる.

5. 考察

本実験結果について, 前述の予測困難度の考え方を導入してクラスタリングの妥当性を考察する.

各クラスタに属するサンプルの受注頻度の平均値と, 各クラスタにおける予測困難度(*DRE*)との関係を図 7 にプロットする. 横軸が受注頻度, 縦軸が *DRE* を示している. 図 7 によると, 平均受注頻度が高いクラスタほど, *DRE* が低い傾向にあり, 明確な相関があることが分かる. サンプル単位で考えた場合, 受注頻度が低い商品について, 予測が困難であろうことは経験的に納得できる. しかし, クラスタリングの際に *DRE* の因子を明示的に考慮していないため, クラスタ単位で考えた場合, 受注頻度と *DRE* との相関に蓋然性はない. したがって, 予測困難度の観点から, 本実験におけるクラスタリングは妥当性があると結論付けられる. また, 図 7 から, 受注頻度が 0.3 を下回る辺りから, 急速に *DRE* が上昇していることが分かり, これらのクラスタに属するサンプル(商品)は予測が困難であるといえる. 表1に, 各クラスタの受注頻度と *DRE*(ともにクラスタ内平均値)を示す.

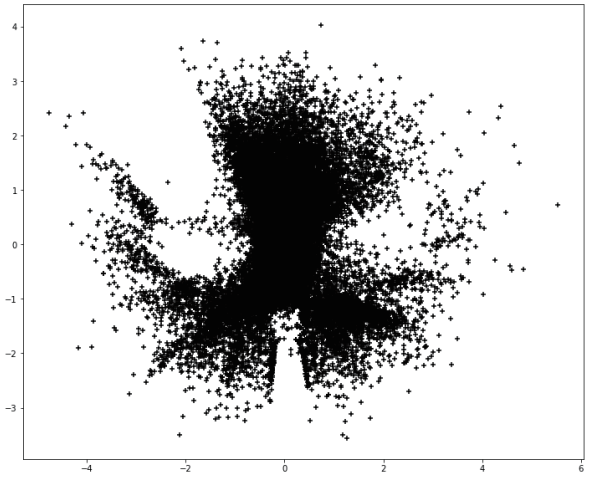

図 3: 潜在空間の分布

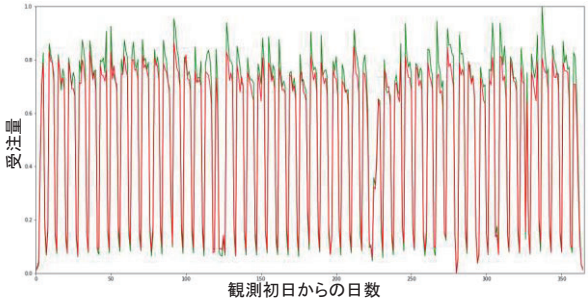

図 4: 再構築されたデータと観測データ
(緑線:観測データ, 赤線:再構築データ)

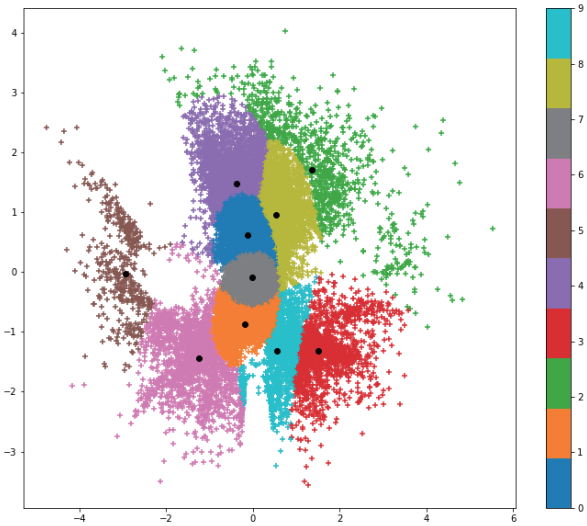

図 5: GMM によるクラスタリング結果

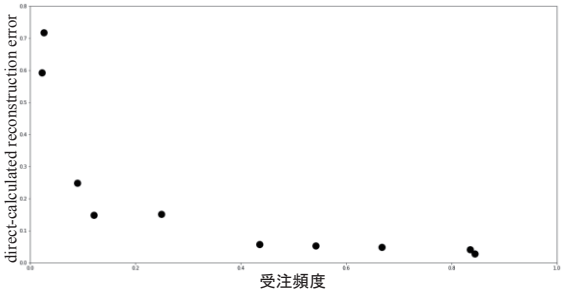

図 7: 受注頻度と *DRE* との関係

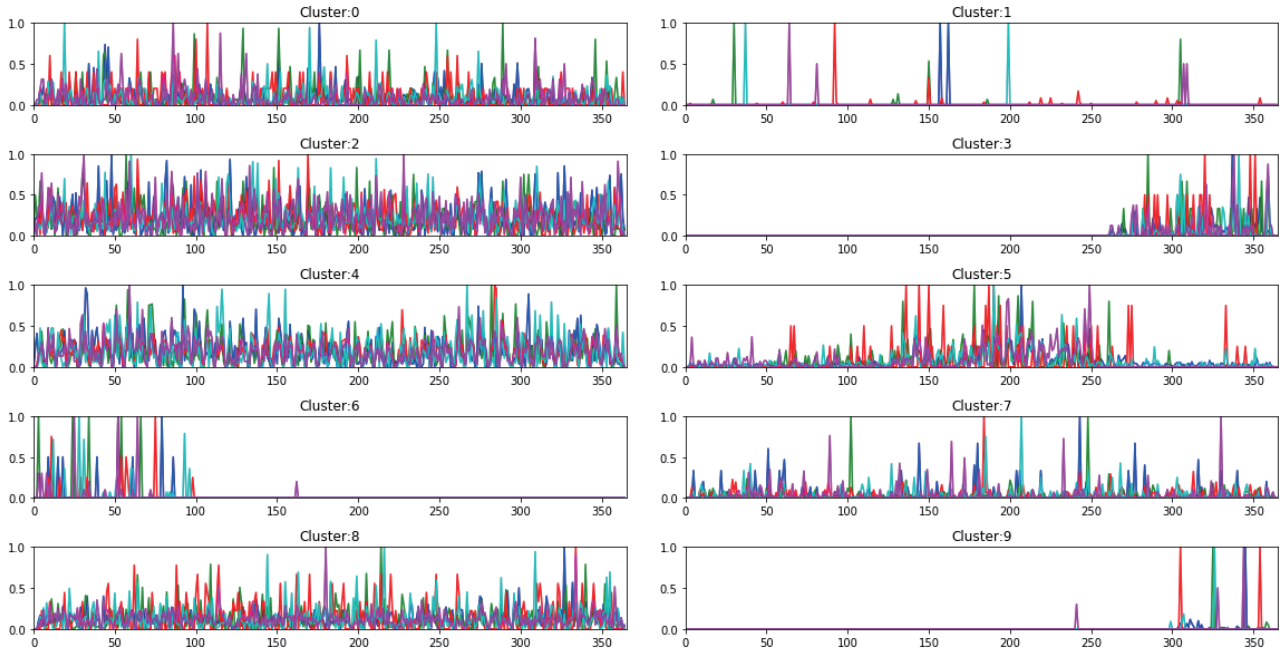

図 6: 各クラスタに属するサンプルの観測データ例（横軸：観測初日からの日数，縦軸：正規化した受注量）

表 1: 各クラスタの受注頻度と *DRE*

クラスタ	受注頻度	*DRE*
0	0.542	0.052
1	0.025	0.717
2	0.835	0.041
3	0.121	0.149
4	0.844	0.027
5	0.435	0.057
6	0.089	0.248
7	0.249	0.151
8	0.667	0.048
9	0.023	0.593

6. おわりに

商品受注の年間推移データに対して VAE を適用して，潜在空間に写像したベクトルに対してクラスタリングを行うことで，受注傾向に基づく商品クラスタリングの実験を行った．実験の結果，類似する受注傾向を持つ商品が同一のクラスタに属することを確認した．また，予測の困難性を示す指標として direct-calculated reconstruction error を提案し，これを用いることでクラスタリングの妥当性を確認した．

今後は，本稿の結果に基づいて，各クラスタに属する商品群を用いて個別に予測モデルの学習を行い，受注量予測の精度を検証する．また，予測困難度については，現状では予測困難度として算出した direct-calculated reconstruction error の数値が，我々の意図通り．予測困難度に依存して大きくなっているのか，あるいは，今回学習した VAE の再現性能が十分でないために誤差が大きくなったのかを精査し切れていない．今後は， VAE の表現性能の向上，および VAE 以外のモデルを用いた実験を行い，同様の結果が得られることを確認する．

参考文献

[Aljalbout 18] Aljalbout, E., Golkov, V., Siddiqui, Y., Strobel, M. and Cremers, D.: Clustering with Deep Learning: Taxonomy and New Methods, *arXiv 1801.07648,* (2018).

[Kingma 14a] Kingma, D. P. and Welling, M.: Auto-Encoding Variational Bayes., *in Proc. 2nd International Conference on Learning Representations,* (2014).

[Dilokthanakul 16] N. Dilokthanakul, P. A. Mediano, M. Garnelo, M. C. Lee, H. Salimbeni, K. Arulkumaran and M. Shanahan, Deep unsupervised clustering with gaussian mixture variational autoencoders, *arXiv:1611.02648*, (2016).

[Kingma 14b] Kingma, D. and Jimmy B.: Adam: A method for stochastic optimization., *arXiv:1412.6980*, (2014).

[鈴木 16] 鈴木 雅大, 松尾 豊: 深層生成モデルを用いたマルチモーダル学習, *人工知能学会全国大会,* (2016).