ガウシアンを用いたディープな教師なしクラスタリング混合変分オートエンコーダー

Nat Dilokthanakul[1],*, Pedro A. M. Mediano[1], Marta Garnelo[1], Matthew C. H. Lee[1]、Hugh Salimbeni[1]、Kai Arulkumaran[2]、Murray Shanahan[1] [1] コンピューティング学科、[2]バイオエンジニアリング学科
インペリアル・カレッ
ジ・ロンドン
ロンドン、イギリス
*n.dilokthanakul14@imperial.ac.uk

ABSTRACT

本研究では、事前分布にガウス混合物を用いた変分自己符号化モデル（VAE）を研究し、深層生成モデルによる教師なしクラスタリングを行うことを目的とする。その結果、通常のVAEで発生することが知られている過剰正則化の問題が、このモデルでも発生し、クラスタの縮退につながることがわかった。VAEでこの問題を軽減することが示されている最小情報制約と呼ばれるヒューリスティックが、我々のモデルを用いた教師なしクラスタリングの性能を向上させるためにも適用できることを示す。さらに、このヒューリスティックの効果を分析し、様々なプロセスを視覚化して直感的に理解できるようにした。最後に、合成データであるMNISTとSVHNを用いて我々のモデルの性能を実証し、得られたクラスタが明確で、解釈可能であり、その結果、教師なしクラスタリングにおいて最先端の結果に匹敵する性能を達成できることを示す。

1 イントロダクション

教師なしのクラスタリングは，機械学習研究の基本的な課題である．*k*-means やガウス混合モデル (GMM) (Bishop, 2006) などの古くから確立された手法は，現在でも多くのアプリケーションの中核をなしているが (Aggarwal & Reddy, 2013)，これらの類似性測定は，データ空間の局所的な関係に限定されるため，潜在的な空間に隠された階層的な依存性を捉えることができない．一方、深層生成モデルは、豊富な潜在的構造を符号化することができる。教師なしクラスタリング問題に*直接*適用されることはあまりないが、結果として得られる低次元空間に古典的なクラスタリング技術を適用して、次元の縮小に使用することができる(Xie et al., 2015)。これは、次元性削減技術の基礎となる仮定が、一般にクラスタリング技術の仮定とは独立しているため、不満足なアプローチである。

深層生成モデルは、観測されたデータの密度を、その潜在的な構造、すなわち隠れた原因に関するいくつかの仮定の下で推定しようとするものである。これにより、純粋に教師付き学習によって学習されたモデルよりも、より複雑な方法でデータを推論することができます。しかし、複雑な潜在的構造を持つモデルの推論は難しい。最近の近似推論の進歩は、扱いやすい推論アルゴリズムを構築するためのツールを提供している。差分可能なモデルと変分推論を組み合わせることで、以前の推論手法では不可能だったサイズのデータセットまで推論を拡張することが可能になった（Rezende et al.2014）。このフレームワークの下で人気のあるアルゴリズムの1つが、変分自動エンコーダー（VAE）である（Kingma & Welling, 2013; Rezendeetal., 2014）。

本論文では、VAEのフレームワークを用いて教師なしクラスタリングを行うアルゴリズ

ムを提案する。そのために、観測データがマルチモーダルな事前分布から生成されると仮定することで、生成モデルを教師なしクラスタリング用に調整することができると仮定し、それに対応して、再パラメータ化トリックを用いて直接最適化することができる推論モデルを構築する。また、VAEにおける過剰正則化の問題は、クラスタリングの性能に深刻な影響を与える可能性があり、Kingmaら（2016）が導入した最小情報制約によって緩和できることを示す。

1.1 関連する仕事

教師なしクラスタリングは、潜在変数を分離する問題のサブセットと考えることができ、教師なしで潜在空間の構造を見つけることを目的としています。最近では、データの様々な変動要因に対応する潜在変数を分離したモデルを学習することに取り組んでいる。腹側視覚ストリームの学習圧力にヒントを得て、Higginsら（2016）は、VAEの下限に正則化係数を加えることで、画像から離接された特徴を抽出することができました。VAEと同様に、生成的敵対ネットワーク（GAN）から分離された特徴を得るための努力も行われています（Goodfellow et al.、2014）。これは最近、InfoGANs（Chen et al., 2016a）で達成されたもので、構造化された潜在変数がノイズベクトルの一部として含まれ、これらの潜在変数と生成器分布の間の相互情報が、2つのネットワーク間のミニマックスゲームとして最大化されます。同様に、反復的償却グルーピングとラダーネットワークを組み合わせたTagger（Greffetal., 2016）は、入力を反復的にノイズ除去し、再構成の一部を異なるグループに割り当てることで、画像内のオブジェクトを知覚的にグループ化することを目指している。Johnsonら（2016）は、構造化VAEsと呼ばれるアルゴリズムにおいて、償却済み推論と確率的変分推論を組み合わせる方法を紹介した。Structured VAEsは、事前分布としてGMMを用いたディープモデルの学習が可能です。Shuら（2016）は、マルチモーダルな事前分布を持つVAEを導入し、標準的な変分目的に対する変分近似を最適化して、ビデオ予測タスクでの性能を示しました。

我々の研究と最も密接に関連している研究は、Kingmaら（2014）による積層生成半教師付きモデル（M1+M2）である。主な違いの1つは、彼らの事前分布が連続変数と離散変数のニューラルネットワーク変換であり、それぞれガウス型とカテゴリー型の事前分布を持っていることです。一方、我々のモデルの事前分布は、ガウス変数のニューラルネットワーク変換であり、ガウスの混合物の平均と分散をパラメトリックに表現し、混合物の成分にはカテゴリー変数を使用しています。重要なのは、Kingmaら（2014）は彼らのモデルを半教師付き分類タスクに適用しているのに対し、我々は教師なし分類に焦点を当てていることです。そのため、我々の推論アルゴリズムは後者に特化したものとなっている。

我々の結果を、ディープジェネレーティブモデルを用いた教師なしクラスタリングにおけるいくつかの直交する最先端技術と比較する：Deep embedded Clustering (DEC) (Xie et al, 2015)、Adversar- ial autoencoder (AAE) (Makhzani et al, 2015)、categorial GANs (CatGANs) (Springenberg, 2015)。

2 変形オートエンコーダー

VAEは、変分ベイズ法と、ニューラルネットワークが提供する柔軟性と拡張性を組み合わせたものです（Kingma&Welling, 2013; Rezendeetal, 2014）。変分ベイズ法を用いると、難解な推論問題を最適化問題に変えることができ(Wainwright & Jordan, 2008)、推論のための利用可能なツールのセットを最適化技術も含めて拡張することができます。しかし、古典的な変分推論では、ほとんどの問題を最適化するために、尤度と事前分布が共役でなければならないという制限があり、このようなアルゴリズムの適用範囲が限られてしまう。変分オートエンコーダは、条件付事後を出力するためにニューラルネットワークを導入し(Kingma&Welling, 2013)、その結果、確率的勾配降下法や標準的なバックプロパゲーションによって変分推論の目的を扱いやすく最適化することができます。reparametrisation trickとして知られるこの技術は、連続的な確率変数を介した逆伝播を可能にするために提案されました。通常の状況では、確率変数を介したバックプロパゲーションは、モンテカルロ法を用いなければ不可能ですが、決定論的な関数と別個のノイズ源の組み合わせによって潜在変数を構成することで、これを回避しています。詳細はKingma&Welling (2013)を参照してください。

3 ガウス混合変量オートエンコーダー

通常のVAEでは、潜在変数に対する事前処理は、一般的に等方性ガウシアンです。このような事前の選択により、多変量ガウスの各次元は、データからの個別の連続的な変動要因の学習に向けて押し出され、構造化されて分離された学習された表現になる可能性があります。これにより、より解釈しやすい潜在的な変数を得ることができますが(Higginsetal.,
2016)、ガウス型事前分布は、学習された表現が単峰性しか持たないため、制限があります。

は、より複雑な表現を許容しない。その結果、VAEの数多くの拡張が開発され、ますます複雑なプライアを指定することで、より複雑な潜在的表現を学習することができるようになりました(Chungetal., 2015; Gregoretal., 2015; Eslamiet al., 2016)。

本論文では，事前分布としてガウスの混合分布を選択する．これは，単一モードのガウス事前分布を直感的に拡張したものである．観測されたデータがガウスの混合物から生成されたと仮定すると、データポイントのクラスを推論することは、データポイントが潜在分布のどのモードから生成されたかを推論することと同じです。これにより、潜在空間を異なるクラスに分離することが可能になりますが、このモデルでの推論は自明ではありません。一般的にVAEに使用される再パラメトリック化のトリックは、離散変数には直接適用できないことがよく知られています。離散変数の勾配を推定するためのいくつかの可能性が提案されている(Glynn, 1990; Titsias &Lázaro-Gredilla, 2015)。また、Graves（2016）は、GMMを介したバックプロパゲーションのアルゴリズムを提案した。その代わりに、標準VAEのアーキテクチャを調整することで、我々のガウス混合変分オートエンコーダー（GMVAE）の変分下限の推定量を、再パラメトリック化トリックを通じて標準バックプロパゲーションでオプティマイズすることができ、推論モデルをシンプルに保つことができることを示す。

3.1 生成・認識モデル

生成モデル$p_{\theta,\vartheta}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{w}, \boldsymbol{z}) = p(\boldsymbol{w})p(\boldsymbol{z}|\boldsymbol{w})p_\theta(\boldsymbol{x}|\boldsymbol{w}, \boldsymbol{z})p_\vartheta(\boldsymbol{y}|\boldsymbol{x})$を考えてみましょう。ここでは、観測されたサンプル$\boldsymbol{y}$が、潜在変数$\boldsymbol{x}$, $\boldsymbol{w}$, $\boldsymbol{z}$のセットから以下のプロセスで生成されます。

$$\boldsymbol{w} \sim \mathrm{N}(0, \boldsymbol{I}) \tag{1a}$$

$$\boldsymbol{z} \sim \mathrm{Mult}(\boldsymbol{\pi}) \tag{1b}$$

$$\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{w} \sim \prod_{k=1}^{K} \mathrm{N}\left(\boldsymbol{\mu}_{z_k}(\boldsymbol{w}; \theta), \mathrm{diag}\,\boldsymbol{\sigma}^2_k(\boldsymbol{w}; \theta)\right)^{z_k} \tag{1c}$$

$$\boldsymbol{y}|\boldsymbol{x} \sim \mathrm{N}\left(\mu(\boldsymbol{x}; \vartheta), \mathrm{diag}\,\boldsymbol{\sigma}^2(\boldsymbol{x}; \vartheta)\right) \text{ or } \mathrm{B}\left(\mu(\boldsymbol{x}; \vartheta)\right). \tag{1d}$$

ここで、Kは混合物に含まれる成分の事前定義された数であり、$\boldsymbol{\mu}_{z_k}(\cdot; \theta)$、$\boldsymbol{\sigma}^2_k(\cdot; \theta)$、$\boldsymbol{\mu}(\cdot; \vartheta)$、および$\boldsymbol{\sigma}^2(\cdot; \vartheta)$は，それぞれパラメータ$\theta$と$\vartheta$を持つニューラルネットワークで与えられる．すなわち，観測された標本$\boldsymbol{y}$は，ϑでパラメータ化されたニューラルネットワークの観測モデルと継続的な潜在変数$\boldsymbol{x}$から生成され，さらに，$\boldsymbol{x}$の分布$\boldsymbol{w}$は，θでパラメータ化された別のニューラルネットワークモデルと入力$\boldsymbol{w}$で指定された平均と分散を持つガウス混合分布である．

具体的には、θでパラメータ化されたニューラルネットワークは、入力として$\boldsymbol{w}$が与えられると、K個の平均値$\boldsymbol{\mu}_{z_k}$とK個の分散値$\boldsymbol{\sigma}^2_k$のセットを出力します。ワンショット・ベクトル$\boldsymbol{z}$は，混合確率$\boldsymbol{\pi}$からサンプリングされ，ガウス混合から1つの成分を選択します．$\boldsymbol{z}$が一様に分布するようにパラメータ$\pi_k=K^{-1}$を設定します。このモデルの生成的な見方と変分的な見方を図1に示します。

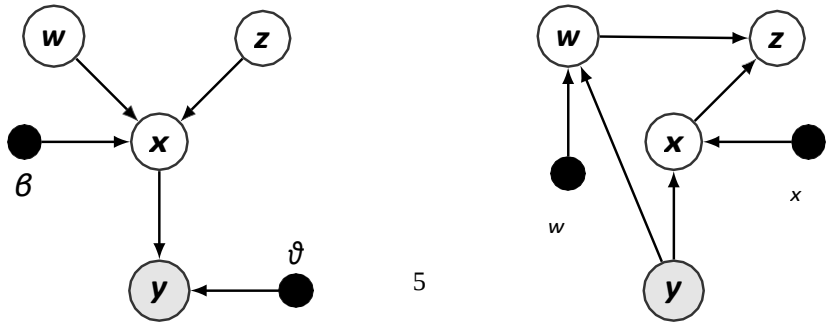


図1：ガウス混合変分自動符号化器（GMVAE）の生成モデル（左）と変分族（右）を示すグラフモデル

3.2 認識モデルによる推論

生成モデルの学習は，変分推論の目的，すなわち，次のように書ける対数信頼性下限値（ELBO）を用いて行われる．

$$\mathcal{L}_{ELBO}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{w}, \boldsymbol{z}) = \mathbb{E}_{q(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{z}|\boldsymbol{y})} \frac{p_{\theta,\vartheta}}{q(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{z}|\boldsymbol{y})} \tag{2}$$

ここでは、平均フィールド変分群$q(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{z}|\boldsymbol{y})$をポステリアの代理として仮定し$q_{\varphi_x}(\boldsymbol{x}_i|\boldsymbol{y}_i)q_{\varphi_w}(\boldsymbol{w}_i|\boldsymbol{y}_i)p_\theta(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{w}_i)$，ここで$i$これを因子分解する。さらに表記を簡単にするために、iを削除し、一度に1つのデータポイントを考慮します。各変分係数を認識ネットワークφ_xとφ_wでパラメータ化し、変分分布のパラメータを出力し、その形式をガウスの後置と指定します。z事後を導き出しました。$p_\theta(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{w})$、として。

$$p_\theta(z_j = 1|\boldsymbol{x}, \boldsymbol{w}) = \frac{p(z_j = 1)p(\boldsymbol{x}|_j z = 1, \boldsymbol{w})}{\sum_{k=1}^{K} p(z_k = 1)p(\boldsymbol{x}|_{z^j} = 1, \boldsymbol{w})} = \frac{\pi_j \mathcal{N}(\boldsymbol{x}|\mu_j(\boldsymbol{w}; \theta), \sigma_j(\boldsymbol{w}; \theta))}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\mu_k(\boldsymbol{w}; \theta), \sigma_k(\boldsymbol{w}; \theta))}. \tag{3}$$

下限値は、次のように書くことができます。

$$\mathcal{L}_{ELBO} = \mathbb{E}_{q(\boldsymbol{x}|\boldsymbol{y})} \log p_\vartheta(\boldsymbol{y}|\boldsymbol{x}) - \mathbb{E}_{q(\boldsymbol{w}|\boldsymbol{y})p(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{w})}(q_\varphi(\boldsymbol{x}|\boldsymbol{y}) p_\theta(\boldsymbol{x}|\boldsymbol{w}, \boldsymbol{z})) - KL(q_{\varphi_w}(\boldsymbol{w}|\boldsymbol{y})||p(\boldsymbol{w})) - \mathbb{E}_{q(\boldsymbol{x}|\boldsymbol{y})q(\boldsymbol{w}|\boldsymbol{y})} KL(p_\theta(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{w})||p(\boldsymbol{z})). \tag{4}$$

ここでは、下界の項をそれぞれ、再構成項、条件付き事前項、w-prior項、z-prior項と呼ぶ。

3.2.1 条件付先行項

再構成項は，$q(\boldsymbol{x}|\boldsymbol{y})$からモンテカルロサンプルを抽出して推定することができ，その際，標準的な再パラメータ化トリックを用いて勾配を逆伝播させることができる（Kingma & Welling, 2013）．w-prior項は解析的に計算できます。

重要なのは、このようにモデルを構築することで、離散分布$p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{w})$からサンプリングしなくても、式(5)を用いて条件付き事前項を推定できることです。

$$\mathbb{E}_{q(\boldsymbol{w}|\boldsymbol{y})p(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{w})} KL(q_{\varphi_x}(\boldsymbol{x}|\boldsymbol{y})||p_\theta(\boldsymbol{x}|\boldsymbol{w}, \boldsymbol{z})) \approx \frac{1}{M}\sum_{j=1}^{M}\sum_{k=1}^{K} p_\theta(z_k = 1|\boldsymbol{x}^{(j)}, \boldsymbol{w}^{(j)}) KL(q_{\varphi_x}(\boldsymbol{x}|\boldsymbol{y})||p_\theta(\boldsymbol{x}|\boldsymbol{w}^{(j)}, z_k = 1)) \tag{5}$$

$p_\theta(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{w})$は1回のフォワードパスですべての$\boldsymbol{z}$に対して計算できるので、その期待値は簡単な方法で計算でき、通常通りバックプロパゲートすることができます。$q_{\varphi_w}(\boldsymbol{w}|\boldsymbol{y})$に対する期待値は、$M$個のモンテカルロサンプルを用いて推定することができ、その勾配は再パラメータ化トリックを用いてバックプロパゲーションすることができます。この期待値の計算方法は、Kingmaら（2014）のmarginalisationアプローチと似ていますが、微妙な違いがあります。Kingmaら（2014）は、z-posteriorの各成分を得るために複数のフォワードパスを必要とします。我々の方法は、θでパラメータ化されたニューラルネットワークのより広い出力層を必要とするが、1つのフォワードパスしか必要としない。どちらの手法もクラスタ数に応じてリニアにスケールアップします。

3.3 離散型潜在変数のklコスト

我々のELBOで最も変わった項はz-prior項である。z-prior項

は、xとwの値から直接クラスタリング割り当て確率を計算し、wによって生成された各クラスタ位置からx が どれだけ離れているかを問う。したがって、z-prior項は、クラスタの位置と符号化された点xを同時に操作することによって、z-posteriorと一様事前の間のKL発散を減少させることができる。直観的には、クラスタ間の重なりを最大にし、平均値を近づけることによってクラスタを統合しようとするだろう。この項は、他のKL正則化項と同様に、再構成項と緊張関係にあり、学習データ量の増加に伴って過剰な力を発揮することが予想されます。

3.4 オーバーレギュレーション問題

VAEのトレーニングにおける正則化項の影響が強すぎる可能性については、VAEの文献で何度も説明されています（Bowmanら、2015年、Sønderbyら、2016年、Kingmaら、2016年、Chenら、2016b）。事前の強い影響の結果、得られた潜在的な表現は、しばしば過度に単純化され、データの基本的な構造を十分に表現していない。これまで、この効果を克服するために2つの主要なアプローチがあった：1つの解決策は、KL項からの正則化をゆっくりと組み込む前に、再構成項がオートエンコーダーネットワークを訓練するようにすることで、訓練中にKL項をアニーリングすることである（Sønderbyら、2016）。もう一つの主なアプローチは、KL項の効果がある閾値を下回ったときに取り除くカットオフ値を設定することで、目的関数を修正することである（Kingma et al.2016）。以下の実験セクションで示すように、この過剰正則化の問題は、GMVAEクラスターの割り当てにおいても広まっており、大きな縮退クラスターに現れます。Kingmaら（2016）が提案した第2のアプローチは、確かにこのマージ現象を緩和することを示していますが、過剰正則化問題の解決策を見つけることは、依然として挑戦的な未解決問題です。

4 実験の様子

本実験の主な目的は、提案モデルの精度を評価するだけでなく、データの意味のある異なる潜在的な表現の構築に関わる最適化のダイナミクスを理解することである。このセクションは3つの部分に分かれています。

1. 我々はまず、低次元の合成データセットにおける推論プロセスを研究し、特に、過剰正則化問題がGMVAEのクラスタリング性能にどのように影響するか、また、この問題をどのように軽減するかに焦点を当てる。
2. そして、MNISTの教師なしクラスタリングタスクで、我々のモデルを評価します。
3. 最後に，潜在変数の異なる値を条件として，我々のモデルから生成された画像を示し，GMVAEが分離された解釈可能なラテント表現を学習できることを示します。

このセクションでは、以下のデータセットを使用しています。

- **合成データ**です。Johnsonら（2016）の発表を模倣した合成データセットを作成します。これは、5つの円の弧から作成された10,000個のデータポイントを持つ2Dデータセットです。
- **MNIST**．28×28のグレースケール画像で構成され，60,000個のトレーニングサンプルと10,000個のテストサンプルからなる標準的な手書き数字データセット(LeCunetal, 1998).
- **SVHN**:
  32x32のハウスナンバーの画像集（Netzeretal.標準セットと追加のトレーニングセットを合わせて、約60万枚の画像を使用しています。

4.1 シンセティック・データ（SYNTHETIC DATA

クラスタリングの性能は，式（6）で表されるz-prior項の大きさを学習中にプロットすることで定量化する。この量は、異なるクラスタがどれだけ重なり合っているかの尺度と考えることができます。我々の目的は、潜在空間において意味のあるクラスタリングを行うことなので、モデルが別々のクラスタを学習するにつれて、この量は減少すると予想されます。

$$L_z = -E_{q(\boldsymbol{x}|\boldsymbol{y})q(\boldsymbol{w}|\boldsymbol{y})}\ \boldsymbol{KL}(p_\theta(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{w})||p(\boldsymbol{z})) \tag{6}$$

しかし、経験的には、そうではないことがわかりました。我々のモデルが収束させる潜在表現は、図2dと図3aに見られるように、異なるクラスタに関する情報を表すのではなく、すべてのクラスを同じ大きなクラスタに統合します。その結果、各データポイントはどのクラスターにも同じように属する可能性があり、潜在表現はクラス構造に関して全く情報を提供しません。

この現象は、*z*-prior項による過剰な正則化の結果と解釈できると主張する。この量は、下界のKL項の最適化によって押し上げられていることを考えると

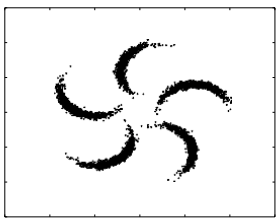
は、クラスに関する情報のエンコードを確実にするために、学習によって減少するのではなく、可能な限り最大の値であるゼロに達します。これは、初期の学習段階で事前の影響が強すぎて、モデルのパラメータが局所的な最適値になってしまい、後に再構成項で追い出すのが難しくなっているのではないかと考えられます。

この観察結果は、概念的にはregu-lar VAEで遭遇する過剰正則化問題と非常によく似ており、したがって、同様のヒューリスティクスを適用することで問題を軽減できるはずだと仮説を立てています。図2fに示すように、Kingmaら（2016）が先に述べた下界プロポーズへの修正を用いることで、*z*-priorによる過剰正則化を回避できることを示しています。これは、*z*-priorからのコストを、その閾値を超えるまで一定の値λで維持することで達成されます。形式的には、修正*z*-prior項は次のように書かれます。

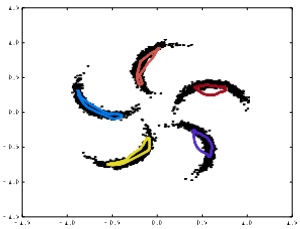
$$L_z' = -\max(\lambda, E_{q(\boldsymbol{x}|\boldsymbol{y})q(\boldsymbol{w}|\boldsymbol{y})} KL(p_\theta(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{w}) || p(\boldsymbol{z}))) \tag{7}$$

この修正は、すべてのクラスターを統合するという*z*-priorの初期効果を抑制し、*z-prior* のコストが十分に高くなるまで、クラスターが分散するようにします。この時点で、その効果は大幅に減少し、十分にオーバーラップしている個々のクラスターをマージすることにほぼ限定されます。このことは、図2eと図2fを見れば一目瞭然です。前者は、*z*-priorコストが考慮される前のクラスターを示しており、そのためクラスターは拡散することができました。*z*-priorが有効になると、図2fに見られるように、非常に近接したクラスターが結合されます。

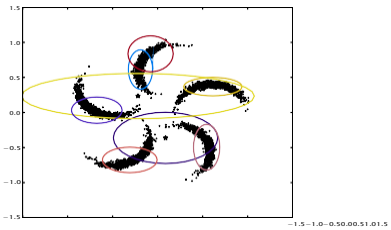
最後に、分布の変換にニューラルネットワークを使用することの利点を示すために、我々のモデルによって観測された密度（図2c）と、データ空間における通常のGMM（図2c）を比較します。図に示されているように、GMVAEは、通常のGMMよりもはるかに豊かな、したがってより正確な表現を可能にし、したがって、非ガウスデータのモデル化に成功しています。



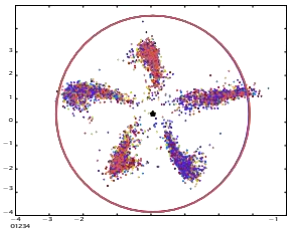
(a) データ空間におけるデータポイント



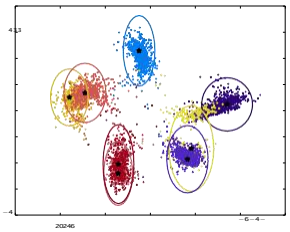
(b) GMVAEの密度



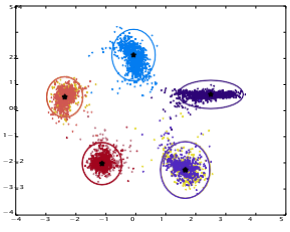
(c) GMMの密度



(d) 最適な状態でない場合の潜在的な空間



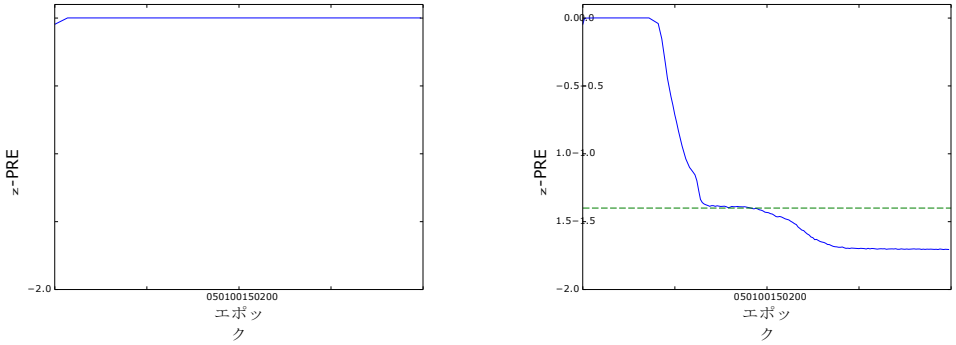
(e) 潜在的な空間、クラスターの広がり



(f) 潜在的な空間、収束時

図2：**合成データセットの視覚化**。(a) データは2次元のデータ空間に5つのモードで分布している。(b)

GMVAEは、データ空間内の非ガウス分布の混合物を用いてデータをモデル化できる密度モデルを学習する。(c)
GMMは、ガウス分布を仮定しているため、データをうまく表現できません。(d)
GMVAEは、過剰正則化の問題があり、潜在空間を見たときに貧弱な最小値になる可能性があります。(e）ELBO（Kingma et al., 2016）の修正を使用すると、クラスターが広がることができます。(f)
モデルが収束すると、z-prior項が有効になり、最終段階で過剰なクラスターをマージすることでクラスターを正則化します。



(a) z 通常のELBOでの先行期間　　(b) z 修正を加えた先行期間

図3: ***z*-prior項のプロット**。(a) 情報制約がない場合、GMVAEはKLコストを回避するためにすべてのクラスタを結合する貧弱な最適値に収束するため、過剰な正則化に悩まされます。(b) 閾値(点線)に達する前に、*z*-prior項からの勾配をオフにして、クラスタが一緒に引き寄せられるのを防ぐことができます(詳細は本文を参照)。閾値に到達するまでに、クラスターは十分に分離されています。この時点では、*z*-prior termからのグラデーションを有効にしても、非常に重なり合ったクラスターを一緒にするだけです。*z-prior* の値は、その勾配を有効にした後も、意味のあるクラスタとより良い最適性をもたらす他の用語に圧倒され、減少し続けます。

4.2 教師なしの画像クラスタリング

次に，データに含まれる離散的な情報を表現するモデルの能力を，画像のクラスタリングタスクで評価する．MNISTトレーニングデータセットでGMVAEを学習し，テストデータセットでGMVAEのクラスタリング性能を評価する．GMVAEによって与えられたクラスタ割り当てを真の画像ラベルと比較するために、Makhzanietal.（2015）の評価プロトコルに従うが、ここでは分かりやすくするために要約する。この方法では，テストセットの中からクラスタi***に属する***確率が最も高い要素を見つけ，そのラベルをiに属する他のすべてのテストサンプルに割り当てます。これをすべてのクラスタ$i = 1, \ldots, K$について繰り返し，割り当てられたラベルを真のラベルと比較して，教師なし分類エラー率を求めます。

合成データセットでGMVAEを学習するとクラスタ縮退の問題が発生するが、MNISTデータセットではこの問題は発生しない。そこで、ELBOを直接利用してGMVAEを最適化することにしました。MNISTベンチマークにおいて、GMVAEと最近の他の手法を用いて得られた結果の概要を表1に示す。GMVAEは，最新の技術に匹敵する分類スコアを達成した．[1]AAE（Adversarial Autoencoders）を除いては，最先端の技術に匹敵する分類スコアを得ることができました．この理由は，やはりVAEの目的語であるKL項に関係していると思われます。Hoffmanらが示したように、敵対的オートエンコーダーの目的における重要な違いは、ELBOのKL項を、潜在空間をより慎重に操作できるような敵対的損失に置き換えたことである(Hoffman & Johnson, 2016)。これらの実験で使用されたネットワークアーキテクチャの詳細は、付録Aに記載されています。

経験的には、図4に示すように、モンテカルロサンプルの数とクラスタの数を増やすことで、GMVAEは初期化に対してよりロバストになり、より安定することがわかりました。使用するサンプル数やクラスター数が少ない場合、GMVAEは劣悪なローカルミニマ

ムに早く収束し、データ分布のモードの一部を見逃すことがあります。

[1]私たちが最初に投稿した直後に、Rui Shuがガウス混合VAEに関する分析を行ったブログ記事（http://ruishu.io/2016/12/25/gmvae/）を公開したことは注目に値します。この記事では、前述のM2アルゴリズムとの比較に加えて、比較的シンプルなネットワークアーキテクチャを使用して、競争力のあるクラスタリングスコアを達成するバージョンを実装しています。重要なのは、ラベルなしで学習した場合、モデルM2が離散的な潜在変数を使用しないことを示していることです。GMVAEでこの問題がそれほど深刻ではない理由は、彼のブログで論じられているように、生成プロセスにおけるより制限的な仮定が最適化を助けているからかもしれません。

表1：クラスタ数（K）を変えた場合のMNISTの教師なし分類精度（正しいラベルの割合として報告されている

方法	K	ベストランアベレージラン
CatGAN (Springenberg, 2015)		2090.30- です。
AAE(Makhzanietal., 2015)		± 16-90.452.05
AAE(Makhzanietal., 2015)		± 30-95.901.13
DEC (Xieetal., 2015)		1084.30-.
gmvae (m =	1)1087.	3177.78 ± 5.75
gmvae (m =	10)1088.	5482.31 ± 3.75
gmvae (m =	1)1689.	0185.09 ± 1.99
gmvae (m =	10)1696.	9287.82 ± 5.33
gmvae (m =	1)3095.	8492.77 ± 1.60
gmvae (m =	10)3093.	2289.27 ± 2.50

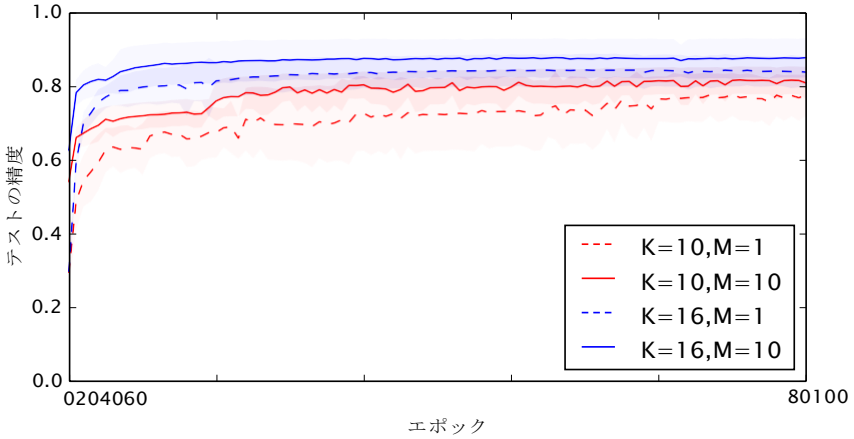


Figure 4: **クラスタ数(K)とモンテカルロサンプル数を変えた場合のクラスタリング精度(中)**：数回のエポックの後、GMVAEは解に収束する。クラスターの数を増やすことで、解の質が大幅に向上します。

4.2.1 画像生成

これまで我々は，GMVAEがデータセット内の自然なクラスタを拾い，これらのクラスタが画像の実際のクラスと何らかの構造を共有していることを論じてきた．ここでは、潜在空間の分布において学習された成分が実際にデータの意味のある特性を表していることを示すために、MNISTにおいて$K = 10$でGMVAEを訓練します。まず，GMVAEからサンプリングする際には，次の2つの確率的な要因があることに注意してください．

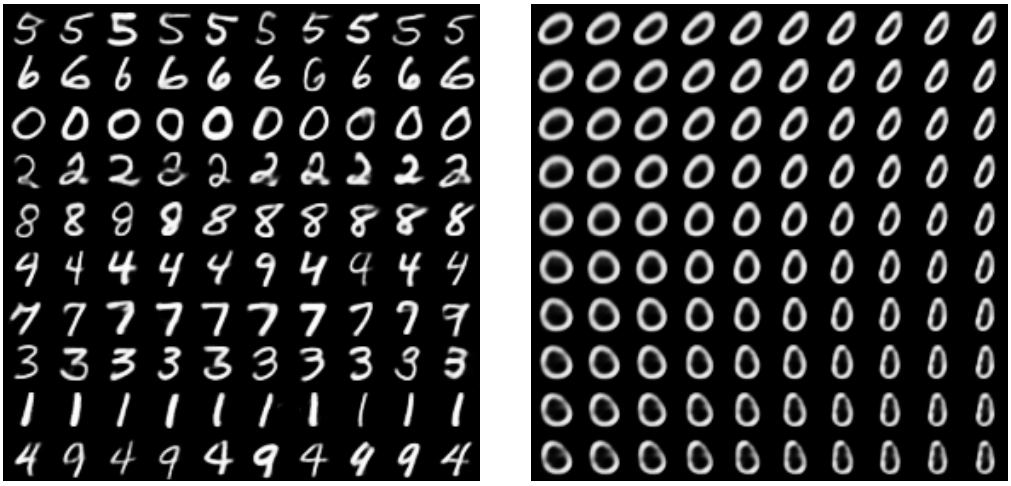
1. その事前情報から$\boldsymbol{w}$をサンプリングし、ニューラルネットワークθを介して$\boldsymbol{x}$の平均値と分散値を生成します。
2. $\boldsymbol{w}$と$\boldsymbol{z}$で決まるガウス混合物から$\boldsymbol{x}$をサンプリングし、ニューラルネットワークϑによって画像を生成する。

図5aでは，$\boldsymbol{w}$=0に設定して，結果として得られるガウス混合物から複数回サンプリングすることで，後者の選択肢を検討しています．図5aの各行は、ガウス混合物の異なる成分からのサンプルに対応しており、同じ成分からのサンプルは、一貫して同じクラスの数字の画像になることがはっきりとわかります。これにより、学習された潜在表現には

、よく区別されたクラスターが含まれており、1つの数字に1つのクラスターが含まれていることが確認できます。さらに、図5bでは、ガウス混合成分を滑らかに変化させることで、生成された画像の感度を調べています。

wと同じ成分からサンプリングしています。***z***は生成される画像のクラスを確実に制御するのに対し、***w***は桁の「スタイル」を設定することがわかります。

最後に、図6はSVHNを用いて学習したGMVAEから抽出した画像を示しており、GMVAEが視覚的に類似した画像をまとめていることがわかります。



(a) 変化するz　　(b) 変化するw

図5：**生成されたMNISTのサンプル**。(a)
各行には，ガウス混合物の異なるガウス成分からランダムに生成された10個のサンプルが含まれている．GMVAEは、離散的な潜在変数zが直接桁の値に対応する意味のある生成モデルを、教師なしで学習する。(b)
w空間をトラバースして生成されたサンプル。wの各位置は数字の特定のスタイルに対応する。

図6：**生成されたSVHNサンプル**。各行は，異なるガウス成分からランダムに生成された10個のサンプルに対応する．GMVAEは、視覚的に類似した画像をグループ化します。

5　結論を言うと

我々は、潜在的な符号化空間の1つのレベルがガウス混合モデルの形をしている変分オー

トエンコーダのクラスを導入し、以下を可能にする生成プロセスを指定しました。

を用いて、変分ベイズ最適化の目的を定式化します。次に、VAEにおける過剰な正則化の問題について議論する。我々のモデルでは、この問題はクラスターの縮退という形で現れることを示します。重要なのは、この問題が標準的なヒューリスティックスで解決できることです。

このモデルを、一般的なデータセットを用いた教師なしのクラスタリングタスクで評価したところ、現在の技術水準と比較して満足のいく結果が得られた。最後に、生成モデルからのサンプリングにより、潜在表現で学習されたクラスターが可視データの意味のある特徴に対応することを示す。潜在空間の同じクラスタから生成された画像は、関連する高レベルの特徴を共有している（例えば、同じMNISTの数字に対応している）が、完全に教師なしで学習されている。

GMVAEは、wの事前分布をガウス混合分布とすることで、積み重ねることができることは注目に値します。深層GMVAEは、層の数と層ごとのクラスタの数の両方に関して組み合わせ可能であることから、クラスタの数に応じてはるかに優れたスケーリングが可能です。このように、階層型クラスタリングのための深層GMVAEに関する将来の研究は可能性がありますが、そのためにはVAEに関連する永続的な最適化の課題にも取り組むことが重要です。

謝辞

実験に使用したGeForce GTX Titan Zを寄贈してくださったNVIDIA Corporationに謝意を表したいと思います。また，有用なコメントをいただいたJason Rolfe氏，Rui Shu氏，および査読者の方々に感謝いたします。また，本稿で使用した変分法ファミリーは，匿名の査読者によって提案されたものであることをお伝えしたいと思います。

参考文献

Charu C Aggarwal and Chandan K Reddy.*Data Clustering: algorithms and applications*.CRC Press, 2013.

クリストファー・M・ビショップパターン認識と機械学習.2006.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Ben- gio.Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel.Info-gan:information maximizing generative adversarial nets による Interpretable representation learning. *arXiv preprint arXiv:1606.03657*, 2016a.

Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel.Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016b.

J.J.Chung, K.Kastner, L.Dinh, K.Goel, A.Courville, and Y.Bengio.A Recurrent Latent Variable Model for Sequential Data.*ArXiv e-prints*, June 2015.

SM Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, Koray Kavukcuoglu, and Geoffrey E Hinton.Attend, infer, repeat: Fast scene understanding with generative models. *arXiv preprint arXiv:1603.08575*, 2016.

PW Glynn.確率的システムのための尤度比勾配推定．*ACMの通信*，33(10):75-84，1990.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio.Generative adversarial nets.*Advances in Neural Information Processing Systems*, pp.2672-2680, 2014で紹介しています。

アレックス・グレイブスStochastic backpropagation through mixture density distributions. *arXiv preprint arXiv:1607.05690*, 2016.

Klaus Greff、Antti Rasmus、Mathias Berglund、Tele Hotloo Hao、Jürgen Schmidhuber、Harri Valpola。Tagger:Deep unsupervised perceptual grouping. *arXiv preprint arXiv:1606.06724*, 2016.

カロル・グレゴール、イヴォ・ダニヘルカ、アレックス・グレイブス、ダニーロ・レゼンデ、ダーン・ウィアストラDraw:A recurrent neural network for image generation.In *Proceedings of The 32nd International Conference on Machine Learning*, pp.1462-1471, 2015.

I.Higgins, L. Matthey, X. Glorot, A. Pal, B. Uria, C. Blundell, S. Mohamed, and A. Lerchner.Unsupervised Deep Learningによる初期のビジュアルコンセプト学習。*ArXiv e-prints*, June 2016.

Matthew D. Hoffman and Matthew J. Johnson.エルボ手術：変分証拠下限を切り分けるためのまだ別の方法。*Workshop in Advances in Approximate Bayesian Inference, NIPS*, 2016.

Matthew J Johnson, David Duvenaud, Alexander B Wiltschko, Sandeep R Datta, and Ryan P Adams.Composing graphical models with neural networks for structured representation and fast infer- ence. *arXiv preprint arXiv:1603.06277*, 2016.

Diederik KingmaとJimmy Ba.Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling.自動符号化変分ベイズ. *arXiv preprint arXiv:1312.6114*, 2013.

Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling.深層生成モデルによる半教師付き学習。In *Advances in Neural Information Processing Systems*, pp.3581-3589, 2014.

Diederik P Kingma, Tim Salimans, and Max Welling.Improving variational inference with inverse autoregressive flow. *arXiv preprint arXiv:1606.04934*, 2016.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner.勾配ベースの学習の文書認識への応用.*Proceedings of the IEEE*, 86(11):2278-2324, 1998.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian Goodfellow.Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng.教師なしの特徴学習による自然画像中の数字の読み取り.2011.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra.Stochastic backpropagation and approximate inference in deep generative model. *arXiv preprint arXiv:1401.4082*, 2014.

R.Shu, J. Brofos, F. Zhang, M. Ghavamzadeh, H. Bui, and M. Kochenderfer.条件付き密度推定を用いた確率的なビデオ予測.In *European Conference on Computer Vision (ECCV) Workshop on Action and Anticipation for Visual Learning*, 2016.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther.How to train deep variational autoencoders and probabilistic ladder networks. *arXiv preprint arXiv:1602.02282*, 2016.

Jost Tobias Springenberg.カテゴリカル生成敵対ネットワークを用いた教師なし・半教師付き学習. *arXiv preprint arXiv:1511.06390*, 2015.

Michalis Titsias, Miguel Lázaro-Gredilla. ブラックボックス変分推論のための局所期待勾配. In *Advances in Neural Information Processing* Systems, pp.2638-2646, 2015.

Martin J WainwrightとMichael I Jordan。グラフィカルモデル、指数関数族、および変分法推論を行う。*Foundations and Trends®in Machine Learning*, 1(1-2):1-305, 2008.

Junyuan Xie, Ross Girshick, and Ali Farhadi.クラスタリング分析のための教師なしのディープエンベッディング。*arXiv preprint arXiv:1511.06335*, 2015.

ネットワークパラメータ

最適化には，Adam（Kingma & Ba, 2014）を使用し，学習率は10^{-4}，標準的なハイパーパラメータ値は$\beta_1 = 0.9$，$\beta_2 = 0.999$，$\epsilon(\Theta')$である$^{-8}$．実験に使用したモデル・アーキテクチャを表A.1，A.2，A.3に示す.

表A.1: $q_\varphi(\boldsymbol{x}, \boldsymbol{w})$のニューラルネットワークアーキテクチャモデル。隠れた層は$q(\boldsymbol{x})$と$q(\boldsymbol{w})$で共有されていますが、出力層はニューラルネットワークが4つの出力ストリームに分割されており、2つは次元N_x、残りの2つは次元Nとなっています。$_w$分散成分の値を正に保つために指数化します。アスタリスク(*)は，バッチ正規化とReLU非線形性の使用を示す．畳み込み層の場合，括弧内の数字はストライドパディングを示す.

データセット	入力	隠し	出力
合成	2	fc 120 ReLU 120 ReLU	N_w = 2, N_w = 2 (Exp), N_x = 2, N_x = 2 (Exp)
MNIST	28x28	コンビネーション 16x6x6* (1-0) 32x6x6* (1-0) 64x4x4* (2-1) 500* (2-1)	N_w＝150、N_w＝150（Exp）。 N_x = 200, N_x= 200 (Exp)
SVHN	32x32	64x4x4* (2-1) 128x4x4* (2-1) 246x4x4* (2-1) 500* (2-1)	N_w＝150、N_w＝150（Exp）。 N_x = 200, N_x= 200 (Exp)

表A.2: $p_\theta(\boldsymbol{x} \mid \boldsymbol{w}, \boldsymbol{z})$のニューラルネットワークアーキテクチャモデル。出力層は$2K$個の出力ストリームに分割され，K個のストリームは平均値を返し，残りのK個のストリームはすべてのクラスタの分散を出力する.

データセット	入力	隠し	出力
合成	2	fc 120 Tanh	$\{N_x = 2\}_{2K}$
MNIST	150	fc 500 Tanh	$\{N_x = 200\}_{2K}$
SVHN	150	fc 500 Tanh	$\{N_x = 200\}_{2K}$

表 A.3: $p_\theta(\boldsymbol{y} \mid \boldsymbol{x})$のニューラルネットワークアーキテクチャモデル。ネットワークの出力は，合成データではガウスパラメータ，MNISTとSVHNではベルヌーイパラメータで，ベルヌーイパラメータの値を0から1の間に保つためにロジスティック関数を使用している．アスタリスク（*）は，バッチ正規化とReLU非線形性を使用していることを示す．畳み込み層の場合，括弧内の数字はストライドパディングを示す.

データセット	入力	隠し	出力
合成	2	fc 120 ReLU 120 ReLU	$\{2\}_2$

MNIST	200	500*フルコンブ 64x4x4* (2-1) 32x6x6* (1-0) 16x6x6* (1-0)	28×28（シグモイド
SVHN	200	500*フルコンブ 246x4x4* (2-1) 128x4x4* (2-1) 64x4x4* (2-1)	32×32（シグモイド