社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

Nonparametric Bayesian Deep Visualization

石塚　治也†　　持橋　大地††

† 株式会社ブリヂストン　〒104–8340 東京都中央区京橋 3–1–1
†† 統計数理研究所　〒190–8562 東京都立川市緑町 10–3
E-mail: †haruya.ishizuka@bridgestone.com, ††daichi@ism.ac.jp

あらまし　高次元データを散布図で可視化する場合，次元削減により観測値を圧縮する必要がある．t-SNE に代表される，観測値間の類似度を元に次元削減を行う類似度ベース次元削減は，可視化の際に広く用いられている．しかし，観測値のベクトル表現によっては，観測値間の類似度と真の類似度が乖離するため，可視化精度が低下する．これに対して，ニューラルネットワーク (NN) を用いる深層潜在変数モデルは，観測値ベクトルよりも正確に特徴を反映する潜在表現を推定ができる可能性があり，ベクトル表現が原因で前者が機能しないときの有効な代替案になる．一方で性能の最適化には，NN のモデル構造など多くの超パラメータを調整する必要があり，その試行の中で，NN の大量のパラメータの学習を繰り返すため，計算時間が増大しやすい．また，可視化結果は設定された超パラメータの探索範囲によって変化する．本稿では，これらの問題点に対処するため，Nonparametric Bayesian Deep Visualization (NPDV) を提案する．NPDV は，NN による潜在表現の推定と可視化を同時に行う確率モデルであり，無限混合ガウスモデル，無限ユニット NN を併用することで，少数の超パラメータでモデルが構成される．さらに，無限ユニット NN は少数のパラメータで定義されるため，パラメータ数も既存の深層潜在変数モデルと比較して少ない．本稿では提案手法の詳細と，実験結果について報告する．

キーワード　データ可視化，ガウス過程，ノンパラメトリックベイズモデル，深層学習

Haruya ISHIZUKA† and Daichi MOCHIHASHI††

† Bridgestone Corporation 3–1–1 Kyobashi, Chuo-ku, 〒104–8340 Japan
†† The Institute of Statistical Mathematics 10–3 Midori–cho, Tachikawa–shi, 190–8562 Japan
E-mail: †haruya.ishizuka@bridgestone.com, ††daichi@ism.ac.jp

Key words　Data Visualization, Gaussian Processes, Nonparametirc-Bayes, Deep Learning

1. ま え が き

データ可視化は，データセットの特徴を把握するための探索的データ分析 (Exploratory Data Analysis : EDA) における有効な手段であり，散布図は広く使われているものの一つである．散布図により個体間の関係を俯瞰できるが，観測値ベクトルが 4 次元以上の場合，それらを直接可視化することはできない．この場合，次元削減により，我々が直観的に把握できる，2 または 3 次元の可視化表現に観測値ベクトルを圧縮する必要がある．その際に用いられる手法は大きく二つに分類される．

(1) 類似度ベース次元削減：このカテゴリに属する手法では，観測値ベクトル間の類似度をもとに次元削減を行う．その先駆的手法である Multi-Dimensional Scaling (MDS) [1] をはじめ，多くの手法が提案されている [2]．中でも，t-SNE [3] や UMAP [4] は，その可視化性能の高さから，高次元データの可視化に標準的に用いられている．しかし，入力となる観測値のベクトル表現によっては，観測値ベクトル間の類似度と個体間の真の類似度が乖離し，可視化精度が低下し得る．文書データは，使用されるベクトル表現が原因で二つの類似度に差が出るデータの一例である．文書データでは，単語の相対的な重要度を定量化する TF-IDF (Term Frequency-Inverse Document Frequency) [5] が各文書のベクトル表現として頻繁に用いられる．しかし，TF-IDF 表現では，good と well のような類義語が異なる変数として扱われるため，意味の似た文書同士でも，二者間の類似度が低く評価される場合がある．この場合，意味が近い文書同士が散布図上で互いに遠くに配置される．

(2) 確率的潜在変数モデル：このカテゴリに属する手法では，観測値よりも低次元かつ潜在的な確率変数を仮定し，その変数を推定することで次元削減を行う．このカテゴリに関しても，Probabilistic PCA [6] をはじめ，多数の手法が提案されている [7]．特に，ニューラルネットワーク (NN) で観測値ベクトルを圧縮する，Variational Autoencoder (VAE) [8] のような深

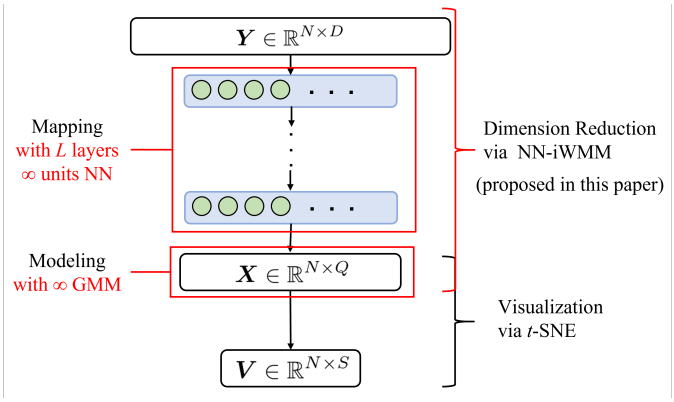


図 1: NPDV の次元削減と可視化フロー.

層潜在変数モデル (Deep Latent Variable Model:DLVM) [9]〜[11] は (1) の課題に対処する方法になり得る．DLVM では，NN の強力な非線形変換により，観測値ベクトルよりも正確に個体の特徴を反映する潜在表現を得られる可能性がある．実際に，教師無しクラスタリングでは，観測値ベクトルを NN で圧縮して得られた潜在表現をクラスタリングすることで，観測値ベクトルを直接クラスタリングするよりも，精度が大きく改善した例が報告されている [12], [13]．そのため，観測値のベクトル表現が原因で (1) の手法が機能しないとき，DLVM は有効な代替案となり得る．一方で，DLVM の性能を最適化するには，ユニット数などの NN のモデル構造をはじめ，多くの超パラメータをチューニングする必要がある．チューニングでは，超パラメータ毎に設定された探索範囲の中で，指標が最良となる値の組み合わせを決定する．その組み合わせ数は，超パラメータ数に対して指数的に増加するため，DLVM はチューニングに多くの試行を要する．また，各試行で NN の大量のパラメータを学習するため，DLVM のチューニングは多くの時間を必要とする．加えて，超パラメータの探索範囲は分析者が設定するため，その範囲が変わると可視化結果が変化する．近年では，超パラメータをデータから推定するノンパラメトリックベイズモデル [14] と DLVM を組み合わせた Nonparametric Bayesian DLVM (NP-DLVM) [15], [16] が提案されているが，依然として多くの超パラメータやパラメータを持つ．例えば，最新の NP-DLVM である VSB-DVM [17] は，潜在変数の事前分布に無限混合ガウス分布 [18] を仮定し，潜在空間内の分布の混合数をデータから推定する．一方で，観測値のエンコーダーなど，モデルが複数のネットワークで構成されており，それぞれのレイヤー数とユニット数が超パラメータとなる．また，大量の重みやバイアスを推定する必要があり，その際の学習率やその減衰率も分析者が指定する．

本稿では，これらの問題に対処するため，Nonparametric Bayesian Deep Visualization (NPDV) を提案する．その次元削減と可視化フローを図 1 に示す．NPDV は，本稿で導入する新たな DLVM である Neural-Network infinite Warped Mixture Model (NN-iWMM) による観測値 $\boldsymbol{Y}$ から潜在変数 $\boldsymbol{X}$ への次元削減，$\boldsymbol{X}$ を入力とした t-SNE による可視化表現 $\boldsymbol{V}$ の推定という，次元削減，可視化を統合した確率モデルである．NN-iWMM では，VSB-DVM と同様に無限混合ガウス分布から潜在変数を生成するが，観測値は NNGP カーネル [19] にもとづくガウス過程で生成される．このガウス過程は，ユニット数を考慮する必要がない，L 層無限ユニット NN と等価な非線形変換を行う．そのため，NN-iWMM は，無限混合分布による潜在変数のモデル化，無限ユニット NN による非線形変換という二つのノンパラメトリック性を持つ DLVM となる．また，t-SNE は，一つの超パラメータのみで可視化表現を推定する．NPDV は，両者を統合することで，NN を利用するにもかかわらず，既存の DLVM よりも少ない超パラメータで次元削減，可視化を行うことができる．さらに，NNGP カーネルは，2 個のパラメータのみで L 層 ∞ ユニット NN による非線形変換を定義する．そのため，NPDV は，通常の NN を使用する既存手法と比較して，パラメータ数も少ないモデルになっている．以下では，提案手法の詳細を説明した後，実証実験の結果を紹介し，まとめを述べる．

2. 予 備 知 識

本節では，NPDV の前提となる t-SNE [3] と infinite Warped Mixture Model (iWMM) [20] を紹介する．以下では，D 次元の N 個の観測値を $\boldsymbol{Y} \in \mathbb{R}^{N\times D}$，対応する Q 次元の潜在変数を $\boldsymbol{X} \in \mathbb{R}^{N\times Q}$，可視化表現を $\boldsymbol{V} \in \mathbb{R}^{N\times S}$，$S \in 2, 3$ とする．

2.1 t-SNE

t-SNE は，類似度ベース次元削減の一つであり，$\boldsymbol{Y}$，$\boldsymbol{V}$ 双方の類似度を確率分布を用いて評価する．観測値 $\boldsymbol{y}_i$， $\boldsymbol{y}_j$ の類似度 p^Y_{ij} は，ガウス分布を用いて，次式で与えられる．

$$p^Y_{j|i} = \frac{\exp(-\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2/2\tau_i^2)}{\sum_{\ell\neq i}\exp(-\|\boldsymbol{y}_i - \boldsymbol{y}_\ell\|^2/2\tau_i^2)}, \quad p_{ij} = \frac{p^Y_{i|j} + p^Y_{j|i}}{2N} \tag{1}$$

ガウス分布の分散 τ_i^2 は，パープレキシティ ρ を元にしたバイナリサーチにより決定される．可視化表現 $\boldsymbol{v}_i$，$\boldsymbol{v}_j$ の類似度 p^V_{ij} は，t 分布を用いて，次式で与えられる

$$p^V_{ij} = \frac{(1 + \|\boldsymbol{v}_i - \boldsymbol{v}_j\|^2)^{-1}}{\sum_{\ell,s,\ell\neq s}(1 + \|\boldsymbol{v}_\ell - \boldsymbol{v}_s\|^2)^{-1}} \tag{2}$$

そして，$\boldsymbol{p}^Y = \{p^Y_{ij}\}_{i,j}$，$\boldsymbol{p}^V = \{p^V_{ij}\}_{i,j}$ 間の KL 情報量

$$\mathrm{KL}[\boldsymbol{p}^Y\|\boldsymbol{p}^V] = \sum_{i,j,i\neq j} p^Y_{ij}\log\frac{p^Y_{ij}}{p^V_{ij}} \tag{3}$$

を最小化することで $\boldsymbol{V}$ を推定する．

2.2 infinite Warped Mixture Model

iWMM は，ガウス過程により潜在変数を観測空間へ写像するガウス過程潜在変数モデル (GPLVM) [21] から派生したノンパラメトリックベイズモデルである．GPLVM では，潜在変数 $\boldsymbol{X}$ にカーネル関数 $k(\boldsymbol{x}, \boldsymbol{x}')$ を適用して得られるグラム行列 $K_{NN} \in \mathbb{R}^{N\times N}$ と，精度 $\beta \in \mathbb{R}$ から定義される共分散から，$\boldsymbol{Y}$ の列ベクトル $\boldsymbol{y}_d \in \mathbb{R}^N$ を次元 d 毎に独立に生成する．なお，NN は行列のサイズを表す．$\boldsymbol{X}$ を所与としたときの $\boldsymbol{Y}$ の 条件付き分布は，次式で定義される．

$$p(\boldsymbol{Y}|\boldsymbol{X}) = \prod_{d=1}^{D} \mathcal{N}(\boldsymbol{y}_d|\boldsymbol{0}, K_{NN} + \beta^{-1}\boldsymbol{I}_N) \tag{4}$$

$\mathcal{N}(\boldsymbol{m}, \boldsymbol{C})$ は平均 $\boldsymbol{m}$，共分散 $\boldsymbol{C}$ を持つガウス分布を，$\boldsymbol{I}_N$ は N 次元単位行列を表す．$\boldsymbol{X}$ は未観測の確率変数であり，その事前分布を設定できる．[21] では $\boldsymbol{X}$ の事前分布に標準ガウス分布を仮定するが，iWMM では以下の無限混合ガウス分布を仮定する．

$$p(\boldsymbol{X}) = \sum_{k=1}^{\infty} \pi_k \mathcal{N}(\boldsymbol{m}_k, \boldsymbol{R}_k^{-1}) \tag{5}$$

π_k, $\boldsymbol{m}_k$, $\boldsymbol{R}_k$ はそれぞれ k 番目の分布の混合比率, 平均, 精度行列を表す．本稿では $\boldsymbol{R}_k$ が対角行列であると仮定する．iWMM では式 (5) の分布で $\boldsymbol{X}$ を生成し，それを入力として K_{NN} を評価した後，式 (4) の分布から $\boldsymbol{Y}$ を生成する．

3. Nonparametric Bayesian Deep Visualization

本節では，提案手法である NPDV の定式化について紹介する．NPDV は，本稿で導入する新たな NP-DLVM である NN-iWMM による $\boldsymbol{Y}$ から潜在変数 $\boldsymbol{X}$ への圧縮，$\boldsymbol{X}$ を入力とする t-SNE による可視化表現 $\boldsymbol{V}$ の推定という二つの圧縮を，Regularized Bayesian Inference [22] の枠組みで統合した確率モデルである．前半部では NN-iWMM の定式化を，後半部では NN-iWMM と t-SNE を統合する方法を紹介する．

3.1 NN-iWMM

ガウス過程はカーネル関数を変更することで様々な非線形関数を表現できる．[19] は L 層 ∞ ユニット NN と等価なガウス過程モデルを構築できる NNGP カーネルを提案した．いま，隠れ層数が L の全結合 NN において，ℓ 層におけるバイアスの要素を b_i^ℓ，重み行列の要素を W_{ij}^ℓ，活性化関数を $\phi(\cdot)$ とする．また，任意の i, j, ℓ において，b_i^ℓ は $\mathcal{N}(0, \sigma_b^2/N_\ell)$ から，W_{ij}^ℓ は $\mathcal{N}(0, \sigma_w^2/N_\ell)$ から独立に生成されたと仮定する．N_ℓ は ℓ 層のユニット数を表す．このとき，入力 $\boldsymbol{x}_n$ から計算される ℓ 層の i 番目の出力 $a_i^\ell(\boldsymbol{x}_n)$ は，$\ell - 1$ 層でのアクティベーション後の値の線形結合として，以下の式で表すことができる．

$$a_i^\ell(\boldsymbol{x}_n) = b_i^\ell + \sum_{j=1}^{N_\ell} W_{ij}^\ell \phi(a_j^{\ell-1}(\boldsymbol{x}_n)) \tag{6}$$

$\{W_{ij}^\ell\}_{j=1}^{N_\ell}$ が独立同一分布に従うので，上式の和の中は i.i.d な確率変数の和となる．そのため，$N_\ell \to \infty$ とすると，中心極限定理から，$a_i^\ell(\boldsymbol{x}_n)$ はガウス分布に従う．これが任意の $\boldsymbol{x}_j$ で成り立つため，ℓ 層の J 個の出力 $\{a_i^\ell(\boldsymbol{x}_j)\}_{j=1}^J$ は多変量ガウス分布に従う．よって，ℓ 層の出力 $\{a_i^\ell(\boldsymbol{x}_n)\}_{n=1}^N$ はガウス過程 $\mathcal{GP}(0, K_{NN}^\ell)$ に従う確率変数となる．このとき，K_{NN}^ℓ の各要素 $K_{NN}^\ell(\boldsymbol{x}, \boldsymbol{x}')$ は次式で与えられる..

$$\begin{aligned} K_{NN}^\ell(\boldsymbol{x}, \boldsymbol{x}') &= \sigma_b^2 + \sigma_w^2 \mathbb{E}_{z_i^{\ell-1} \sim \mathcal{GP}(|\boldsymbol{0}, K_{NN}^{\ell-1})}[\Phi(\boldsymbol{x}, \boldsymbol{x}')] \\ \Phi(\boldsymbol{x}, \boldsymbol{x}') &\equiv \phi(z_i^{\ell-1}(\boldsymbol{x}))\phi(z_i^{\ell-1}(\boldsymbol{x}')) \end{aligned} \tag{7}$$

再帰式の一段目は，$K^0(\boldsymbol{x}, \boldsymbol{x}') = \sigma_b^2 + \sigma_w^2 \frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}'}{D}$ で与えられる．そして，上式に従って，グラム行列の各要素を L 回再帰的に計算することで，隠れ層数が L，ユニット数が ∞ の全結合 NN と等価なガウス過程モデルを構築できる．以下では，この再帰式で定義されるカーネル関数を NNGP カーネルと呼び，$k^L(\boldsymbol{x}, \boldsymbol{x}')$ と表す．また，$k^L(\boldsymbol{x}, \boldsymbol{x}')$ を使用して構築されたグラム行列を K_{NN}^L と表記する．

本稿で提案する NN-iWMM は，iWMM におけるグラム行列を K_{NN}^L とすることで，潜在変数を上記の L 層 ∞ ユニット NN と等価なガウス過程で観測値に変換する．その生成過程を図 2 に示す．まず棒折り過程 GEM(α) [23] からクラスター k の混合比率 π_k を生成し，対応するガウス分布の平均 $\boldsymbol{m}_k$ を $\mathcal{N}(\boldsymbol{m}_0, \boldsymbol{R}_0)$ から，$\boldsymbol{R}_k$ の対角成分 r_{kq} を $\mathrm{Gam}(a_0, b_0)$ から生成する．なお，本稿では $\alpha = 1$, $\boldsymbol{m}_0 = \boldsymbol{0}$, $\boldsymbol{R}_0 = \boldsymbol{I}_Q$, $a_0 = b_0 = 1$ で固定する．その後，$\{\pi_k\}_{k=1}^\infty$ をパラメータに持つカテゴリカル分布からクラスター割り当て z_n を生成した後，対応する分布から潜在変数 $\boldsymbol{x}_n$ をサンプリングする．そして，$\boldsymbol{X}$，NNGP カーネルの分散パラメータ σ_w，σ_b から K_{NN}^L を構築し，式 (4) で $\boldsymbol{Y}$ の列ベクトル $\boldsymbol{y}_d$ を生成する．上記手順で定義される NN-iWMM は，∞ 混合分布による潜在変数のモデル化，L 層 ∞ ユニット NN による $\boldsymbol{X}$ の変換という二つのノンパラメトリック性を持つ DLVM となる．また，NNGP を利用することで，重みやバイアスを用いる通常の NN と異なり，σ_w，σ_b という二つのパラメータのみで NN による変換を定義できる．

カーネル関数

K_{NN}^L は，式 (7) に期待値計算が含まれるため，一般の活性化関数に対して解析的に計算することはできない．しかし，恒等写像のような単純な関数や，ReLU 関数などの Polynomial rectified nonlinear functions 族に属する非線形関数を活性化関数として使用する場合，その値を解析的に評価できることが知られている [24]．本稿では，解析的に K_{NN}^L を構築するため，恒等写像，ReLU 関数に対応するカーネル関数を用いる．恒等写像 $\phi(x) = x$ の場合は，

$$\mathbb{E}[\phi\phi'] = K^{\ell-1}(\boldsymbol{x}, \boldsymbol{x}') \tag{8}$$

ReLU 関数 $\phi(x) = \max\{0, x\}$ の場合は，

$$\begin{aligned} K^\ell(\boldsymbol{x}, \boldsymbol{x}') = \sigma_b^2 \quad &+ \frac{\sigma_w^2}{2\pi}\sqrt{K^{\ell-1}(\boldsymbol{x}, \boldsymbol{x})K^{\ell-1}(\boldsymbol{x}', \boldsymbol{x}')} \\ &\times \left(\sin\theta_{\boldsymbol{x},\boldsymbol{x}'}^{\ell-1} + (\pi - \theta_{\boldsymbol{x},\boldsymbol{x}'}^{\ell-1})\cos\theta_{\boldsymbol{x},\boldsymbol{x}'}^{\ell-1}\right) \\ \theta_{\boldsymbol{x},\boldsymbol{x}'}^{\ell-1} &= \cos^{-1}\left(\frac{K^{\ell-1}(\boldsymbol{x}, \boldsymbol{x}')}{\sqrt{K^{\ell-1}(\boldsymbol{x}, \boldsymbol{x})K^{\ell-1}(\boldsymbol{x}', \boldsymbol{x}')}}\right) \end{aligned} \tag{9}$$

で各要素の再帰計算を行う．

3.2 NN-iWMM と t-SNE の統合

NPDV を定式化するには，確率モデルである NN-iWMM と，非確率的手法である t-SNE という設計方針の異なる二つの方法を統合する必要がある．以下では，Regularized Bayesian Inference (RegBayes) [22] の枠組みで両者を確率モデルに統合する方法を紹介する．RegBayes では，ベイズ公式を最適化問題として再定式化することで，事後分布に関する制約を考慮し

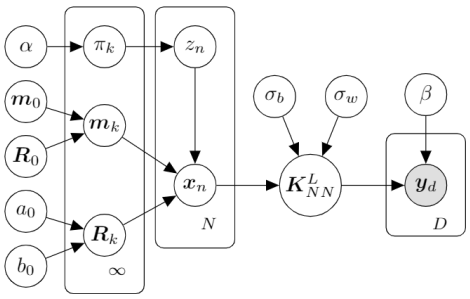


図 2: NN-iWMM の生成過程.

たベイズモデルの作成を可能にする．パラメータを $\boldsymbol{\theta}$，その事前分布を $p(\boldsymbol{\theta})$，観測値を $\boldsymbol{Y}$，尤度を $p(\boldsymbol{Y}|\boldsymbol{\theta})$ としたとき，その事後分布はベイズ公式により $p(\boldsymbol{\theta}|\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ で与えられる．[25] は，$p(\boldsymbol{\theta}|\boldsymbol{Y})$ が $p(\boldsymbol{Y}|\boldsymbol{\theta})$，$p(\boldsymbol{\theta})$ から成る以下の制約付き最適化問題の解と一致することを示した．

$$\begin{cases} \min\limits_{q(\boldsymbol{\theta})} & \mathrm{KL}[q(\boldsymbol{\theta})\|p(\boldsymbol{\theta})] - \displaystyle\int q(\boldsymbol{\theta}) \log p(\boldsymbol{Y}|\boldsymbol{\theta}) d\boldsymbol{\theta} \\ \text{s.t.} & q(\boldsymbol{\theta}) \in \mathcal{P} \end{cases} \tag{10}$$

$\mathcal{P}$ は確率分布の集合を表す．RegBayes では，式 (10) に $\boldsymbol{\theta}, \boldsymbol{Y}$ から構成される正則化項 $\mathcal{R}(\boldsymbol{\theta}, \boldsymbol{Y})$ の期待値 $E_{q(\boldsymbol{\theta})}[\mathcal{R}(\boldsymbol{\theta}, \boldsymbol{Y})]$ に関する制約を加えた，以下の最適化問題を考える．

$$\begin{cases} \min\limits_{q(\boldsymbol{\theta})} & \mathrm{KL}[q(\boldsymbol{\theta})\|p(\boldsymbol{\theta})] - \displaystyle\int q(\boldsymbol{\theta}) \log p(\boldsymbol{Y}|\boldsymbol{\theta}) d\boldsymbol{\theta} \\ \text{s.t.} & E_{q(\boldsymbol{\theta})}[\mathcal{R}(\boldsymbol{\theta}, \boldsymbol{Y})] \leq 0,\ q(\boldsymbol{\theta}) \in \mathcal{P} \end{cases} \tag{11}$$

制約条件を踏まえたときの最適分布 $q^*(\boldsymbol{\theta})$ は変分法により求めることができて，次式で与えられる．

$$q^*(\boldsymbol{\theta}) \propto p(\boldsymbol{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \exp(-\lambda \mathcal{R}(\boldsymbol{\theta}, \boldsymbol{Y})) \tag{12}$$

λ はラグランジュ未定乗数を表す．上式より，$q^*(\boldsymbol{\theta})$ に対応する $\boldsymbol{\theta}, \boldsymbol{Y}$ の非正規化同時分布はその右辺で定義できる．

NPDV では, 式 (12) で定義される確率モデルにおいて, $p(\boldsymbol{\theta})$ が $p(\boldsymbol{X}, \boldsymbol{z}, \{\boldsymbol{m}_k, \boldsymbol{r}_k, \pi_k\}_{k=1}^{\infty})$ に，$p(\boldsymbol{Y}|\boldsymbol{\theta})$ が $p(\boldsymbol{Y}|\boldsymbol{X})$ に対応する．$\boldsymbol{z} = \{z_n\}_{n=1}^N$ は潜在クラスタ割り当ての集合を，$\boldsymbol{r}_k = \{r_{kq}\}_{q=1}^Q$ は k 番目の分布の精度行列 $\boldsymbol{R}_k$ の対角成分を表す．NPDV では t-SNE で $\boldsymbol{X}$ を $\boldsymbol{V}$ に圧縮するため，制約条件 $\mathcal{R}(\boldsymbol{Y}, \boldsymbol{\theta})$ は $\boldsymbol{X}$，$\boldsymbol{V}$ 間の t-SNE のコスト関数 $\mathrm{KL}[\boldsymbol{p}^X\|\boldsymbol{p}^V]$ に対応する．$\boldsymbol{V}$ はカーネル関数のパラメータ σ_w^2，σ_b^2，$p(\boldsymbol{Y}|\boldsymbol{X})$ の精度 β と同じく，確率分布を仮定しない．さらに，NN-iWMM の生成過程における条件付き独立性から，NPDV の非正規化同時分布は,

$$\begin{aligned} &p(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{z}, \{\boldsymbol{m}_k, \boldsymbol{R}_k \pi_k\}_{k=1}^{\infty} | \boldsymbol{V}) \\ &\propto p(\boldsymbol{Y}|\boldsymbol{X}) p(\boldsymbol{X}|\boldsymbol{z}, \{\boldsymbol{m}_k, \boldsymbol{r}_k, \pi_k\}_{k=1}^{\infty}) p(\boldsymbol{z}|\{\pi_k\}_{k=1}^{\infty}) \\ &\times p(\{\boldsymbol{m}_k\}_{k=1}^{\infty}) p(\{\boldsymbol{r}_k\}_{k=1}^{\infty}) p(\{\pi_k\}_{k=1}^{\infty}) \\ &\times \exp(-\lambda \mathrm{KL}[\boldsymbol{p}^X\|\boldsymbol{p}^V]) \end{aligned} \tag{13}$$

となる．簡単のため $\boldsymbol{V}$ 以外のパラメータは省略した．λ は，NN-iWMM の尤度関数と t-SNE のコスト関数をバランスする項であり，本研究では超パラメータとして扱う．

4. 変分ベイズ法による学習アルゴリズム

提案手法では，変分ベイズ法を用いて学習を行う．前半部では，学習の目的関数となる周辺対数尤度 $\log p(\boldsymbol{Y}|\boldsymbol{V})$ の下界 (Evidence Lower BOund : ELBO) の導出と評価方法を紹介する．その後，学習アルゴリズムの要約と，既存の NP-DLVM と比較したときの利点を示す.

4.1 ELBO の導出

変分ベイズ法では，変分分布 $\mathcal{Q}$ と Jensen の不等式から導出される式 (13) の ELBO

$$\mathcal{L} = \mathbb{E}_{\mathcal{Q}}\left[\frac{p(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{z}, \{\boldsymbol{m}_k, \boldsymbol{r}_k, \pi_k\}_{k=1}^{\infty}, |\boldsymbol{V})}{\mathcal{Q}}\right] \tag{14}$$

を最大化するようにパラメータを推定する．NPDV は潜在変数の事前分布に無限混合モデルを持つが, [18] に倣って，混合モデルに関する変分分布に有限混合ガウスモデルを仮定する．また，$\mathcal{Q}$ には平均場仮定を与えた以下の分布を設定する．

$$\begin{aligned} \mathcal{Q} &= q(\boldsymbol{X}, \boldsymbol{z}, \{\boldsymbol{m}_k, \boldsymbol{r}_k, \pi_k\}_{k=1}^K) \\ &= q(\boldsymbol{x}_n) q(z_n) \prod_{k=1}^K q(\pi_k) q(\boldsymbol{m}_k) q(\boldsymbol{r}_k) \prod_{i=1}^N \end{aligned} \tag{15}$$

なお, 混合モデルに関する変分分布 $q(\pi_k)$, $q(\boldsymbol{m}_k)$, $q(\boldsymbol{r}_k)$, $q(z_n)$ の定義は [18] と同一である．$q(\boldsymbol{x}_n) = \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{S}_n)$ であり，$\boldsymbol{S}_n$ を対角行列とする．$\mathcal{Q}$ の平均場仮定と，NN-iWMM の条件付き独立性から，式 (14) は次の 4 項に分解できる．

$$\begin{aligned} \mathcal{L} &= \mathbb{E}_{q(\boldsymbol{X})}[\log p(\boldsymbol{Y}|\boldsymbol{X})] - \mathbb{E}_{q(\boldsymbol{X})}[\log q(\boldsymbol{X})] \\ &+ \mathbb{E}_{q(\boldsymbol{X}, \boldsymbol{z}, \boldsymbol{m}, \boldsymbol{\Sigma}, \boldsymbol{\phi})}\left[\frac{\log p(\boldsymbol{X}, \boldsymbol{z}, \{\boldsymbol{m}_k, \boldsymbol{r}_k, \pi_k\}_{k=1}^K)}{q(\boldsymbol{z}, \{\boldsymbol{m}_k, \boldsymbol{R}_k\}_k, \boldsymbol{\pi})}\right] \\ &- \lambda \mathbb{E}_{q(\boldsymbol{X})}[\mathrm{KL}[\boldsymbol{p}^X\|\boldsymbol{p}^V]] \\ &= \mathcal{L}_1 + \mathcal{H}(q(\boldsymbol{X})) + \mathcal{L}_2 - \lambda \mathcal{R} \end{aligned} \tag{16}$$

$\mathcal{H}(q(\boldsymbol{X}))$ は正規分布のエントロピーであり，解析的に評価できる．以下では，$\mathcal{L}_1$，$\mathcal{L}_2$，$\mathcal{R}$ の評価方法について紹介する．

$\mathcal{L}_1$ の評価に関して，$\boldsymbol{Y}$ が次元 d 毎に独立なので，$\mathcal{L}_1$ は d に関する和に分解できる．

$$\mathcal{L}_1 = \sum_{d=1}^D \int q(\boldsymbol{X}) \log p(\boldsymbol{y}_d|\boldsymbol{X}) d\boldsymbol{X} = \sum_{d=1}^D \mathcal{L}_1^{(d)} \tag{17}$$

$\mathcal{L}_1^{(d)}$ に含まれる積分は解析解を持たないため，数値積分が必要になる．数値積分を行う場合，$\log p(\boldsymbol{y}_d|\boldsymbol{X})$ の評価に計算量が $\mathcal{O}(N^3)$ となる K_{NN}^L の逆行列計算が含まれるため，N が大きいとき学習に多大な時間を必要とする．そこで, [26] で導入された GPLVM に対する誘導点法を利用し、逆行列の計算量を削減した上で $\mathcal{L}_1^{(d)}$ を近似する. [26] は，潜在空間で定義される $M(\ll N)$ 個の擬似入力 $\boldsymbol{\zeta} \in \mathbb{R}^{M \times Q}$ と，対応するガウス過程の出力である誘導点 $\boldsymbol{v}_d \in \mathbb{R}^M$ を用いて，$\mathcal{L}_1^{(d)}$ の下限が次式で与えられることを示した.

$$\begin{aligned} \mathcal{L}_1^{(d)} &\geq \log\left[\frac{\beta^{\frac{N}{2}} |K_{MM}^L|^{\frac{1}{2}}}{(2\pi)^{\frac{N}{2}} |\beta \boldsymbol{\Psi}_2 + K_{MM}^L|^{\frac{1}{2}}} e^{-\frac{1}{2} \boldsymbol{y}_d^{\mathrm{T}} \boldsymbol{G} \boldsymbol{y}_d}\right] \\ &- \frac{\beta \psi_0}{2} + \frac{\beta}{2} \mathrm{tr}((K_{MM}^L)^{-1} \boldsymbol{\Psi}_2) \end{aligned} \tag{18}$$

K^L_{MM} は $\boldsymbol{\zeta}$ 間で計算される $M\times M$ 行列を，K^L_{MN} は $\boldsymbol{\zeta}, \boldsymbol{X}$ 間で計算される $M\times N$ 行列を，K^L_{NM} は K^L_{MN} の転置行列を表す．これらの行列はいずれも NNGP カーネルから計算される．また，$\boldsymbol{G}=\beta\boldsymbol{I}_N-\beta^2\boldsymbol{\Psi}_1(\beta\boldsymbol{\Psi}_2+K^L_{MM})^{-1}\boldsymbol{\Psi}_1^{\mathrm{T}}$, $\psi_0=\mathrm{tr}(\mathbb{E}_{q(\boldsymbol{X})}[K^L_{NN}])$, $\boldsymbol{\Psi}_1=\mathbb{E}_{q(\boldsymbol{X})}[K^L_{MN}]$, $\boldsymbol{\Psi}_2=\mathbb{E}_{q(\boldsymbol{X})}[K^L_{MN}K^L_{NM}]$ である．式 (18) では，右辺の第 3 項と $\boldsymbol{G}$ に逆行列計算が含まれるが，そのサイズはともに $M\times M$ である．そのため，上式で $\mathcal{L}_1^{(d)}$ を近似することで，逆行列の計算量を $\mathcal{O}(N^3)$ から $\mathcal{O}(M^3)$ に削減できる．次に，ψ_0，$\boldsymbol{\Psi}_1$，$\boldsymbol{\Psi}_2$ の評価方法を説明する．ψ_0，$\boldsymbol{\Psi}_1$ の (n,m) 成分，$\boldsymbol{\Psi}_2$ の (m,m) 成分は，n に関する和に分解できる．

$$\begin{aligned}\psi_0&=\textstyle\sum_{n=1}^N\mathbb{E}_{q(\boldsymbol{X})}[k^L(\boldsymbol{x}_n,\boldsymbol{x}_n)]\\(\boldsymbol{\Psi}_1)_{nm}&=\textstyle\sum_{n=1}^N\mathbb{E}_{q(\boldsymbol{X})}[k^L(\boldsymbol{x}_n,\boldsymbol{\zeta}_m)]\\(\boldsymbol{\Psi}_2)_{mm'}&=\textstyle\sum_{n=1}^N\mathbb{E}_{q(\boldsymbol{X})}[k^L(\boldsymbol{x}_n,\boldsymbol{\zeta}_m)k^L(\boldsymbol{\zeta}_{m'},\boldsymbol{x}_n)]\end{aligned}\tag{19}$$

和中の期待値は，ガウス分布に対する Reparameterization trick [8] を用いて，モンテカルロ近似する．モンテカルロ標本 $\widetilde{\boldsymbol{x}}_n$ は，$q(\boldsymbol{x}_n)$ の平均 $\boldsymbol{S}$，分散 $\boldsymbol{S}_n$ を用いて次式で生成される．

$$\widetilde{\boldsymbol{x}}_n=\boldsymbol{\mu}_n+\boldsymbol{S}_n\boldsymbol{\epsilon}_n,\ \boldsymbol{\epsilon}_n\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_Q)\tag{20}$$

そして，得られたモンテカルロ標本 $\widetilde{\boldsymbol{X}}=\{\widetilde{\boldsymbol{x}}_n\}_{n=1}^N$ を式 (19) に代入し，近似値 $\widetilde{\psi}_0$，$\widetilde{\boldsymbol{\Psi}}_1$，$\widetilde{\boldsymbol{\Psi}}_2$ を得る．その後，これらの量を式 (18) に代入し，$\mathcal{L}_1^{(d)}$ の近似値 $\widetilde{\mathcal{L}}_1^{(d)}$ を得る．$\mathcal{L}_1$ は $\widetilde{\mathcal{L}}_1^{(d)}$ を用いて次式で近似される．

$$\widetilde{\mathcal{L}}_1=\sum_{d=1}^D\widetilde{\mathcal{L}}_1^{(d)}\tag{21}$$

次に $\mathcal{L}_2$ の評価方法を説明する．$\mathcal{L}_2$ は，$\widetilde{\boldsymbol{X}}$ を所与とすると，

$$\widetilde{\mathcal{L}}_2=\mathbb{E}_{q(\boldsymbol{z},\boldsymbol{m},\boldsymbol{\Sigma},\boldsymbol{\phi})}\left[\frac{\log p(\widetilde{\boldsymbol{X}},\boldsymbol{z},\boldsymbol{m},\boldsymbol{\Sigma},\boldsymbol{\phi})}{q(\boldsymbol{z},\boldsymbol{m},\boldsymbol{\Sigma},\boldsymbol{\phi})}\right]\tag{22}$$

となる．これは，$\widetilde{\boldsymbol{X}}$ を観測と見なせば，[18] 中の式 (10) で定義される ELBO と一致する．具体形は [18] を参照されたい．

最後に $\mathcal{R}$ の評価方法について紹介する．$\mathcal{R}$ も，$\mathcal{L}_1$，$\mathcal{L}_2$ の場合と同様に，$\widetilde{\boldsymbol{X}}$ を用いて近似する．$\mathcal{R}$ では，$\boldsymbol{p}^V$ が $q(\boldsymbol{X})$ に関して定数となるため，$\boldsymbol{p}^X$ の要素 p^X_{ij} の近似のみを考えれば良い．このとき，p^X_{ij} は式 (1) から

$$\widetilde{p}^X_{j|i}=\frac{\exp(-\|\widetilde{\boldsymbol{x}}_i-\widetilde{\boldsymbol{x}}_j\|^2/2\tau_i^2)}{\sum_{\ell\neq i}\exp(-\|\widetilde{\boldsymbol{x}}_i-\widetilde{\boldsymbol{x}}_\ell\|^2/2\tau_i^2)},\ \widetilde{p}^X_{ij}=\frac{\widetilde{p}^X_{i|j}+\widetilde{p}^X_{j|i}}{2N}\tag{23}$$

と近似できる．そして，$\{\widetilde{p}^X_{ij}\}_{i,j}$ を用いて $\mathcal{R}$ を次式で近似する．

$$\widetilde{\mathcal{R}}=\sum_{i,j,i\neq j}\widetilde{p}^X_{ij}\log\frac{\widetilde{p}^X_{ij}}{p^V_{ij}}\tag{24}$$

その後，$\widetilde{\mathcal{L}}_1$，$\widetilde{\mathcal{L}}_2$，$\widetilde{\mathcal{R}}$ を式 (16) に代入して $\mathcal{L}$ を近似する．

4.2 学習アルゴリズムの要約

学習アルゴリズムの概要を図 3 に示す．NPDV の学習に先立って，ガウス分布の混合数を 1 とする NN-iWMM の学習を行い，$q(\boldsymbol{X})$ の変分パラメータ $\{\boldsymbol{\mu}_n,\boldsymbol{S}_n\}_{n=1}^N$，疑似入力 $\boldsymbol{\zeta}$，カーネル関数のパラメータ σ_w^2，σ_b^2，精度 β を初期化する．その後，$\{\boldsymbol{\mu}_n\}_{n=1}^N$ に t-SNE を適用し，可視化表現 $\boldsymbol{V}$ を初期化する．

Algorithm 1 Learning algorithm of NPDV

Input observations $\boldsymbol{Y}$
`/* 1. Pre-training */`
Initialize $\boldsymbol{\Pi}_0=[\{\boldsymbol{\mu}_n,\boldsymbol{S}_n\}_{n=1}^N,\boldsymbol{\zeta},\sigma_b^2,\sigma_w^2,\beta]$
for i=1,2,... **do**
  update $\boldsymbol{\Pi}_0$ with a grandient-based method
end for
Obtain initial $\boldsymbol{V}$ by applying t-SNE to $\{\boldsymbol{\mu}_n\}_{n=1}^N$
`/* 2. Training NPDV */`
Initialize $\boldsymbol{\Pi}_1=\{\{\boldsymbol{m}_k,\boldsymbol{r}_k,\pi_k\}_{k=1}^K,\boldsymbol{z}\}$
for i=1,2,... **do**
  Generate $\widetilde{\boldsymbol{X}}$ with Eq.(20)
  Approximate ψ_0, $\boldsymbol{\Psi}_1$, $\boldsymbol{\Psi}_2$, $\{p^{\boldsymbol{X}}_{ij}\}_{i,j}$ based on $\widetilde{\boldsymbol{X}}$
  Update $\boldsymbol{\Pi}_1$ with VB-EM in [18]
  Update $\boldsymbol{\Pi}_2=\{\boldsymbol{\mu}_n,\boldsymbol{S}_n\}_{n=1}^N,\boldsymbol{\zeta},\boldsymbol{V},\sigma_b^2,\sigma_w^2,\beta\}$ with a gradient-based method
end for

図 3: NPDV の学習アルゴリズム．

NPDV の学習では，まず式 (20) に従ってモンテカルロ標本 $\widetilde{\boldsymbol{X}}$ を生成し，ELBO の評価に必要な統計量を近似する．NPDV の学習にあたって，推定する変数を，混合モデルに関する $\boldsymbol{\Pi}_1=[\boldsymbol{z},\{\boldsymbol{m}_k,\boldsymbol{r}_k,\pi_k\}_{k=1}^K]$ と，それ以外の $\boldsymbol{\Pi}_2=[\{\boldsymbol{\mu}_n,\boldsymbol{S}_n\}_{i=1}^N,\boldsymbol{\zeta},\boldsymbol{V},\sigma_b^2,\sigma_w^2,\beta]$ に分割する．$\boldsymbol{\Pi}_1$ は，モンテカルロ標本を所与としたときに，$\widetilde{\mathcal{L}}_2$ が [18] の学習の目的関数と一致するので，変分ベイズ EM アルゴリズム (VB-EM) により値が更新できる．VB-EM による更新では，勾配法による更新とは異なり，ELBO を最大化する位置に $\boldsymbol{\Pi}_1$ を遷移させることができる．$\boldsymbol{\Pi}_2$ は通常の NN と同じく勾配法で更新する．

4.3 提案手法の超パラメータ数

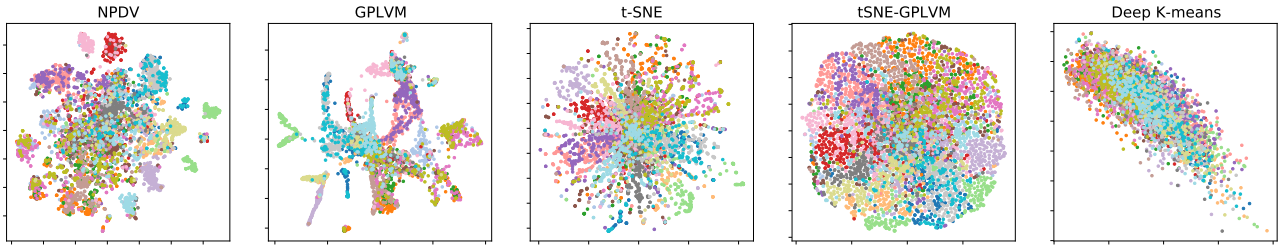
本節では，最新の NP-DLVM である VSB-DVM [17] と提案手法の超パラメータ数，パラメータ数を比較し，提案手法の利点を述べる．なお，両者はともに潜在変数を無限混合ガウス分布でモデル化しているため，以下ではそれ以外の点を比較する．

VSB-DVM の学習は，観測値のエンコーダー，デコーダーに加え，$\boldsymbol{x}_n$ のクラスター所属確率を計算する際に，Stick-Breaking ネットワークというもう一つの NN が使用される．エンコーダー，デコーダー，Stick-Breaking ネットワークの層数をそれぞれ L_1，L_2，L_3 とすると，合計で $\sum_{i=1}^3 L_i$ 層分のユニット数を設定する必要がある．また，層数 $\{L_i\}_{i=1}^3$ 自身も超パラメータであり，モデル構造に関する超パラメータ数は $3+\sum_{i=1}^L L_i$ となる．加えて，学習率 η，その減衰率 ϵ も分析者が指定するため，学習に必要な超パラメータは $5+\sum_{i=1}^3 L_i$ 個となる．その学習では，3 個のネットワークの重み，バイアスが推定され，パラメータ数の合計が数百万以上になることも多い．

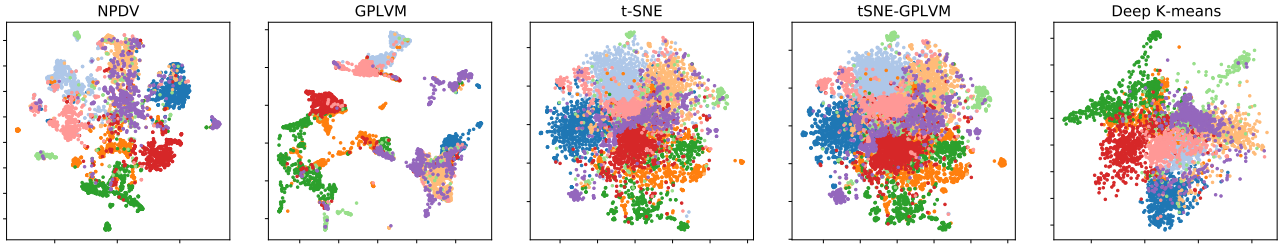
NPDV の超パラメータは，NN-iWMM における層数 L，$\boldsymbol{X}$ の次元数 Q，バランス項 λ，誘導点の数 M，t-SNE のパープレキシティ ρ，学習率 η の 6 個である．その内の ρ はモデル性能に関する感度が低いことが知られている [3]．また，M，η

表 1: k 近傍による推定精度．各手法の推定結果は確率的に変動するため，乱数シードを変えた上で 5 回実験を行った．表 1 には，5 回の平均値を記載している．太字の数値は最高精度を，*がつく数値は次点の精度を表す．

	MNIST			**20 Newsgroups**			**livedoor**		
	k=10	k=20	k=30	k=10	k=20	k=30	k=10	k=20	k=30
t-SNE	**0.931**	0.921*	0.913*	0.559	0.512	0.492	0.793	0.767	0.747
UMAP	0.921	0.921	0.913	0.339	0.264	0.230	0.540	0.495	0.463
GPLVM	0.798	0.787	0.781	0.527	0.496	0.487	0.809*	0.798*	0.790*
t-SNE-GPLVM	0.930*	0.919	0.931	0.593*	0.529*	0.500*	0.793	0.765	0.747
Deep k-means	0.817	0.803	0.795	0.383	0.347	0.330	0.755	0.746	0.742
VSB-DVM	0.864	0.850	0.838	0.051	0.052	0.052	0.116	0.116	0.119
NPDV	**0.931**	**0.923**	**0.918**	**0.599**	**0.552**	**0.529**	**0.844**	**0.822**	**0.809**



(a) **20 Newsgroups** の可視化結果．各点の色は 20 個のラベルに対応している．



(b) **livedoor** 可視化結果．各点の色は 7 個のラベルに対応している．

図 4: NPDV および各手法による可視化結果の比較．

は，予備実験で $M = 50$，$\eta = 0.01$ で良好な性能が得られている．そのため，調整が必要な超パラメータは L，Q，λ の 3 個であり，VSB-DVM と比較して超パラメータ数が少ない．また，NN-iWMM では，NNGP カーネルを利用することで，σ_w，σ_b という 2 つのパラメータで NN による変換を定義できる．観測値の精度 β を加えても，NN-iWMM のパラメータ数は 3 個である．それらに加えて，可視化表現 $\boldsymbol{V} \in \mathbb{R}^{N\times S}$ は NS 個のパラメータを含むので，NPDV の合計パラメータ数は $3 + NS$ 個となる．可視化では，S は 2，または 3 であるため，多くの場合，数百万個以上のパラメータを持つ VSB-DVM よりも，NPDV のパラメータ数は少なくなる．

モデル性能の最適化には超パラメータのチューニングが必要だが，超パラメータが多いとき，チューニングに必要な試行回数は増加する．また，各試行の中で，NN 内の大量のパラメータを学習することになり，VSB-DVM の学習は計算時間が増大しやすい．一方で提案手法は，VSB-DVM と比較して超パラメータ，パラメータが少ないため，より短時間で最適化されたモデルによる可視化を行える可能性がある．

5. 実証実験

本節では，提案手法の性能確認のために行った実験の結果を報告する．前半部では，ラベルが付与されているデータセットの可視化実験により，他モデルと提案手法の性能比較を行う．後半部では，ラベルがないデータセットの可視化を通じて，ラベルが利用できない状況での提案手法の有効性を検証する．

5.1 他モデルとの比較実験

5.1.1 実験設定

実験では，3 つのデータセットを使用した．**MNIST** は，手書き数字画像 6 万枚の構成されるデータセットで，各画像には 0 から 9 のいずれかがラベルとして与えられている．**20 Newsgroups** は，約 2 万件の英語記事から構成されいていて，20 個のラベルのいずれかが各記事に与えられている．各記事はストップワード(注1)を削除した後，lemmatization を行い語幹を抽出して解析に使用した．**livedoor** は，約 7 千件の日本語のブログの記事を含み，各記事には 7 個のラベルのいずれかが与えられている．各記事は，MeCab を用いて形態素解析を行い，名詞，形容詞を抽出して解析に使用した．また，**20 Newsgroups**，**livedoor** は，TF-IDF が大きくなった上位 1,000 語を抽出し，その TF-IDF 表現を観測値ベクトルとした．本実験では，3 データからそれぞれ 5,000 件を抽出して可視化を行った．

(注1)：https://gist.github.com/sebleier/554280

評価は，可視化空間における k 近傍法での分類精度による定量評価と，得られた散布図の定性評価を行った．k 近傍法での分類精度は，同一ラベルを持った個体が互いに近くにプロットされているほど向上する．そのため，この指標が高いことは，モデルがラベルの違いを反映した可視化表現を推定できたことを意味する．近傍数は $k = [10, 20, 30]$ とした．

また，モデル比較のため，t-SNE，UMAP，GPLVM，t-SNE-GPLVM [27]，VSB-DVM，Deep k-means [11] の 6 手法でも可視化を行った．t-SNE-GPLVM は，GPLVM の対数尤度に式 (3) で定義される t-SNE のコストを正則化項として加えた確率モデル，Deep k-means は最新の深層クラスタリングモデルである．NPDV における NNGP カーネルは，全データで共通で L=6 とし，式 (8)(9) を用いて，1，6 層目が恒等写像，それ以外の層が ReLU 関数に対応するように設計した．他の超パラメータも，全データ共通で，式 (16) 内のバランス項を $\lambda = ND$，$\boldsymbol{X}$ の次元数を $Q = 50$，誘導点の数を $M = 50$，GMM の混合数を $K = 50$，t-SNE のパープレキシティを $\rho = 30$，学習率を $\eta = 0.01$ とした．事前学習，学習の反復回数はともに 1500 回とした．

5.1.2 実験結果

表 1 に k 近傍法による分類精度の評価結果を示す．3 データセット全てで提案手法が最高精度を記録した．特に，TF-IDF 表現を用いる二つの文書データセットにおいて，**20 Newsgroups** では，$k = 20, 30$ のときに NPDV と次点の t-SNE の間に 2.5%，**livedoor** では $k = 10$ のときに次点の GPLVM と 3.5%，$k = 20$ のときに 2.5% と大きな差を記録している．ラベルの違いは個体の特徴の違いを表すため、提案手法が最も正確に個体の違いを反映した可視化を行っている．

20 Newsgroups，**livedoor** の可視化結果を図 3(a)，(b) に示す．散布図上の各点は推定された可視化表現を，点の色の違いはラベルの違いを表す．**20 Newsgroups** では，t-SNE，t-SNE-GPLVM，Deep k-means から得られた散布図では，同一ラベルを持つ点同士でクラスターを形成する一方，異なるクラスター同士が重なり合っている．特に Deep k-means においてその傾向が顕著である．GPLVM ではその三手法と比較すると，薄緑，薄紫のクラスターが他のクラスターとよく分離している．NPDV では，それらのクラスターに加えて赤，薄桃のクラスターがより確認しやすくなっている．**livedoor** でも同様に，t-SNE，t-SNE-GPLVM，Deep k-means ではクラスターが重なり合っている．GPLVM では，クラスターの分離は確認できるが，青，紫，薄橙のクラスターが混合している．NPDV では，GPLVM と比較して，その 3 つのクラスターの混合が改善されている．また，**MNIST** でも同様の傾向を確認した．

以上より，既存手法と比較して，提案手法が個体の特徴の違いをより正確に反映し，クラスター分離が明確な可視化表現を推定し得る，という有効性が示された．

5.2 教師無し設定での可視化実験

5.2.1 実験設定

実験には，ラベルが付与されていない約 4 万個の文を収録する京都大学コーパス [28] を用いた．本実験では，5,000 文を抽出して可視化を行った．文データは，一文に含まれる単語数が少ないものがあり，TF-IDF の性能が低下する可能性がある [29]．そこで，300 次元日本語エンティティベクトル(注2) を元に，各文を [30] の方法に従って文ベクトルに変換し，ℓ^2 ノルムが 1 となるように正規化してモデルの入力とした．NPDV の学習時の設定は §5.1.1 に記載した設定と同一である．比較のため t-SNE でも可視化を行った．

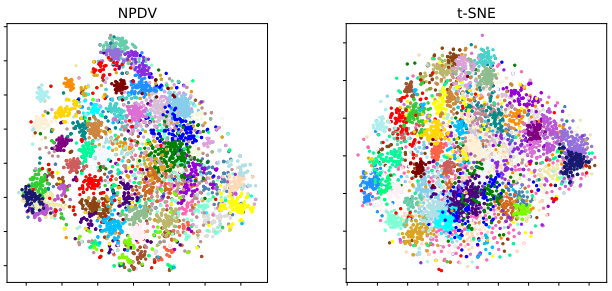


図 5: 京都大学コーパス可視化結果．各点の色は，NPDV では推定された潜在クラスター割り当てに，t-SNE では，GMM により推定された観測空間でのクラスター割り当てに対応している

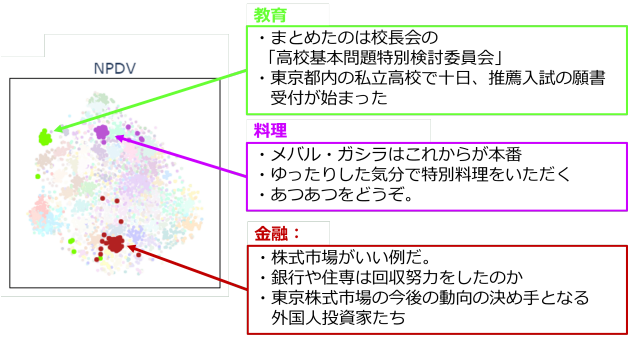


図 6: NPDV で推定されたクラスター内のテキスト例．

5.2.2 可視化結果

NPDV，t-SNE で得られた散布図を図 5 に示す．各点の色は，NPDV では NN-iWMM により推定された潜在クラスター割り当てを，t-SNE では，$K = 50$ とした GMM を観測値に適用して得られた，観測空間でのクラスター割り当を示す．NPDV は，t-SNE と比較してクラスターの境界が明確であり，視認性の高い結果が得られている．また，各クラスターは意味的類似性が高い文から形成されている．その例を図 6 に示す．図内の例のように，料理，教育や金融などのクラスターを確認した．それ以外にも，国際関係や裁判など，意味の類似した文が集まってクラスターが形成されていた．文データでは，文意は個体の特徴を表すため，NPDV では，特徴の似た個体が集まってクラスターを形成する，という性質が示された．

図 7 は t-SNE の散布図であり，図 6 の NPDV の料理クラスターに含まれていた 58 文を強調表示している．右側の図はその拡大図である．t-SNE ではこれらの文は 7 クラスターに分割されており，その一部は料理とは無関係のクラスターに含まれていた．NPDV では，「メバル，ガシラはこれからが本番」，「あつあつをどうぞ。」という文は料理クラスターに属している．しかし，t-SNE では，前者は芸能クラスターに，後者は様々な文が入り混じるクラスターに含まれていた．NPDV では，t-SNE

(注2)：http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/

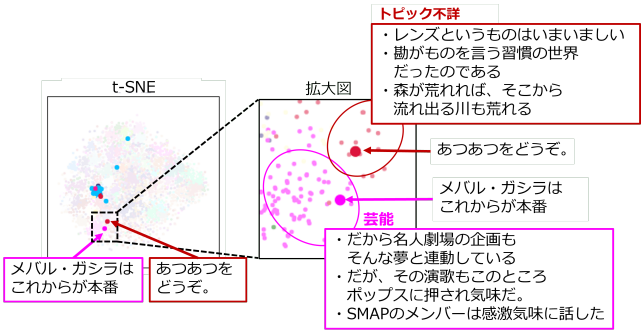


図 7: *t*-SNE において，NPDV では図 5 の料理クラスターに含まれた文を強調表示した散布図とその拡大図

とは異なり，観測値ではなく NN-iWMM で推定された潜在変数を圧縮して可視化表現を得る．NN-iWMM を用いた圧縮により，観測値よりも文意を正確に反映する潜在変数が推定され，それを再び圧縮したことで，より誤謬の少ない可視化につながった可能性がある．

データセットに内在するクラスターを把握することは，データ分析を行う上で重要な情報となる．ラベルデータはクラスター構造を知るための強力なヒントになるが，実際のデータ分析ではラベルデータを利用できないことも多い．提案手法は，*t*-SNE 等と異なり，NN-iWMM で推定されたクラスター割り当てを可視化に反映できるという利点がある．さらに本実験からは，それらのクラスターが類似した個体同士で形成される，妥当性が高いものであることを確認した．加えて，クラスターの境界が明確で視認性の高い散布図が得られるため，ラベルデータが利用できない状況でも，提案手法による可視化は，データセットのクラスター構造の直感的な理解を促進し得る．

6. ま と め

本稿では，新たな深層潜在変数モデルである NN-iWMM と *t*-SNE を統合した，可視化のための確率モデルである NPDV を提案した．提案手法は，NNGP カーネルに基づく L 層無限ユニット NN の利用や，無限混合ガウス分布による潜在変数のモデル化により．従来の深層潜在変数モデルよりも少ない超パラメータで学習が行われる．また，重み，バイアスを用いる通常の NN とは異なり，NNGP カーネルは 2 つのパラメータで NN による変換を定義できるため，NPDV は学習されるパラメータ数も少ない．実験では，既存手法と比較して，NPDV が個体の特徴の違いを高精度に反映した可視化を行うことが示された．加えて，文データの可視化実験から，NPDV では，特徴の似た個体が集まってクラスターを形成する，クラスターの境界が明確で視認性の高い散布図が得られるという性質を確認した．これらの性質から，NPDV による可視化は，ラベルデータが利用できない状況でも，データセット内のクラスター構造の直感的な理解を促進する．

文 献

[1] J. B. Kruskal. “Multidimensional Scaling by Pptimizing Goodness of Fit to a Nonmetric Hypothesis,” Psychometrika, vol.29, no.1, 1964.

[2] D. Lunga et al., ”Manifold-learning-based feature extraction for classification of hyperspectral data: a review of advances in manifold learning,” IEEE Signal Process. Mag., vol.31, no.1, pp.55-–66 2014.

[3] van der Maaten and G. Hinton, “Visualizing data using *t*-SNE,” Journal of Machine Learning Research, vol.9, pp.2579-–2605, Nov., 2008.

[4] L. McInnes, J. Healy, N. Saul and L. Großberge, “UMAP: Uniform Manifold Approximation and Projection,” The Journal of Open Source Software, 3(29):861, 2018.

[5] S. Jones, “A Statistical Interpretation of Term Specificity and its Application in Retrieval.” Journal of Documentation, vol.28, no.1, pp.11-–21, 1972.

[6] C. M. Bishop, “Variational principal components,” In Proceedings 9th International Conference on Artificial Neural Networks, vol.1, pp.509—514, 1999.

[7] Z. Ge, “Process Data Analytics via Probabilistic Latent Variable Models: A Tutorial Review,” Ind. Eng. Chem. Res., vol. 57, pp. 12646–12661, 2018.

[8] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,”, arXiv preprint arXiv:1312.6114, Dec., 2013.

[9] 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, Nov., 2016.

[10] Y. Zheng, H. Tan, B. Tang, H. Zhou, et al, “Variational Deep Embedding: A Generative Approach to Clustering,” arXiv preprint arXiv:1611.05148, Nov., 2016.

[11] M. M. Fard, T. Thonet and E. Gaussier, “Deep *k*-means: Jointly Clustering with *k*-means and Learning Representations,” Pattern Recognition Letters, vol.138, pp.185–192, 2020.

[12] J. Xie, R. Girshick and A. Farhadi, “Unsupervised Deep Embedding for Clustering Analysis,” the 33rd International Conference on Machine Learning, vol.48, pp.478–487, June, 2016.

[13] B. Yang, X. Fu, N. D. Sidiropoulos and M. Hong, “Towards k-means-friendly Spaces: Simultaneous Deep Learning and Clustering,” the 34th International Conference on Machine Learning, vol.78, pp.3861–3870, Aug., 2017.

[14] S. Ghosal, and van der Vaart, "Fundamentals of Nonparametric Bayesian Inference" (Vol. 44), Cambridge, UK: Cambridge University Press, 2017.

[15] E. Abbasnejad, A. Dick and van den Hengel, “Infinite Variational Autoencoder for Semi-supervised Learning,” arXiv preprint arXiv:1611.07800, nov., 2016.

[16] R. Singh, J. Ling and F. Doshi-Velez, “Structured Variational Autoencoders for the Beta-Bernoulli Process,” Advances in Neural Information Processing Systems, 2017.

[17] X. Yang, Y. Yan, K. Huang and R. Zhang, “VSB-DVM: An End-to-End Bayesian Nonparametric Generalization of Deep Variational Mixture Model,” IEEE International Conference on Data Mining, Nov., 2019.

[18] D. Blei and M. Jordan, “Variational nference for Dirichlet Process Mixtures,” Journal of Bayesian Analysis, vol.1 no.1, pp.121-–144, Mar., 2006.

[19] J. H. Lee, Y. Bahri, R. Novak, S. Schoenholz, J. Pennington and J. Sohl-Dickstein, “Deep Neural Networks as Gaussian Processes,”, International Conference on Learning Representation, Feb., 2018.

[20] T. Iwata, D. Duvenaud and Z. Ghahramani, “Warped mixtures for nonparametric cluster shapes,” International Conference on Uncertainty in Artificial Intelligence, pp. 311–319, 2013.

[21] N.D. Lawrence, “Gaussian Process Latent Variable Models for Visualisation of High Dimensional Data,” the 16th Advances in Neural Information Processing Systems, pp.329—336, 2004.

[22] J. Zhu, N.Chen, and E. P. Xing, “Bayesian Inference with Posterior Regularization and Applications to Infinite Latent SVMs,”, Journal of Machine Learning Research, vol.15, no.1, pp.1799—1847, 2014.

[23] J. Sethuraman, “ Constructive Definition of Dirichlet Priors,” Statistica Sinica, vol.4, pp.639-–650, 1994.

[24] Y. Cho and L. K. Saul, “Kernel methods for deep learning,” the 22th In Advances in Neural Information Processing Systems, pp.342—350, 2009.

[25] A. Zellner, ”Optimal information processing and Bayestheorem,”, American Statistician, vol.42, pp.278–280, 1988.

[26] M. K. Titsias and N. D. Lawrence, “Bayesian Gaussian Process Latent Variable Model,”, the 13th International Workshop on Artificial Intelligence and Statistics, vol.9, pp.844-–851, 2010.

[27] van der Maaten, “Preserving Local Structure in Gaussian Process Latent Cariable Models,”, the 18th Annual Belgian-Dutch Conference on Machine Learning, pp.81-–88, 2009.

[28] 黒橋禎夫, 長尾眞, “京都大学テキストコーパス・プロジェクト”, 言語処理学会 第 3 回年次大会, pp.115–118, 1997.

[29] X. Yan, J. Guo, S. Liu, X. Cheng and Y. Wang, “Clustering Short Text Using Ncut-weighted Non-negative Matrix Factorization,”, International Conference on Information and Knowledge Management, pp.2259-–2262, 2012.

[30] S. Arora, Y. Liang and T. Ma, “A Simple but Tough-to-beat Baseline for Sentence Embeddings,” In Internatonal COnference on Learning Representation, 2017.