確率的潜在変数モデルに基づくデータマイニング

岩田　具治

データに内在する隠れた構造を抽出するために確率的潜在変数モデルが広く利用されている．本稿では，まず，確率的潜在変数モデルの基礎と代表的なモデルである，混合モデル，確率的主成分分析，変分オートエンコーダについて説明する．そして，データマイニング分野における応用例として，教師なしオブジェクトマッチング，集約されたデータからの人流推定，ゼロショットドメイン適応に関する研究を紹介する．

キーワード：潜在変数モデル，混合モデル，確率的主成分分析，変分オートエンコーダ，教師なし学習，機械学習

1. はじめに

われわれの身の周りにあるデータはさまざまな隠れた構造をもっている．文書データでは，同じトピックの文書では同じ単語が使われやすい．たとえば，スポーツのトピックの文書では，「勝利」「試合」「スタジアム」といった単語がよく使われる．購買データでは，年齢や職業が近い消費者は，似た商品を買いやすい．写真データでは，同じ写真で写っていやすいモノ（たとえば「包丁」と「まな板」）がある．ソーシャルネットワークデータでは，友達の友達は友達になりやすい．

このような隠れた構造を自動的に見つけることは，そのデータの理解や活用につながる．たとえば，文書データのトピックを抽出することにより，大量の文書を整理し，似た内容の文書を見つけることができるようになる．また，消費者を購買行動が似たグループで分けることにより，それぞれのグループに特化した効果的なマーケティング戦略が立てられる．

隠れた構造を見つけるための手法として，確率的潜在変数モデルが広く利用されている．確率に基づく手法であるため，実世界データに含まれるノイズを適切に扱うことができ，また，長年蓄積されている確率統計における推定，データ統合，モデル化などに関する技術が利用可能になる．

2. 確率的潜在変数モデルの基礎

確率的潜在変数モデルでは，隠れた構造を潜在変数として表現し，潜在変数 z からデータ x が生成される過程を，条件付き確率 $p(x|z)$ でモデル化する．たとえば，文書データの場合，トピックが潜在変数 z，文書がデータ x，トピックごとにどのような単語が使われやすいかが条件付き確率 $p(x|z)$ に対応する．データ x の潜在変数 z を知りたい場合，条件付き確率 $p(z|x)$ を推定できればよい．この確率 $p(z|x)$ は，生成過程 $p(x|z)$ からベイズ則を使って導くことができる [1].

$$p(z|x) = \frac{p(z)p(x|z)}{p(x)} \tag{1}$$

ここで $p(z)$ はデータ x が与えられる前の潜在変数の確率（事前確率）であり，$p(z|x)$ は事後確率と呼ばれる．

確率的潜在変数モデルに基づくデータマイニングは，

1. 潜在変数 z からデータ x が生成される過程 $p(z), p(x|z)$ をモデル化する
2. モデルのパラメータを推定する
3. ベイズ則に従い事後確率 $p(z|x)$ を推定する

という流れで，実行される．

以下に，三つの代表的な確率的潜在変数モデルである，離散潜在変数をもつ混合モデル（2.1 節），連続潜在変数をもつ確率的主成分分析（2.2 節），潜在変数とデータとの間に非線形な関係を考える変分オートエンコーダ（2.3 節）を紹介する．

2.1 混合モデル

潜在変数が離散の場合 $z \in \{1, 2, \ldots, K\}$ を考える．データごとの離散潜在変数を推定することにより，データをクラスタリングできる．データの確率は，混合モデル

$$p(x) = \sum_{z=1}^{K} p(z)p(x|z) \tag{2}$$

で与えられる．

条件付き確率 $p(x|z)$ として正規分布を用いた場合，混合正規分布

いわた　ともはる
NTT コミュニケーション科学基礎研究所
tomoharu.iwata.gy@hco.ntt.co.jp

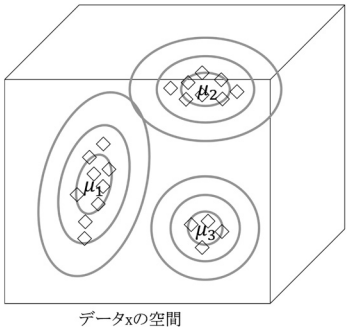


図 1　混合正規分布

$$p(x) = \sum_{z=1}^{K} \pi_z \mathcal{N}(x|\mu_z, \Sigma_z) \tag{3}$$

となる（図 1）．ここで $\pi_z = p(z)$ は潜在変数 z が選ばれる事前確率，$\mathcal{N}(\mu, \Sigma)$ は平均 μ，共分散 Σ の正規分布を表す．

推定すべき未知パラメータは $\{\pi_z, \mu_z, \Sigma_z\}_{z=1}^{K}$ である．これらのパラメータは，尤度を最大化する最尤法によって推定できる．最大化すべき目的関数である対数尤度は，N 個のデータ $X = \{x_n\}_{n=1}^{N}$ が与えられたとき，

$$L = \sum_{n=1}^{N} \log \sum_{z=1}^{K} \pi_z \mathcal{N}(x_n|\mu_z, \Sigma_z) \tag{4}$$

となる．データごとの潜在変数の事後確率は，ベイズ則を用い

$$p(z|x_n) = \frac{\pi_z \mathcal{N}(x_n|\mu_z, \Sigma_z)}{\sum_{z'=1}^{K} \pi_{z'} \mathcal{N}(x_n|\mu_{z'}, \Sigma_{z'})} \tag{5}$$

で得られる．

2.2　確率的主成分分布

潜在変数が連続の場合 $z \in \mathbb{R}^K$ を考える．データごとの連続潜在変数を推定することにより，データの特徴抽出，次元削減，可視化ができる．データの確率は，離散の場合式 (2) の和を積分にした，

$$p(x) = \int p(z)p(x|z)dz \tag{6}$$

となる．

条件付き確率 $p(x|z)$ として平均が潜在変数の線形変換である正規分布 $p(x|z) = \mathcal{N}(x|Wz + \mu, \sigma^2 I)$，事前分布として標準正規分布 $p(z) = \mathcal{N}(z|0, I)$ を用いたとき，確率的主成分分析となる [2]（図 2）．ここで $W \in \mathbb{R}^{D\times K}$ は線形射影行列，D はデータの次元 $x \in \mathbb{R}^D$，I は単位行列を表す．共役事前分布を用いたため，潜在変数に関する積分を解析的に計算でき，データの確率 (6) は，

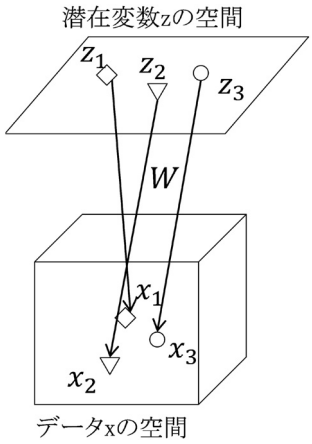


図 2　確率的主成分分析

$$\begin{aligned} p(x) &= \int \mathcal{N}(z|0, I)\mathcal{N}(x|Wz + \mu, \sigma^2 I)dz \\ &= \mathcal{N}(x|\mu, WW^\top + \sigma^2 I) \end{aligned} \tag{7}$$

と潜在変数 z を含まない正規分布で表現できる．推定すべきパラメータは W, μ, σ である．混合正規分布と同様に，対数尤度 $L = \sum_{n=1}^{N} \log p(x_n)$ を最大化することにより，パラメータを推定できる．事後確率は，ベイズ則により，正規分布

$$\begin{aligned} p(z|x) = \mathcal{N}((W^\top W + \sigma^2 I)^{-1} W^\top (x - \mu), \\ \sigma^{-2} W^\top W + I) \end{aligned} \tag{8}$$

で得られる．

2.3　変分オートエンコーダ

確率的主成分分析では，潜在変数 z とデータ x は線形関係にあると仮定していた．しかしながら，画像など複雑なデータの場合，線形関係であるとは限らない．非線形関係を扱うことが可能な手法として，変分オートエンコーダ [3] が提案されている．変分オートエンコーダでは，条件付き確率 $p(x|z)$ として平均が潜在変数の非線形変換 $f(\cdot)$ である正規分布 $p(x|z) = \mathcal{N}(x|f(z), \sigma^2 I)$，事前分布として標準正規分布 $p(z) = \mathcal{N}(z|0, I)$ を用いる（図 3）．非線形変換 $f(\cdot)$ としてニューラルネットが用いられる．このとき，データの確率は

$$p(x) = \int \mathcal{N}(z|0, I)\mathcal{N}(x|f(z), \sigma^2 I)dz \tag{9}$$

となる．

変分オートエンコーダでは，イェンセンの不等式を用いて導出される対数尤度の下限を最大化することによってパラメータを推定する．

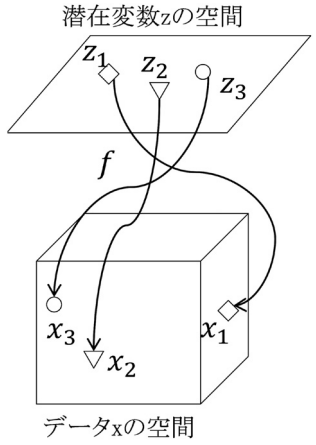


図 3　変分オートエンコーダ

$$
\begin{aligned}
L &= \sum_{n=1}^{N} \log \int p(z)p(x_n|z)dz \\
&= \sum_{n=1}^{N} \log \int \frac{q(z|x_n)}{q(z|x_n)} p(z)p(x_n|z)dz \\
&\geq \sum_{n=1}^{N} \int q(z|x_n) \log \frac{p(z)p(x_n|z)}{q(z|x_n)} dz \qquad (10)
\end{aligned}
$$

ここで $q(z|x_n)$ は近似事後確率を表す．近似事後確率として平均と分散がデータ x_n を入力とするニューラルネットでモデル化される正規分布

$$
q(z|x_n) = \mathcal{N}(z|g(x_n), g'(x_n)) \qquad (11)
$$

を用いる．これらのニューラルネット $g(\cdot), g'(\cdot)$ は $f(\cdot)$ と同様に対数尤度の下限を最大化することにより推定する．確率的主成分分析とは異なり変分オートエンコーダでは積分を解析的に計算できないため，モンテカルロ積分を用いて積分を近似する．

変分オートエンコーダは，ニューラルネットを生成過程 $p(x|z)$ および事後確率 $p(z|x)$ に導入することにより，複雑なデータを扱うことができる確率的潜在変数モデルを実現している．

2.4　拡張

本節では基本的なモデルとして，混合モデル，確率的主成分分析，変分オートエンコーダを紹介したが，さまざまな拡張が考えられる．これらのモデルでは条件付き確率 $p(x|z)$ として正規分布を仮定したが，与えられたデータに応じて別の分布を用いることもできる．たとえば，データが連続値の場合は正規分布，二値の場合はベルヌーイ分布，離散値の場合は多項分布，非負整数値の場合はポアソン分布，非負連続値の場合はガンマ分布，外れ値がある場合はスチューデント t 分布 [4] などと使い分けることによって，データのノイズに適したクラスタリング・次元削減が可能になる．変分オートエンコーダのニューラルネットとして，系列データの場合はリカレントニューラルネット [5]，画像データの場合は畳み込みニューラルネット [6]，グラフデータの場合はグラフ畳み込みニューラルネット [7] などと，データの構造に応じた使い分けも重要である．また，複数の確率モデルを組み合わせる拡張も考えられる．たとえば，混合モデルと変分オートエンコーダを組み合わせることにより，特徴抽出した空間でのクラスタリングが可能となる [8]．

3.　確率的潜在変数モデルの応用

本節では，確率的潜在変数モデルのデータマイニング分野における応用例として，教師なしオブジェクトマッチング（3.1 節），集約されたデータからの人流推定（3.2 節），ゼロショットドメイン適応（3.3 節）を紹介する．

3.1　教師なしオブジェクトマッチング

オブジェクトマッチングとは，異なるドメインのオブジェクトを対応づけるタスクである．たとえば，英語の単語を同じ意味の日本の単語に対応づけたり，異なるデータベースで同一のユーザを対応づけたりする．教師情報（たとえば，単語の対応づけの場合は辞書や対訳文）がある場合，教師あり学習によって対応を見つけることができる．しかしながら，プライバシー保護や，コストの理由で対応が得られない場合も多々ある．そこで，教師なしで異なるドメインのオブジェクトを対応づけるための確率的潜在変数モデル [9] を紹介する．

データとしてネットワークを想定し，異なるネットワークのノードを対応づける状況を考える．D 個のネットワークのデータ $X = \{\{\{x_{dnm}\}_{n=1}^{N_d}\}_{m=1}^{N_d}\}_{d=1}^{D}$ が与えられたとする．ここで，N_d はネットワーク d のノード数，ネットワーク d のノード n とノード m の間にリンクがある場合 $x_{dnm} = 1$，ない場合 $x_{dnm} = 0$ である．

各ノードが離散潜在変数をもち，その潜在変数に基づいてリンクが生成される過程をモデル化する．具体的には，ノード間のリンクは，そのノードの潜在変数に依存した確率に基づいて生成されると仮定する．

$$
p(x_{dnm}|z_{dn}, z_{dm}) = \theta_{z_{dn}z_{dm}}^{x_{dnm}} (1 - \theta_{z_{dn}z_{dm}})^{1-x_{dnm}} \qquad (12)
$$

ここで z_{dn} はネットワーク d のノード n の潜在変数，

表 1　英語とドイツ語の文書単語ネットワークの単語ノードクラスタリング結果

クラスタ 1	英語	prize laureates fields nobel
	ドイツ語	nobelpreis preisverleihung nobelstiftung hochschullehrer
クラスタ 2	英語	basketball nba draft weight sportspeople guard pro kg lb
	ドイツ語	draft basketball cm rebounds punkte forward playoffs lakers
クラスタ 3	英語	youth cup goals clubs premier counted footballers app gls fa caps
	ドイツ語	englischer fc united tore kader angegeben league manchester
クラスタ 4	英語	chemist otto carl chemical frederick harold kurt chemists doi irving
	ドイツ語	chemie chemiker datensatz pnd individualisierter vorhanden

$\theta_{k\ell} \in [0,1]$ は潜在変数 k のノードと潜在変数 ℓ のノードの間にリングがある確率である．ポイントは，潜在変数 $\{1,2,\ldots,K\}$ とリンク確率 $\theta_{k\ell}$ がすべてのネットワークで共有されている点である．これにより，異なるネットワークのノードも同じ潜在変数を割り当てることができ，同じ潜在変数が割り当てられたノードが，対応するノードと推定できる．

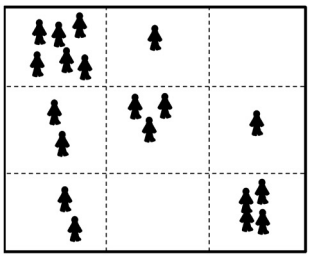
英語とドイツ語の文書単語ネットワークを入力とし，各単語ノードを潜在変数に基づいてクラスタリングした結果を表 1 に示す．クラスタ 1 はノーベル賞関連，クラスタ 2 はバスケットボール関連，クラスタ 3 はサッカー関連，クラスタ 4 は化学関連の単語が，辞書や対訳文なしに言語をまたいでクラスタリングできている．

3.2　集約されたデータからの人流推定

マーケティング，災害対策，公衆衛生，都市計画など，さまざまな分野において，人の流れの解析は重要となる．携帯電話やセンサ機器の普及に伴い，人の移動軌跡が計測可能になってきている．しかしながら，プライバシー保護のために，集約されたデータのみが得られる場合が多い．たとえば，図 4(a)(b) のように，個人ごとの位置情報が，ある範囲に存在する人口に集約される．このような集約された人口データから人の流れを推定する手法 [10] を紹介する．

人口の時系列データ $X = \{\{x_{t\ell}\}_{\ell\in\mathcal{L}}\}_{t=1}^{T}$ が与えられたとする．ここで，$x_{t\ell}$ は時刻 t の場所 ℓ の人口，$\mathcal{L}$ は場所集合，T は時刻数である．タスクは，時刻ごとの人の流れ $Z = \{\{\{z_{t\ell\ell'}\}_{\ell'\in\mathcal{E}_\ell}\}_{\ell\in\mathcal{L}}\}_{t=1}^{T-1}$（図 4(c)）を推定することである．ここで，$z_{t\ell\ell'}$ は時刻 t において場所 ℓ から ℓ' に移動する人数，$\mathcal{E}_\ell \subseteq \mathcal{L}$ は場所 ℓ の近傍の場所の集合である．

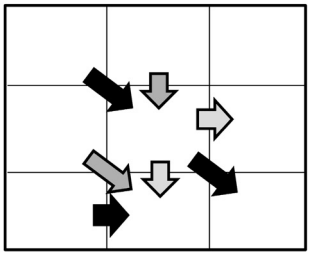
人流データ Z は観測できないため，これを潜在変数と考え，人口データ X から人流データ Z が生成される過程をモデル化する．具体的には，ある場所の人口 $x_{t\ell}$ が割合 $\theta_{\ell\ell'}$ に応じて，近傍の場所へ移動すると考え，以下の多項分布を用いる．



(a) 人の位置データ

6	1	0
2	3	1
2	0	4

(b) 集約された人口データ: X



(c) 人の流れデータ: Z

図 4　集約されたデータからの人流推定

$$p(z_{t\ell}|x_{t\ell}) = \frac{x_{t\ell}!}{\prod_{\ell'\in\mathcal{E}_\ell} z_{t\ell\ell'}!} \prod_{\ell'\in\mathcal{E}_\ell} \theta_{\ell\ell'}^{z_{t\ell\ell'}} \qquad (13)$$

ここで $\theta_{\ell\ell'}$ は場所 ℓ から ℓ' へ移動する割合である．

これまでと同様に最尤推定によって未知パラメータを推定したいが，人口データよりも人流データのほうがパラメータ数が多いため，尤度最大化だけでは，推定することができない．そこで，人口データと人流データの間の以下の関係式を用いる．

$$x_{t\ell} = \sum_{\ell' \in \mathcal{E}_\ell} z_{t\ell\ell'}, \quad x_{t+1,\ell} = \sum_{\ell' \in \mathcal{E}_\ell} z_{t\ell'\ell} \tag{14}$$

第 1 式はある場所から流出する人の数の総和はその場所の人口と一致する，第 2 式はある場所へ流入する人の数の総和は次の時刻のその場所の人口と一致することを表す．対数尤度にこれらの関係式を正則化として加えたもの

$$\begin{aligned} L =& \sum_{t-1} \sum_{\ell} \log p(z_{t\ell}|\theta_{t\ell}, x_{t\ell}) \\ &+ \lambda \sum_{t=1}^{T-1} \sum_{\ell=1}^{L} (x_{t\ell} - \sum_{\ell' \in \mathcal{E}_\ell} z_{t\ell\ell'})^2 \\ &+ \lambda \sum_{t=1}^{T-1} \sum_{\ell=1}^{L} (x_{t+1,\ell} - \sum_{\ell' \in \mathcal{E}_\ell} z_{t\ell'\ell})^2 \end{aligned} \tag{15}$$

を最大化することによって，人流データ Z および遷移確率 $\{\{\theta_{\ell\ell'}\}_{\ell' \in \mathcal{E}_\ell}\}_{\ell \in \mathcal{L}}$ が推定可能となる．

前節までは，潜在変数 z からデータ x が生成される過程をモデル化してきたが，この応用例のように，観測されるデータから観測できない潜在変数が生成される過程をモデル化する場合もある．

3.3 ゼロショットドメイン適応

通常の教師あり学習では，学習するデータの性質（ドメインと呼ぶ）とテストするデータのドメインが異なる場合，性能が劣化してしまう．たとえば，一眼レフカメラで撮った写真で物体認識器を学習し，携帯電話で撮った写真でテストした場合，認識精度が下がる．このような性能劣化を抑える技術としてドメイン適応が提案されている．ドメイン適応では，ドメイン間の分布の差がなくなるように，学習ドメインとテストドメインのデータを使って学習する．しかしながら，テストドメインのデータが学習時に与えられない場合もある．そこで，確率的潜在変数モデルに基づく，学習ドメインのデータのみを用いたゼロショットドメイン適応手法 [11] を紹介する．

D 個のドメインのラベルつきデータ $\{(x_{dn}, y_{dn})_{n=1}^{N_d}\}_{d=1}^{D}$ が与えられたとする．ここで x_{dn} はドメイン d の n 番目のデータ，$y_{dn} \in \{1, \ldots, C\}$ はそのラベルである．テスト時には $X_{d'} = \{x_{d'n}\}_{n=1}^{N_{d'}}$ が与えられ，テストデータでの性能が高い分類器を学習したい．

ドメイン d を潜在変数 $z_d \in \mathbb{R}^K$ で表現する．そして，潜在変数に応じて，ドメインごとに異なる分類器を生成するモデルを考える．具体的には，ドメイン d のデータ x_{dn} のラベル y は，以下の確率で生成されると仮定する（図 5）．

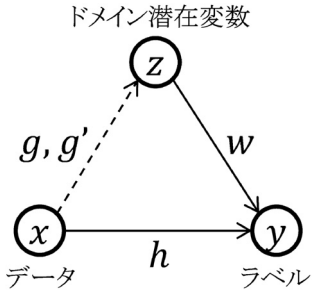


図 5 ゼロショットドメイン適応
実線矢印は生成，点線矢印は推定を表す．

$$p(y|x_{dn}, z_d) = \frac{\exp(w_y(z_d)^\top h(x_{dn}))}{\sum_{y'=1}^{C} \exp(w_{y'}(z_d)^\top h(x_{dn}))} \tag{16}$$

ここで $w_y(\cdot)$ はラベル y に関するドメイン潜在ベクトルに依存した線形識別パラメータを生成するニューラルネット，$h(\cdot)$ はデータから特徴抽出をするニューラルネットである．

変分オートエンコーダと同様に対数尤度の下限の最大化によってパラメータを求める．

$$\begin{aligned} L =& \sum_{d=1}^{D} \log \int p(z_d) \prod_{n=1}^{N_d} p(y_{dn}|x_{dn}, z_d) dz_d \\ \geq& \sum_{d=1}^{D} \int q(z_d|X_d) \log \frac{p(z_d) \prod_{n=1}^{N_d} p(y_{dn}|x_{dn}, z_d)}{q(z_d|X_d)} dz_d \end{aligned} \tag{17}$$

ここで $q(z_d|X_d)$ はドメイン d のデータ $X_d = \{x_{dn}\}_{n=1}^{N_d}$ が与えられたときの潜在変数 z_d の近似事後確率であり，平均と分散が X_d を入力とするニューラルネットである正規分布

$$q(z_d|X_d) = \mathcal{N}(g(X_d), g'(X_d)) \tag{18}$$

でモデル化する．このようにモデル化することにより，学習時に与えられていないテストデータ $X_{d'}$ でもそのドメイン潜在変数 $z_{d'}$ を推定できる．X_d は集合であるため，$g(\cdot), g'(\cdot)$ として集合を入力とするニューラルネットである deep set [12]

$$g(X_d) = \phi(\sum_{n=1}^{N_d} \rho(x_{dn})) \tag{19}$$

を用いる．これにより，ドメインごとにオブジェクト数が異なる場合でも適切に潜在変数の分布を推定できるようになる．

4. おわりに

確率的潜在変数モデルを用いたデータマイニングについて紹介した．深層学習の発展により，教師データが大量にある場合は，汎用的な深層学習モデルを使って高い性能を達成することが可能になってきている．しかしながら，実世界問題では，教師データが少量しか与えられない場合や，まったく手に入らない場合も多々あり，そのような場合は汎用的なモデルでは過学習してしまう．確率的潜在変数モデルに基づいて，データの特性を捉えた生成過程をモデル化することにより，教師データのない隠れた構造の発見や，性能の向上が可能となる．

参考文献

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[2] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **61**, pp. 611–622, 1999.

[3] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," In *Proceedings of International Conference on Learning Representations*, 2013.

[4] H. Takahashi, T. Iwata, Y. Yamanaka, M. Yamada and S. Yagi, "Student-t variational autoencoder for robust density estimation," In *Proceedings of the 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence*, pp. 2696–2702, 2018.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, **9**, pp. 1735–1780, 1997.

[6] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, **25**, pp. 1097–1105, 2012.

[7] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," In *Proceedings of International Conference on Learning Representations*, 2017.

[8] M. Johnson, D. K. Duvenaud, A. Wiltschko, R. P. Adams and S. R. Datta, "Composing graphical models with neural networks for structured representations and fast inference," *Advances in Neural Information Processing Systems*, **29**, pp. 2946–2954, 2016.

[9] T. Iwata, J. R. Lloyd and Z. Ghahramani, "Unsupervised many-to-many object matching for relational data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **38**, pp. 607–617, 2016.

[10] T. Iwata, H. Shimizu, F. Naya and N. Ueda, "Estimating people flow from spatiotemporal population data via collective graphical mixture models," *ACM Transactions on Spatial Algorithms and Systems*, **3**, article number: 2, 2017.

[11] A. Kumagai and T. Iwata, "Zero-shot domain adaptation without domain semantic descriptors," arXiv: 1807.02927, 2018.

[12] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov and A. J. Smola, "Deep sets," *Advances in Neural Information Processing Systems*, **30**, pp. 3391–3401, 2017.