非負値行列分解におけるランク選択

Rank selection for non-negative matrix factorization

数学専攻　羽山 美優

HAYAMA, Miyu

1　はじめに

購買データや文書データなど様々なデータにおける頻出するパタンの抽出や, それを用いた個体のクラスタリング手法として, 特異値分解や主成分分析が広く用いられている. これらは, データを複数の実数値要素の和に分解することによりパタンの抽出を行なっている. したがって, 例えば非負値をとる画像データの場合, これらの手法を用いると分解後の要素が負値を含むことになり, 分解後の行列を画像として表すには要素が非負値となるようにスケーリングを行う必要がある. スケーリング後の 0 は, 非負値における最小値を表す 0 と異なる意味をもち, 得られた基底の解釈が行いにくいなどの問題が生ずる. このような, 分解後の要素も非負値であることが求められる場合に有効な手法の 1 つとして非負値行列分解 (Non-negative Matrix Factorization, NMF)(Lee and Seung (1999, 2001)) がある.

NMF において, 用いる基底数を意味するランクの選択は, 予測精度を左右する重要な問題である. しかし, これまでにランク選択について言及した研究はほとんど存在しない. 本研究では, NMF のランク選択法について検討する.

2　非負値行列分解 (NMF)

ここでは, NMF の定式化を行い, NMF に用いられるダイバージェンスと確率分布の関係性について述べる.

2.1　NMF の定式化

NMF では, 非負値のデータベクトル $\boldsymbol{y}_i$ がいずれも K 個の非負の基底ベクトルの重み付き和によって表されるとして, 全ての $\boldsymbol{y}_i$ を最も良く表現するように K 個の基底ベクトルとその重みとなる非負係数を推定することが目的である. いま, $n \times m$ 非負値データ行列を $Y = (y_{ij})$ として,

$$Y \approx WH^T$$

という分解を考える. W と H はそれぞれ基底行列, 係数行列といい, $W \in \mathbb{R}_{\geq 0}^{n \times K}$, $H \in \mathbb{R}_{\geq 0}^{m \times K}$ である. ここで, $\mathbb{R}_{\geq 0}$ は 0 以上の実数全体を表す. 要素ごとに書き表すと

$$y_{ij} \approx (WH^T)_{ij} = \sum_{k=1}^{K} w_{ik} h_{jk}$$

となる. ただし, 分解するランク K は $K < \min(n, m)$ である.

NMF の主な利点は大きく分けて 2 つある. 1 つ目は, 非負値データ行列を非負値の 2 つの行列に分解するため, 自然な解釈が行えるという点である. 2 つ目は, 要素の表現に足しあわせのみを用いるため非負制約の副次的効果により, 係数行列の要素は 0 に近い値が多くなる傾向がある点である. これより基底行列の情報量が増え, 特徴的な情報を取り出すことが可能である.

NMF は, Y と WH^T の乖離度を表す損失関数 $D(Y|WH^T)$ を用いて, 最適化問題の枠組みで次のように定式化できる:

$$\begin{aligned} &\text{minimize} && D(Y|WH^T) \\ &\text{subject to} && w_{ik} \geq 0,\ h_{jk} \geq 0. \end{aligned}$$

損失関数 $D(Y|WH^T)$ としては, ユークリッド距離や KL ダイバージェンス (Kullback-Leibler divergence), IS ダイバージェンス (板倉齋藤擬距離, Itakura-Saito divergence) とそれら 3 つを統一的に記述できる β-ダイバージェンス (β-divergence) がしばしば用いられる. これらはそれぞれ

$$D_{\mathrm{EU}}(Y|WH^T) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(y_{ij}-\mu_{ij}\right)^2, \tag{2.1}$$

$$D_{\mathrm{KL}}(Y|WH^T) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(y_{ij}\log\frac{y_{ij}}{\mu_{ij}} - y_{ij} + \mu_{ij}\right), \tag{2.2}$$

$$D_{\mathrm{IS}}(Y|WH^T) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{y_{ij}}{\mu_{ij}} - \log\frac{y_{ij}}{\mu_{ij}} - 1\right), \tag{2.3}$$

$$D_{\beta}(Y|WH^T) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(y_{ij}\frac{y_{ij}^{\beta-1}-\mu_{ij}^{\beta-1}}{\beta-1} - \frac{y_{ij}^{\beta}-\mu_{ij}^{\beta}}{\beta}\right) \tag{2.4}$$

で与えられる. ここで $\mu_{ij} = \sum_{k=1}^{K} w_{ik}h_{jk}$ である. また, β は実数であり, $\beta \neq 0$, $\beta \neq 1$ である. いま, $\lim_{\beta\to 0}(\mu_{ij}^{\beta} - y_{ij}^{\beta})/\beta = \log(\mu_{ij}/y_{ij})$ であることから, 式 (2.4) は $\beta \to 2$ のときユークリッド距離, $\beta \to 1$ のとき KL ダイバージェンス, $\beta \to 0$ のとき IS ダイバージェンスとなることが確認できる.

それぞれのダイバージェンスを規準とした NMF のアルゴリズムは補助関数法 (亀岡 (2012), 澤田 (2012)) によって導出できる.

2.2 NMF におけるダイバージェンスと確率分布の関係

NMF を統計モデルの観点からみたとき, データの分布と式 (2.1) から式 (2.4) で述べた規準を関連づけることで, データの性質を反映させた損失関数を選択することが可能である.

式 (2.1) から式 (2.4) を規準とした NMF の最適化問題は, y_{ij} がそれぞれ μ_{ij} を平均とした正規分布, ポアソン分布, 指数分布 (ガンマ分布), Tweedie 分布

$$y_{ij} \sim N\left(y_{ij}|\mu_{ij},\sigma^2\right), \quad y_{ij} \sim \mathrm{Po}\left(y_{ij}|\mu_{ij}\right), \quad y_{ij} \sim \mathrm{Exp}\left(y_{ij}|\mu_{ij}\right), \quad y_{ij} \sim \mathrm{Tweedie}\left(y_{ij}|\mu_{ij},\phi\right)$$

に従って独立に生成されたと仮定した場合の W, H の最尤推定問題と等価である. ここで,

$$\begin{aligned} N\left(z|\mu,\sigma^2\right) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(z-\mu)^2}{2\sigma^2}\right\}, \\ \mathrm{Po}\left(z|\mu\right) &= \frac{\mu^z e^{-\mu}}{z!} \quad (z \geq 0), \\ \mathrm{Exp}\left(z|\mu\right) &= \frac{1}{\mu}e^{-z/\mu} \quad (z \geq 0), \\ \mathrm{Tweedie}\left(z|\mu,\phi\right) &= a_\beta\left(z,\phi\right)\exp\left\{\frac{1}{\phi}\left(z\theta\left(\mu\right) - \kappa\left(\mu\right)\right)\right\} \end{aligned}$$

である. また

$$\theta(\mu) = \begin{cases} \frac{\mu^{\beta-1}-1}{\beta-1} & (\beta \neq 1) \\ \log\mu & (\beta = 1) \end{cases}, \qquad \kappa(\mu) = \begin{cases} \frac{\mu^{\beta}-1}{\beta} & (\beta \neq 0) \\ \log\mu & (\beta = 0) \end{cases}$$

であり, μ, ϕ はそれぞれ期待値と分散パラメータ, $\beta \in (-\infty, 1] \cup [2, \infty)$ は分散を決定する指標である. β-ダイバージェンスがユークリッド距離, KL ダイバージェンス, IS ダイバージェンスの一般化であるように, Tweedie 分布は β の値によって異なる分布を表現することができ, 正規分布 ($\beta = 2$), ポアソン分布 ($\beta = 1$), ガンマ分布 ($\beta = 0$) の一般化となっている. $a_\beta(z, \phi)$ は, β によって異なり, $\beta = 0,\ 1,\ 2$ のときのみ閉じた形で書ける.

式 (2.1) から式 (2.4) を規準とした NMF の最適化問題が, y_{ij} が対応する分布に従って独立に生成されたと仮定した場合の W, H の最尤推定問題と等価であることは, それぞれを y_{ij} を定数とし, μ_{ij} に関する問題とすることで確認できる.

3 Bayesian NMF

ここでは, KL ダイバージェンスを規準とした NMF を階層ベイズモデルとして再定式化した Bayesian NMF (Cemgil (2009)) について述べる.

Y の各要素は自然数とし, 補助変数 $S \in \mathbb{N}^{n \times K \times m}$ を用いて, Y の各要素を

$$y_{ij} = \sum_{k=1}^{K} s_{ikj} \approx \sum_{k=1}^{K} w_{ik} h_{jk}$$

と近似することを考える. すなわち

$$s_{ikj} \approx w_{ik} h_{jk}$$

となる. 行列分解という観点において S は必要ないが, このような補助変数を新たに導入することによって, 効率的なアルゴリズムを導くことができる.

いま, S の各要素の生成モデルが独立なポアソン分布

$$p(S|W, H) = \prod_{i=1}^{n} \prod_{k=1}^{K} \prod_{j=1}^{m} p(s_{ikj}|w_{ik}, h_{jk}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \prod_{j=1}^{m} \mathrm{Po}(s_{ikj}|w_{ik} h_{jk})$$

であるとする. さらに, Y の各要素は独立にデルタ分布に従う, すなわち

$$p(Y) = \prod_{i=1}^{n} \prod_{j=1}^{m} p(y_{ij}) = \prod_{i=1}^{n} \prod_{j=1}^{m} \mathrm{Del}\left(y_{ij} \middle| \sum_{k=1}^{K} s_{ikj}\right)$$

とする. ここで, デルタ分布は

$$\mathrm{Del}(x|z) = \begin{cases} 1 & \text{if } x = z \\ 0 & \text{otherwise} \end{cases}$$

という離散型確率分布である. 生成の観点から考えると, $\sum_{k=1}^{K} s_{ikj}$ として得られた値がそのまま y_{ij} の値として扱われると考えてよい. また, W, H の事前分布は, ポアソン分布の共役事前分布であるガンマ分布

$$p(W) = \prod_{i=1}^{n} \prod_{k=1}^{K} p(w_{ik}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \mathrm{Gam}\left(w_{ik}|a_{ik}^{W}, b_{ik}^{W}\right),\ p(H) = \prod_{j=1}^{m} \prod_{k=1}^{K} p(h_{jk}) = \prod_{j=1}^{m} \prod_{k=1}^{K} \mathrm{Gam}\left(h_{jk}|a_{jk}^{H}, b_{jk}^{H}\right)$$

とする. ただし, ガンマ分布は

$$\mathrm{Gam}(x|a, b) = \exp\left\{(a-1)\log x - \frac{x}{b} - \log\Gamma(a) - a\log b\right\}$$

であり, $a, b \in \mathbb{R}_{\geq 0}$, $\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$ である. これらの確率変数の関係性を整理し, 同時分布として書くと

$$p(Y, S, W, H) = p(Y|S) p(S|W, H) p(W) p(H)$$

となる. Bayesian NMF のアルゴリズムの導出については, ギブスサンプリングを用いた方法などがある.

4 NMF におけるランク選択法

NMF におけるランク K は, 実データ解析においてデータから何種類のパタンを取り出すかを規定するものである. K を大きくすることで $D(Y|WH^T)$ は小さくなるが, データに過適合すると考えられるため, 適切な K を設定することが必要である.

ここでは, Lee らが提案した NMF に対する赤池情報量規準 (Akaike's Information Criterion, AIC) とベイズ情報量規準 (Bayesian Information Criterion, BIC) を用いたランク選択を提案する. NMF のランク選択に用いる AIC と BIC はそれぞれ

$$\mathrm{AIC} = -2\sum_{i=1}^{n}\sum_{j=1}^{m}\left(y_{ij}\log\sum_{k=1}^{K}w_{ik}h_{jk} - \sum_{k=1}^{K}w_{ik}h_{jk} - \log y_{ij}!\right) + 2K(n+m),$$

$$\mathrm{BIC} = -2\sum_{i=1}^{n}\sum_{j=1}^{m}\left(y_{ij}\log\sum_{k=1}^{K}w_{ik}h_{jk} - \sum_{k=1}^{K}w_{ik}h_{jk} - \log y_{ij}!\right) + \log(nm)K(n+m)$$

である. 修士論文では数値実験を行い, AIC と BIC はともに真のランクよりも小さいランクを選択する傾向があることがわかった. これは, モデルの自由パラメータ数で表現されているバイアスの部分が, 大きすぎるためであると考えられる.

また, Bayesian NMF によるランク選択を試みた. ギブスサンプリングにおける平均値, 中央値, 最頻値を推定値とし, 式 (2.1) が最小となるランクを選択するとして数値実験を行った. その結果, どの推定値を用いた場合でも真のランクよりも小さいランクを選択する傾向があることがわかった.

5 まとめと今後の展望

本研究では, はじめに NMF の定式化, ダイバージェンスと分布の関係について述べた. また, NMF を階層構造で表現することにより, ベイズ推論を用いた Bayesian NMF を定式化した. 最後に, AIC と BIC を用いた NMF におけるランク K の選択について数値実験を行い検証した. 今後の課題として, さらにランク選択の精度を向上するために, 条件付き尤度に基づく情報量規準を用いたランク選択や, Bayesian NMF のパラメータを推定する際にランクを同時推定する方法などが考えられる.

参考文献

[1] Cemgil, A. T. (2009) Bayesian inference for nonnegative matrix factorization models, *Computational Intelligence and Neuroscience*.

[2] Lee, D. D. and Seung, H. S. (1999) Learning the parts of objects by non-negative matrix factorization, *Nature*, **401**, 788-791.

[3] Lee, D. D. and Seung, H. S. (2001) Algorithms for non-negative matrix factorization, In *Advances in Neural Information Processing Systems*, **13**, 556-562.

[4] 亀岡弘和 (2012) 非負値行列因子分解とその音響信号処理応用, 信学技報, **EA2012-118**, 53-58.

[5] 澤田宏 (2012) 非負値行列因子分解 NMF の基礎とデータ／信号解析への応用, 電子情報通信学会誌, **95(9)**, 829-833.