経験変分ベイズ法による非負値行列分解

Non-negative matrix factorization via empirical variational Bayes method

数学専攻　杉生　友輝
SUGIU, Tomoki

1　はじめに

非負値行列分解 (non-negative matrix factorization, NMF) は, 画像データのような非負値をとるデータに対し, 分解後の要素が非負値をとる行列の積に分解することで潜在的な特徴の抽出を可能とする手法である (Lee and Seung, 2001). 非負値行列分解は生成モデルとして確率分布により表現し, ベイズ推論へと拡張ができる. Cemgil (2009) は, KL ダイバージェンス規準における NMF を生成モデルとして解釈し, 階層ベイズモデルとして Bayesian NMF を定式化し, 変分ベイズ法を用いることで NMF の次元数の選択を行った.

変分ベイズ法は事前分布の超パラメータの設定によって結果が異なる. そこで中島, 杉山 (2013) は変分ベイズ法と超パラメータの推定を同時に行う経験変分ベイズ法を提案した. 本研究では, Cemgil が提案した Bayesian NMF に対し, 経験変分ベイズ法を用いて分解した 2 つの行列の階数を客観的に選択できることを検証する.

2　非負値行列分解

非負値データ行列 $\boldsymbol{X} = (x_{ij}) \in \mathbb{R}_{\geq 0}^{n \times m}$ に対し,

$$\boldsymbol{X} \approx \boldsymbol{U}\boldsymbol{H}$$

と非負値の要素からなる基底行列 $\boldsymbol{U} \in \mathbb{R}_{\geq 0}^{n \times d}$ および係数行列 $\boldsymbol{H} \in \mathbb{R}_{\geq 0}^{d \times m}$ に分解できると仮定する. ここで, $\mathbb{R}_{\geq 0}^{n \times m}$ は非負の範囲からなる $n \times m$ の実数行列である. 要素ごとに書き直すと

$$x_{ij} \approx \sum_{k=1}^{d} u_{ik} h_{kj}$$

となる. ただし, $d < \min\{n, m\}$ である. このような $\boldsymbol{X}$ の分解を非負値行列分解という.

非負値行列分解は最適化問題の枠組みでは

$$\begin{aligned} \underset{U,H}{\text{minimize}} \quad & D(\boldsymbol{X}||\boldsymbol{U}\boldsymbol{H}) \\ \text{subject to} \quad & u_{ik},\ h_{kj} \geq 0 \end{aligned} \tag{2.1}$$

として定式化することができる. ここで, $D(\boldsymbol{X}||\boldsymbol{U}\boldsymbol{H})$ は $\boldsymbol{X}$ と $\boldsymbol{U}\boldsymbol{H}$ の乖離度を表す規準であり, ユークリッド規準 $D_{\mathrm{EU}}(\boldsymbol{X}||\boldsymbol{U}\boldsymbol{H})$ や KL ダイバージェンス規準 (Kullback-Leibler divergence) $D_{\mathrm{KL}}(\boldsymbol{X}|\boldsymbol{U}\boldsymbol{H})$, 板倉斎藤ダイバージェンス規準 $D_{\mathrm{IS}}(\boldsymbol{X}||\boldsymbol{U}\boldsymbol{H})$ などが用いられ, それぞれ以下のように書ける:

$$D_{\mathrm{EU}}(\boldsymbol{X}||\boldsymbol{U}\boldsymbol{H}) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(x_{ij} - \sum_{k=1}^{d} u_{ik} h_{kj}\right)^2 \tag{2.2}$$

$$D_{\mathrm{KL}}(\boldsymbol{X}||\boldsymbol{UH}) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(x_{ij}\log\frac{x_{ij}}{\sum_{k=1}^{d}u_{ik}h_{kj}} - x_{ij} + \sum_{k=1}^{d}u_{ik}h_{kj}\right) \tag{2.3}$$

$$D_{\mathrm{IS}}(\boldsymbol{X}||\boldsymbol{UH}) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{x_{ij}}{\sum_{k=1}^{d}u_{ik}h_{kj}} - \log\frac{x_{ij}}{\sum_{k=1}^{d}u_{ik}h_{kj}} - 1\right). \tag{2.4}$$

2.1 NMF におけるダイバージェンスと確率分布の関係

(2.2), (2.3), (2.4) などのダイバージェンスの最小化問題は, データに推定の確率分布を仮定した時の最尤法と等価であることが知られている．より具体的には, それぞれの規準における NMF の最適化問題は, x_{ij} がそれぞれ平均 $\mu_{ij} = \sum_{k=1}^{d} u_{ik}h_{kj} > 0$ とした正規分布, ポアソン分布, 指数分布に従う, すなわち,

$$x_{ij} \sim \mathcal{N}(x_{ij}|\mu_{ij}, \sigma^2), \tag{2.5}$$

$$x_{ij} \sim \mathcal{P}(x_{ij}|\mu_{ij}), \tag{2.6}$$

$$x_{ij} \sim \mathrm{Exp}(x_{ij}|\mu_{ij}) \tag{2.7}$$

のときの u_{ik}, h_{kj} の最尤推定問題が等価である．ここで,

$$\mathcal{N}(y|\mu,\sigma^2) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right),$$

$$\mathcal{P}(y|\mu) = \frac{\mu^y e^{-\mu}}{y!} \quad (y \geq 0),$$

$$\mathrm{Exp}(y|\mu) = \frac{1}{\mu}e^{-\frac{y}{\mu}} \quad (y \geq 0)$$

である.

3 Bayesian NMF

ここでは, Cemgil（2009）が提案した KL ダイバージェンス規準における NMF を階層ベイズモデルとして再定式化した Bayesian NMF について述べる．

観測変数を $\boldsymbol{X} \in \mathbb{N}^{n\times m}$ とし, 補助変数 $\boldsymbol{S} \in \mathbb{N}^{n\times m\times d}$ を用いて, $\boldsymbol{X}$ の各成分を

$$x_{ij} = \sum_{k=1}^{d} s_{ijk} \approx \sum_{k=1}^{d} u_{ik}h_{kj}$$

と近似することを考える．

いま, $\boldsymbol{S}$ の各成分の生成モデルが独立なポアソン分布

$$p(\boldsymbol{S}|\boldsymbol{U},\boldsymbol{H}) = \prod_{i=1}^{n}\prod_{j=1}^{m}\prod_{k=1}^{d} p(s_{ijk}|u_{ik},h_{kj}) = \prod_{i=1}^{n}\prod_{j=1}^{m}\prod_{k=1}^{d}\mathcal{P}(s_{ijk}|u_{ik}h_{kj}) \tag{3.1}$$

とする．さらに, $\boldsymbol{X}$ の各成分は独立なデルタ分布

$$p(\boldsymbol{X}|\boldsymbol{S}) = \prod_{i=1}^{n}\prod_{j=1}^{m} p(x_{ij}|s_{ij}) = \prod_{i=1}^{n}\prod_{j=1}^{m}\mathrm{Del}\left(x_{ij}\middle|x_{ij} = \sum_{k=1}^{d}s_{ijk}\right) \tag{3.2}$$

に従うとする．ここで, デルタ分布は

$$\mathrm{Del}\left(x_{ij}\middle|x_{ij}=\sum_{k=1}^{d}s_{ijk}\right)=\begin{cases}1 & \text{if}\ \ x_{ij}=\displaystyle\sum_{k=1}^{d}s_{ijk}\\ 0 & \text{otherwise}\end{cases}$$

である．また, $\boldsymbol{U},\boldsymbol{H}$ 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

$$p(\boldsymbol{U})=\prod_{i=1}^{n}\prod_{k=1}^{d}p(u_{ik})=\prod_{i=1}^{n}\prod_{k=1}^{d}\mathrm{Gam}(u_{ik}|\alpha_{U_0},\beta_{U_0}) \tag{3.3}$$

$$p(\boldsymbol{H})=\prod_{j=1}^{m}\prod_{k=1}^{d}p(h_{kj})=\prod_{i=1}^{n}\prod_{k=1}^{d}\mathrm{Gam}(h_{kj}|\alpha_{H_0},\beta_{H_0}) \tag{3.4}$$

に従うとする．このとき, 周辺尤度は

$$p(\boldsymbol{X})=\int\int\int p(\boldsymbol{X},\boldsymbol{S},\boldsymbol{U},\boldsymbol{H})\ d\boldsymbol{S}d\boldsymbol{U}d\boldsymbol{H} \tag{3.5}$$

である．(3.5) は解析的に陽な形で得ることができないので, 事後分布は計算することができない．

4 経験変分ベイズ法に基づく非負値行列分解

経験変分ベイズ法は変分ベイズ法と事前分布の超パラメータの推定を同時に行う近似手法である．変分ベイズ法は近似分布の表現に平均場近似という確率変数間の独立性制約をつけ, その制約の下, 事後分布を最も近い分布 (変分事後分布) で表現する手法である (Jordan *et al.*, 1999; Jaakkola, 2001).

3 節の KL ダイバージェンス規準における Bayesian NMF モデルに対し, 経験変分ベイズ法を用いる．このとき, 最適化問題は

$$\begin{aligned}&\underset{q,\alpha_{U_0},\beta_{U_0},\alpha_{H_0},\beta_{H_0}}{\text{maximize}} && \mathcal{L}[q(\boldsymbol{U},\boldsymbol{H},\boldsymbol{S}),\alpha_{U_0},\beta_{U_0},\alpha_{H_0},\beta_{H_0}]\\ &\text{subject to} && q(\boldsymbol{U},\boldsymbol{H},\boldsymbol{S})=q_U(\boldsymbol{U})q_H(\boldsymbol{H})q_S(\boldsymbol{S}),\\ &&& \alpha_{U_0},\beta_{U_0},\alpha_{H_0},\beta_{H_0}>0\end{aligned} \tag{4.1}$$

で与えられる．(4.1) の目的関数は変分下限といい, 以下のように定義される:

$$\mathcal{L}[q(\boldsymbol{U},\boldsymbol{H},\boldsymbol{S}),\alpha_{U_0},\beta_{U_0},\alpha_{H_0},\beta_{H_0}]\equiv\int\int\int q(\boldsymbol{U},\boldsymbol{H},\boldsymbol{S})\log\frac{p(\boldsymbol{X},\boldsymbol{S},\boldsymbol{U},\boldsymbol{H}|\alpha_{U_0},\beta_{U_0},\alpha_{H_0},\beta_{H_0})}{q(\boldsymbol{U},\boldsymbol{H},\boldsymbol{S})}d\boldsymbol{U}d\boldsymbol{H}d\boldsymbol{S}.$$

これを解くと, 変分事後分布はそれぞれ

$$\hat{q}_U(\boldsymbol{U})=\prod_{i=1}^{n}\prod_{k=1}^{d}\mathrm{Gam}(u_{ik}|\hat{\alpha}_U^{(i,k)},\hat{\beta}_U),\qquad \begin{cases}\hat{\alpha}_U^{(i,k)}=\alpha_{U_0}+\displaystyle\sum_{j=1}^{m}x_{ij}\hat{\pi}_{ij}^{(k)}\\ \hat{\beta}_U^{(k)}=\beta_{U_0}+\displaystyle\sum_{j=1}^{m}\frac{\hat{\alpha}_H^{(k,j)}}{\hat{\beta}_H^{(k)}}\end{cases} \tag{4.2}$$

$$\hat{q}_H(\boldsymbol{H})=\prod_{j=1}^{m}\prod_{k=1}^{d}\mathrm{Gam}(h_{kj}|\hat{\alpha}_H^{(k,j)},\hat{\beta}_H^{(k)}),\qquad \begin{cases}\hat{\alpha}_H^{(k,j)}=\alpha_{H_0}+\displaystyle\sum_{i=1}^{n}x_{ij}\hat{\pi}_{ij}^{(k)}\\ \hat{\beta}_H^{(k)}=\beta_{H_0}+\displaystyle\sum_{i=1}^{n}\frac{\hat{\alpha}_U^{(i,k)}}{\hat{\beta}_U^{(k)}}\end{cases} \tag{4.3}$$

$$\hat{q}_S(\boldsymbol{S})=\prod_{i=1}^{n}\prod_{j=1}^{m}\prod_{k=1}^{d}\mathrm{Mult}(s_{ijk}|\hat{\pi}_{ij}^{(k)}) \tag{4.4}$$

で与えられる．ここで, $\boldsymbol{S}$ の変分事後分布は多項分布であり, 変分パラメータ $\hat{\pi}_{ij}^{(k)}$ は以下のように書ける:

$$\hat{\pi}_{ij}^{(k)} = \frac{\exp\left[(\psi(\hat{\alpha}_U^{(i,k)}) - \log\hat{\beta}_U^{(k)}) + (\psi(\hat{\alpha}_H^{(k,j)}) - \log\hat{\beta}_H^{(k)})\right]}{\sum_{k=1}^{d}\exp\left[(\psi(\hat{\alpha}_U^{(i,k)}) - \log\hat{\beta}_U^{(k)}) + (\psi(\hat{\alpha}_H^{(k,j)}) - \log\hat{\beta}_H^{(k)})\right]}. \tag{4.5}$$

また, 超パラメータの推定値は

$$\hat{\alpha}_{U_0} = \psi^{-1}\left(\log\beta_{U_0} + \frac{1}{nd}\sum_{i=1}^{n}\sum_{k=1}^{d}\left[\psi(\hat{\alpha}_U^{(i,k)}) - \log\hat{\beta}_U^{(k)}\right]\right), \quad \hat{\beta}_{U_0} = \alpha_{U_0}\left(\frac{1}{nd}\sum_{i=1}^{n}\sum_{k=1}^{d}\frac{\hat{\alpha}_U^{(i,k)}}{\hat{\beta}_U^{(k)}}\right)^{-1} \tag{4.6}$$

$$\hat{\alpha}_{H_0} = \psi^{-1}\left(\log\beta_{H_0} + \frac{1}{md}\sum_{j=1}^{m}\sum_{k=1}^{d}\left[\psi(\hat{\alpha}_H^{(k,j)}) - \log\hat{\beta}_H^{(k)}\right]\right), \quad \hat{\beta}_{H_0} = \alpha_{H_0}\left(\frac{1}{md}\sum_{j=1}^{m}\sum_{k=1}^{d}\frac{\hat{\alpha}_H^{(k,j)}}{\hat{\beta}_H^{(k)}}\right)^{-1} \tag{4.7}$$

となる．以上の結果を用いて, 超パラメータの初期値を設定し, 逐次的最大化を行うことで推定を行う．修士論文では複数の初期値の設定に対して数値実験を行い, 推定評価を行った．

5 まとめと課題点

本論文では, まず, 非負値行列分解の定式化, ダイバージェンス規準と確率分布の関係について述べた．また, NMF を生成モデルとして解釈することにより, ベイズ推論を用いた Bayesian NMF を定式化した．最後に, 経験変分ベイズ法をすることで, NMF の次元数を決定しつつ, 超パラメータの推定をシミュレーションを行い検証した．その結果, 経験変分ベイズ法の方が真の次元数を選択できており, 超パラメータの推定精度も真の次元数に近いほど良いことが確認できた．課題点としては係数行列 $\boldsymbol{H}$ における超パラメータの推定精度が悪さ, 新たな分布を仮定することによるモデルの拡張などがあげられる．現在, 複素データ行列に対し, ベイズ推論の拡張を行なっている．

参考文献

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

[2] Jaakkola, T. (2001) Tutorial on variational approximation methods. *Advances in Mean Field Methods*, 129-159.

[3] Jordan, M., Ghahramani, Z. and Jaakkola, T. (1999) An introduction to variational methods for graphical models, *Machine Learning*, **37**, 183-233.

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

[5] 中島伸一, 杉山将 (2013) 変分ベイズ学習理論の最近動向, 日本応用数理学会論文誌, Vol.23, No.3, 453-483.