3I4-OS-14b-2in

リツイート時系列の 3 パラメータ混合対数正規分布モデルによる分析

Analysis of Retweets Time Series by Mixture Model of Three-Parameter Lognormal Distribution

松澤 有[*1]
Matsuzawa Yu

セーヨー サンティ[*2]
Saeyor Santi

鳥海 不二夫[*1]
Toriumi Fujio

陳 昱[*1]
Chen Yu

大橋 弘忠[*1]
Ohashi Hirotada

[*1] 東京大学大学院工学系研究科 システム創成学専攻
Department of Systems Innovation, School of Engineering, University of Tokyo

[*2] 株式会社ホットリンク 研究開発グループ
Research and Development Group, Hottolink, Inc.

In this research, we discuss a method to analyze statistical characteristics of retweets time series extracted from actual tweets logs. We propose a statistical model which is capable of representing their characteristics, with a required method for estimation of its parameters, and evaluate the superiority of the model by an Information Criteria. Also, we present pattern classification of retweets time series using the proposed model.

1. 導入

2000 年代中盤から普及したブログや SNS などのソーシャルメディアは，いまやインターネットの中核ともいえる存在となっている．インターネットメディア総合研究所によると[公文 2012]，2012 年度の国内インターネット利用者人口に占めるソーシャルメディア利用者割合は52%に達しており，その中でも Twitter の利用率が全サービス中1位である．

Twitter では 2011 年 3 月 11 日の東日本大震災直後，災害に関する様々な情報が爆発的な勢いで交換された．緊急時の通信手段としての有用性や，公共機関の情報発信チャネルとしての意義が認識されると同時に，悪質なデマを急速に広めてしまう危険性も指摘された．これらを踏まえ，関連する研究が盛んに行われるようになっている[白井 2012] [臼井 2012]．

Twitter 上での情報拡散の主要な一形態として，元となるユーザのツイートを引用し，自分のフォロワーに対し紹介するリツイートと呼ばれる行為がある．震災時にも，このリツイートが有益な情報・デマ問わず情報拡散の媒介となった．簡便な手段でありながら短時間で多くのユーザに発言を広める可能性を持つツールだが，その特質を表現し，分類・評価するための時系列分析がまだ不十分である．

本研究ではツイートログからリツイートの時系列データを抽出し，その特徴を表現できる統計モデルと，必要なパラメータを最尤法の枠組みで推定する手法を提案し，情報量規準に照らしてモデルの優位性を評価する．また，モデルの特徴的なパラメータについてクラスタリングを行い，時系列のパターン分類を行う．

連絡先：松澤有
東京大学大学院工学系研究科システム創成学専攻
〒113-8656 東京都文京区本郷 7-3-1 工学部 8 号館 526
TEL: 03-5841-6991
E-mail: matsuzawa@crimson.q.t.u-tokyo.ac.jp

2. 提案モデルと推定手法

2.1 背景となる Twitter ネットワークの想定

リツイートの時系列には，ある短い時間内に爆発的に連鎖したり（バースト），元となるツイートが行われてからある程度の空白期間を経てからバーストが出現し始めたり（時間遅れ），バーストの発生が 1 回にとどまらず，長短の時間遅れを挟んで断続的に発生する（多峰性）といった特徴がある．このような特徴を生じる要因としては，伝播力のあるユーザ（インフルエンサー）が存在しており，かつインフルエンサー間にネットワーク距離があるということが想定できる（図1）．

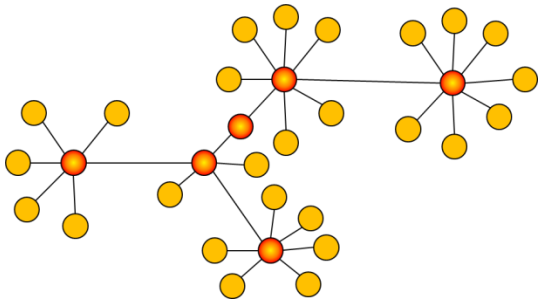

図1 Twitter ネットワークの想定図（各円がユーザ，線がユーザ間のフォロー関係を表す．橙色は伝播力の大きいユーザ）

インフルエンサーが情報をツイート・リツイートすると，そのインフルエンサーから近い距離にいるユーザの集合（いわゆるクラスタ）でリツイートのバーストが起こる．その情報がリツイートの連鎖によって伝搬し，また別のインフルエンサーに到達してリツイートされることで，時間遅れを挟んで複数のバーストが発生する，という想定である．

2.2 3パラメータ混合対数正規分布

2.1 で想定したようなネットワーク上でのリツイートによる情報拡散には感染症伝播の考え方を当てはめることができ，元ツイートあるいはリツイートを受信後に自身がリツイートを行うまでの時間は感染症の潜伏期間になぞらえることができる．この時間分布は統計データから対数正規分布$LN(\mu,\sigma^2)$に従う傾向にあると知られている[Sartwell 1950]．対数正規分布はデータ解析のモデルとして既に用いられているが[鳥海 2012]，ピークが遅れて表れる分布で，立ち上がりの急峻さ，立下りのなだらかさといった特徴が表現できない問題がある．そこで時間遅れパラメータτを導入して平行移動した3パラメータ対数正規分布，

$$f(y|\tau,\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}(y-\tau)}e^{-\frac{(\log(y-\tau)-\mu)^2}{2\sigma^2}} \quad (1)$$

をあるクラスタ内における分布と想定する．時系列全体としてはクラスタの数だけ(1)式を混合した分布となる．

2.3 プロファイル尤度

時間遅れパラメータτは一般的な最尤法の枠組みで構成する尤度方程式では推定が難しいことが知られているが，単一の分布に対してはプロファイル尤度を用いる方法で推定できる[丹後 1998]．

τを所与とした場合のμ,σ^2の最尤推定量は，(1)式の分布関数を基に対数尤度を構成し，それぞれのパラメータで偏微分することで，

$$\hat{\mu}=\hat{\mu}(\tau)=\frac{1}{n}\sum_{i=1}^{n}\log(y_i-\tau)$$

$$\hat{\sigma}^2=\hat{\sigma}^2(\tau)=\frac{1}{n}\sum_{i=1}^{n}\log(y_i-\tau)-\mu^2$$

と求められる．これらを基に対数尤度をτの関数として構成すると，

$$\log L^{**}=-\mathrm{n}\left(\hat{\mu}(\tau)+\frac{1}{2}\log\hat{\sigma}^2(\tau)\right) \quad (2)$$

が得られる． (2)式のような尤度をτのプロファイル尤度と呼び，この尤度を最大にする$\hat{\tau}$が求める最尤推定量となる．$\hat{\tau}$はデータの最小値を上限とする離散的な数値列$\{\tau_1 \le \tau_2 \le \cdots \le \tau_M < \min \boldsymbol{y}\}$を適当に用意して(2)式に代入していき，その中から最大となるプロファイル尤度を探す線形数値探索で簡単に求められる．この方法で$\hat{\tau}$を先に求めておけば，それを所与として残る 2 パラメータを推定することができる．

ただし，プロファイル尤度法は混合分布を想定していないことに注意が必要である．本研究における想定では，単一の分布はあるクラスタ内でのバーストに対応しているものと仮定しているので，時系列データのバーストを検知してクラスタごとに分割し，それぞれのクラスタに対してτを推定するという手順を踏む．

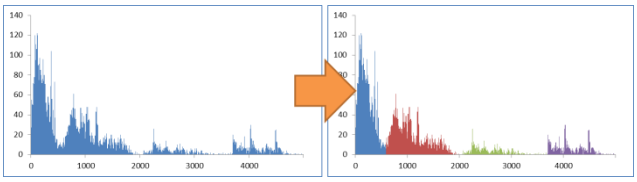

図2 バースト検知による時系列分割のイメージ

2.4 Kleinberg のバースト検知

時系列データのバースト検知の代表的な手法として Kleinberg が考案したアルゴリズムがあり，特定のニュース記事が世の中の出来事に対応して急激に増加するような現象の解析に用いられている[Kleinberg 2002]．このアルゴリズムの最も簡単なモデルでは，非バースト状態q_0とバースト状態q_1の2状態を取りうるオートマトンを定義し，離散的な時間単位で入力されてくる記事（本研究ではツイート）集合における関連記事（リツイート）の割合によってオートマトンの状態が遷移することで，バーストしているか否かを決定する．

ある時刻tに対し，単位時間あたりの総記事数をd_t，関連記事数をr_tとおく．非バースト状態の期待値p_0には，全期間における総記事数合計に占める関連記事数合計の割合，

$$p_0=\frac{\sum_{t=1}^{m}r_t}{\sum_{t=1}^{m}d_t}$$

を割り当てる．バースト状態にはp_0にパラメータsをかけた値である$p_1=p_0 s$を割り当てる．この際sは$s>1$ かつ$p_1 \le 1$であるような値でなければならない．オートマトンの状態は，状態の系列，

$$\boldsymbol{q}=\{q_{i_1},q_{i_2},\cdots,q_{i_m}\}$$

を通るためのコスト計算によって決定する．ある時刻tに状態q_iにいるためのコストは，二項係数を用いて，

$$\sigma(i,r_t,d_t)=-\log\left[\binom{d_t}{r_t}p_i^{r_t}(1-p_i)^{(d_t-r_t)}\right] \quad (3)$$

と定義する．また，閾値付近での状態遷移を自然にするために，異なる状態への遷移にペナルティ，

$$\tau(i,j)=\begin{cases}(j-i)\gamma & (j>i)\\ 0 & (j\le i)\end{cases} \quad (4)$$

を課すことで，頻繁な状態遷移を抑制する．パラメータγは特に理由がない場合は$\gamma=1$とする．(3)(4)式から，状態系列を通るための総コスト関数，

$$c(\boldsymbol{q}|r_t,d_t)=\sum_{t=0}^{m-1}\tau(i_t,i_{t+1})+\sum_{t=0}^{m}\sigma(i_t,r_t,d_t) \quad (5)$$

を構成する．オートマトンの状態系列は(5)式を最小にするものを解とし，バースト時刻は解となる系列における状態q_oからq_1への遷移によって判定する．

Kleinberg のアルゴリズムでは，バースト状態の期待値を決定するパラメータsがバースト判定の閾値として機能し，sを小さくとるとリツイートの割合が少なくてもバーストと見なされやすくなる．実際のリツイート時系列においては，バーストの規模や時間遅れ幅に大きなばらつきがあるため，すべての時系列に対し精度よく検知できるsの決定が難しい．本研究では，

$$p_1=p_0+\frac{1-p_0}{2}$$

となるようなsをそれぞれの時系列に合わせて設定した．

2.5 EM アルゴリズム

ここまでの方法で推定したτを所与とし，混合分布のパラメータ推定手段として確立されている EM アルゴリズム[Dempster 1977][小西 2008]を用い，残りのパラメータを推定する．

不完全な観測によりデータ$\boldsymbol{y}$が得られたとする（以降，小文字は確率変数に対応する実現値を表す）．不完全データの確率変数を$\boldsymbol{Y}$，観測されなかったデータ（欠測）の確率変数を$\boldsymbol{Z}$，完全なデータの確率変数を$\boldsymbol{X^T}=(\boldsymbol{Y^T},\boldsymbol{Z^T})$ とし，その密度関数を

$f^C(\boldsymbol{x}|\boldsymbol{\theta})$ と書く（$\boldsymbol{\theta}$はパラメータのベクトル）. このような条件下で, ある時点で利用可能なパラメータの推定値（初期値, あるいは反復過程の1ステップ前の値）$\boldsymbol{\theta}^{(k)}$ を用いて, $\boldsymbol{Z}$の期待値,

$$E_{\boldsymbol{\theta}^{(k)}}[\boldsymbol{Z}|\boldsymbol{Y}=\boldsymbol{y}]$$

を計算する. 観測されたデータと欠測の期待値とを合わせて擬似完全データ$\boldsymbol{x}^*$を構成し, それを基に完全観測の対数尤度関数,

$$l^C(\boldsymbol{\theta},\boldsymbol{x}^*)=\log f^C(\boldsymbol{x}^*|\boldsymbol{\theta})$$

の最大化を行い, パラメータの最尤推定値を得て, それを$\boldsymbol{\theta}^{(k+1)}$として更新する. このような操作を繰り返して最尤推定値を得る方法が EM アルゴリズムである. 一般的な設定のもとでは, この方法による最尤推定値は観測データ$\boldsymbol{y}$に基づく尤度$\log f(\boldsymbol{y}|\boldsymbol{\theta})$を最大化するパラメータ$\boldsymbol{\theta}^*$に単調に収束することが知られている[Dempster 1977][Wu 1983].

本研究で用いるモデルは混合分布であり, 混合数g, 各分布の混合比（関係するネットワーク全体に占めるそのクラスタの大きさ, 重み）をξとして,

$$f(y|\boldsymbol{\theta})=\sum_{\mathrm{j}=1}^{\mathrm{g}}\xi_j f_j(y|\tau,\mu,\sigma^2)$$

と表わされるような分布である. このようなモデルに対し EM アルゴリズムで最尤推定を行う場合, ある観測値$\boldsymbol{y}$が関係しているネットワーク全体のうち, どの部分集合（クラスタ）Ω_jに属しているかを示す指示データが実際には得られないことから, これが欠測$\boldsymbol{Z}$に当たると考える. 指示データの形式は,

$$\boldsymbol{z}=(z_1,z_2,\ldots,z_g)^T$$

$$z_j=\begin{cases}1 & \text{観測が}\Omega_j\text{から得られた}\\ 0 & \text{それ以外}\end{cases}$$

とする. $\boldsymbol{Y}=\boldsymbol{y}$が与えられたとき, $z_j=1$である場合の条件付き確率密度関数は,

$$f(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta})=\frac{\xi_j f_j(y|\tau_j,\mu_j,\sigma_j^2)}{\sum_{j=1}^{g}\xi_j f_j\left(y\middle|\tau_j,\mu_j,\sigma_j^2\right)} \tag{6}$$

となる. この仮定のもとで完全観測$\boldsymbol{X}$を構成すると, 対数尤度は,

$$l^C(\boldsymbol{\theta},\boldsymbol{x})=\sum_{i=1}^{n}\sum_{j=1}^{g}z_{ij}\log f_j\left(y_j\middle|\tau_j,\mu_j,\sigma_j^2\right)+\sum_{i=1}^{n}\sum_{j=1}^{g}z_{ij}\log\xi_j \tag{7}$$

となる. ただし, z_{ij}は観測y_iに対する指示データベクトル$\boldsymbol{z_i}$のj番目の要素を示す. 欠測は(7)式の通り対数尤度に線形に取り込まれているので, (6)式から構成できるz_{ij}の条件付き期待値をそのまま代入でき, 対数尤度の条件付き期待値,

$$Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(k)}\right)=E_{\boldsymbol{\theta}^{(k)}}[l^C(\boldsymbol{\theta},\boldsymbol{X}|\boldsymbol{Y}=\boldsymbol{y})] \tag{8}$$

が構成できる. この(8)式を反復計算によって最大化すればよい. ξ_jについては和が 1 になることから, ラグランジュの未定乗数法を用いて,

$$\xi_l^{(k+1)}=\frac{1}{\mathrm{n}}\sum_{i=1}^{n}z_{il}^{(k)} \tag{9}$$

と解くことができる. τ_jにはクラスタごとに前項までで別に推定したものを代入し, 残る 2 つのパラメータは(7)式の第二項が非依存であることから,

$$\begin{pmatrix}\frac{\partial}{\partial\mu_l}\\ \frac{\partial}{\partial\sigma_l^2}\end{pmatrix}Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(k)}\right)=\begin{pmatrix}\frac{\partial}{\partial\mu_l}\\ \frac{\partial}{\partial\sigma_l^2}\end{pmatrix}\sum_{i=1}^{n}\sum_{j=1}^{g}z_{ij}^{(k)}\log f_j\left(y_j\middle|\tau_j,\mu_j,\sigma_j^2\right)=\boldsymbol{0} \tag{10}$$

を解けばよい.

EM アルゴリズムをまとめると以下のような計算手順となる.

- 適当なパラメータの初期値$\boldsymbol{\theta}^{(0)}$を用意する.
- E ステップ：所属クラスタの指示データを欠測と考えて, 完全観測に基づく対数尤度の条件付き期待値(8)式を構成する.
- M ステップ：$\boldsymbol{\theta}$の関数として見た$Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(k)}\right)$を最大化するようなパラメータ$\boldsymbol{\theta}^{(k+1)}$を(9)(10)式を解いて求める.
- 収束条件を満足するまで, E ステップと M ステップを繰り返す.

2.6 推定手法まとめ

ここまでの推定手順をまとめると以下のようになる.

- Kleinberg のバースト検知アルゴリズムで, リツイート時系列をクラスタ単位に分割する.
- 分割したクラスタごとにプロファイル尤度法を用いて時間遅れパラメータの最尤推定値$\hat{\boldsymbol{\tau}}$を得る.
- $\hat{\boldsymbol{\tau}}$を既知として, EM アルゴリズムで残るパラメータを推定する. M ステップでは, $\boldsymbol{\xi}$は(9)式で, 残る2パラメータは(10)式を解いた,

$$\mu_l^{(k+1)}=\frac{\sum_{i=1}^{n}z_{il}^{(k)}\log(y_i-\tau_l)}{\sum_{\mathrm{i}=1}^{\mathrm{n}}z_{il}^{(k)}}$$

$$(\sigma_l^2)^{(k+1)}=\frac{\sum_{i=1}^{n}z_{il}^{(k)}(\log(y_i-\tau_l)-\mu_l)^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}}z_{il}^{(k)}}$$

- を用いてそれぞれ更新する.

3. モデル推定結果

3.1 使用するツイートログ

分析に使用したのは, 2011 年 9 月 1 日から同 12 月 31 日までの4ヶ月間に投稿された日本語ツイートのサンプリングデータで, 約 24 億ツイートを含む. このデータから元ツイートごとにリツイート時系列を抽出し, うちリツイート数上位の 1000 件を主に使用する.

また, 提案モデルに対する従来手法として, 時間遅れパラメータを含まない混合対数正規分布によるモデル推定も行い, 尤度の大きさとパラメータの少なさを元にモデルを評価する情報量規準 BIC で結果を比較する.

3.2 推定結果

図3に推定結果の例を示す. 10 分あたり区間頻度分布で表した時系列（淡青, 棒グラフ）に, 推定結果の部分分布と混合分布をそれぞれ点線・実線で重ねてプロットしている.

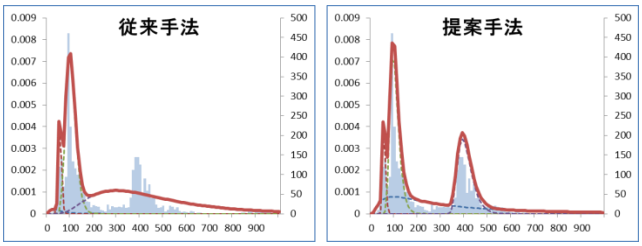

図3　混合数 4 の時系列に対する推定結果.
(横軸:経過時間(分), 左軸:確率密度, 右軸:区間頻度)

3.3 BIC による推定結果の比較

推定結果は, 分布形状においても尤度や BIC などのスコアにおいても従来手法を改善している場合もあるが, 例外も多かった. そこで BIC の変位率をモデルの分布混合数ごとにプロットしたのが図4である. 横軸は混合数で, 最小 1, 最大 20 である. BIC は変位が負であれば改善となる.

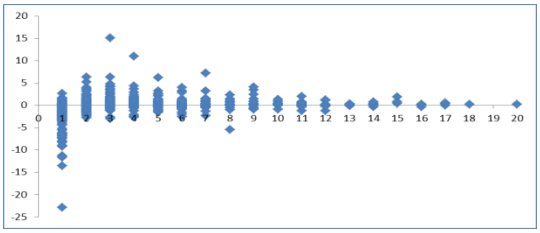

図4　混合数(横軸)別, BIC 変位率(縦軸,%)

平均では−0.22%とわずかに改善した. 混合数 1 のデータは明確に改善したが, それ以外では明確には改善しなかった.

4. クラスタリングによるパターン分類

各時系列データの推定結果から混合比の大きい分布のパラメータを抽出してベクトルを作り, それらを成分ごとに正規化した上で K 平均法によるクラスタリングを行った. 収束したクラスタの中心ベクトルを, クラスタの特徴モデルとして図示したのが図3である.

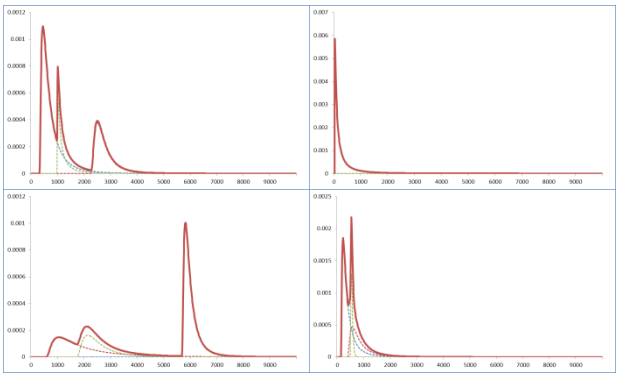

図5　混合比上位3分布から作成した特徴ベクトル群による 4 分割クラスタリング(横軸:経過時間(分), 縦軸:確率密度)

図5の例では, 右上・右下のような初期にバーストが集中し早期収束に至る分布のクラスタにデータ全体の 93%が所属した. このような傾向はパラメータを抽出する分布数, クラスタリングの分割数を変えても変わらず観測された. リツイートの拡散は基本的に元ツイートの配信から約1日以内という短い期間に集中して起こる現象といえる.

5. 結論

情報拡散の一形態として注目されている Twitter 上でのリツイートについて, その時系列データの解析手法とするべく, 3 パラメータ混合対数正規分布モデルを提案した. 同モデルのパラメータ推定手法として, バースト検知アルゴリズム・プロファイル尤度法・EM アルゴリズムを組み合わせた手法を導入し, 全てのパラメータの最尤推定を可能にした. 提案モデルの実用例として, 推定したパラメータを用いて時系列のクラスタリング分類を行い, リツイートの大部分が1日程度の短い時間規模であるという傾向を見出した.

6. 謝辞

研究に際し, Twitter のログデータ収集にご協力頂いた株式会社ホットリンクに感謝する. また, 本研究は科研費(24300064)の助成を受けて行われた.

参考文献

[公文　2012] 公文紫都, 高柳慶太郎, 澤紫臣, インターネットメディア総合研究所:『ソーシャルメディア調査報告書 2012』, 株式会社インプレス R&D, 東京, 2012.

[白井 2012] 白井 嵩士, 榊 剛史, 鳥海 不二夫, 篠田 孝祐, 風間 一洋, 野田 五十樹, 沼尾 正行, 栗原 聡: Twitter におけるデマツイートの拡散モデルの構築とデマ拡散防止モデルの推定, SIG-DOCMAS-B102, 2012.

[臼井 2012] 臼井 翔平, 鳥海 不二夫, 石井 健一郎, 間瀬 健二:　震災による情報伝播ネットワークの変化, 第 26 回人工知能学会全国大会, 2012

[Sartwell　1950] Sartwell, P. E.: The distribution of incubation periods of infectious diseases, American Journal of Hygiene, 51, 310-318, 1950.

[鳥海 2012] 鳥海不二夫, 篠田孝祐, 榊剛史, 栗原聡, 風間一洋, 野田五十樹, 松尾真人: 災害情報支援に向けた大規模データ分析- 異種協調型災害情報支援システム実現に向けた基盤技術の構築, 人工知能学会「社会における AI」研究会第 15 回研究会, 2012

[丹後　1998] 丹後俊郎: 潜伏期間に対数正規分布を仮定した集団食中毒の曝露時点の最尤推定法, 日本公衆衛生雑誌, 45, 129-141, 1998.

[Kleinberg　2002] Kleinberg, J.: Bursty and Hierarchical Structure in Streams, Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002.

[Dempster 1977] Dempster, A. P., Laird, N. M., Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society. Series B(Methodological), Vol. 39, No. 1, 1-38, 1977.

[小西 2008] 小西貞則, 越智義道, 大森裕浩:『計算統計学の方法—ブートストラップ・EM アルゴリズム・MCMC— (シリーズ予測と発見の科学 5)』, 朝倉書店, 東京, 2008.

[Wu 1983] Wu, C. F. J.: On the Convergence Properties of the EM Algorithm, The Annals of Statistics, Vol. 11, No. 1, 95-103, 1983.