

第9章 EMアルゴリズム

渡辺美智子¹

EMアルゴリズムは、不完全データに基づく統計モデル一般に適用される最尤推定導出のためのアルゴリズムである。不完全データを広義にとらえることで、欠損データの問題に限らず、切断や打ち切りデータ、有限混合分布モデル、ロバスト分布モデルや潜在変数モデル、ベイズモデルと、その適用範囲が広いことが特徴である。本章ではEMアルゴリズムの基本的な考え方や理論及び代表的なモデルに対する具体例を解説し、ECMやECMEなどの拡張EMについても概観する。

¹東洋大学経済学部

Contents

| | | |
|-------|-------------------------------------|----|
| 1 | EM アルゴリズム | 1 |
| 1.1 | はじめに | 3 |
| 1.2 | EM アルゴリズムの考え方 | 4 |
| 1.3 | EM アルゴリズムの理論と一般形 | 6 |
| 1.4 | EM アルゴリズムと適用例 | 8 |
| 1.4.1 | 血液型遺伝子に関する発生確率の推定 | 8 |
| 1.4.2 | 多変量データにおける欠測：多変量正規モデル | 9 |
| 1.4.3 | 重回帰モデル：目的変数に欠測がある場合 | 12 |
| 1.4.4 | 中途打ち切りデータに基づく例 | 13 |
| 1.4.5 | 混合分布モデル | 16 |
| 1.5 | EM を利用した漸近分散共分散行列の評価 | 20 |
| 1.5.1 | Louis の方法 | 20 |
| 1.5.2 | Oakes の方法 | 22 |
| 1.5.3 | SEM(Supplemented EM) | 23 |
| 1.6 | EM の特性 | 23 |
| 1.7 | GEM アルゴリズムとその他の拡張型 EM | 25 |
| 1.7.1 | GEM(Generalized EM) | 25 |
| 1.7.2 | ECM アルゴリズム | 26 |
| 1.7.3 | ECME アルゴリズム – Contaminated Normal – | 27 |

CONTENTS

| | |
|------------------------|----|
| 1.7.4 加速化を意識した完全データの探索 | |
| — optimal EM アルゴリズム— | 29 |

1.1 はじめに

EM アルゴリズムは、不完全データから最尤推定値を導くための統一的なアルゴリズムで、Dempster, Laird and Rubin(1977) が最初に、EM という一般的なフレームとその背景となる理論および諸種の統計モデルへの広い適用性を示した。当初、EM アルゴリズムは、最適化アルゴリズムとして既に普及していた Newton-Raphson 法やスコアリング法などと比較し、収束に至るまでの反復回数が多いことなどの問題点が指摘されていたが、アルゴリズムのフレームの単純性と収束の安定性、適用モデルの汎用性などの利点が勝り、今日に至るまでに、工学・医学・社会学・経営学など統計モデルが必要とされるほぼすべての分野で、EM アルゴリズムを推測に利用するモデル構築が数多く発表され続け、尤度に基づく統計分析法を構築する上での汎用的なツールになっている。

EM アルゴリズムの利点は、先述したアルゴリズムのフレームの単純性と収束の安定性、適用モデルの汎用性にある。具体的には、フレームが単純であるため、尤度の定式化から具体的なアルゴリズムの導出およびプログラミングの作成に至るまでの作業効率が比較的に高く、また、尤度の形状にも拠るが、Newton-Raphson 法などと比較した場合に初期値によらず局所的な最適解が得られる安定性がある。そして、EM アルゴリズムの適用事例が広範にわたる最大の理由として、対象とするデータの不完全性の解釈に柔軟性があり、かなり自在に適用モデルを拡張できる汎用性が挙げられる。

不完全データ本来の意味である欠測値問題はもとより、切断分布や打ち切り分布、混合分布、分散分析における変量効果モデル・混合効果モデル、ロバスト分析、潜在変数モデル、ベイズモデルと主要な統計モデルがその範疇に入る。また、範疇に入ったモデル同士を組み合わせさらに複雑なモデルに最尤解を与えることも可能となる。この意味で、EM アルゴリズムは単なる

CONTENTS

計算手段というより、様々な現実課題の様相により適合し説明力のある複雑な統計モデルを一挙に実用段階に押し上げ、さらにモデルの拡張性へのヒントを与え得るアルゴリズムとも言える。

1.2 EM アルゴリズムの考え方

EM アルゴリズムは、不完全データの問題を完全データのフレームワークで逐次的に解決する方法論の一つと位置づけられる。基底をなす考え方は単純で、次のステップでまとめることができる。

1. 手元にあるデータ \mathbf{y}_{obs} だけでは解を導くことがすぐには困難な問題に対して、解決が容易であると考えられるレベルまでデータを完全化し(欠けているデータ部分 \mathbf{y}_{mis} が存在するとして)問題を定式化する。
2. 例えば、パラメータ θ の推定値 $\hat{\theta}$ を導くことが当面の目的であれば、 \mathbf{y}_{mis} に暫定的な値を埋め込み $\hat{\theta}$ を求める。
3. $\hat{\theta}$ を用いて \mathbf{y}_{mis} を改良し、その値を埋め込む。
4. $\hat{\theta}$ の値が収束するまで、前の2つのステップを繰り返す。

例として、血液遺伝子(A,B,O)に関する発生確率 $\theta = (p_A, p_B, p_O)$, ($p_O = 1 - p_A - p_B$) の推定の問題を考えてみる。実際に観測されるデータは、表 1.1 のように血液型に関する頻度データのみで、このデータに基づく直接的な最尤解は数式で明示的に表現することはできず、何等かの反復計算が必要となる。

Table 1.1: 観測データ

CONTENTS

| 血液型 | セル確率 | 観測データ |
|-----|---------------------|----------|
| O | p_O^2 | n_O |
| A | $p_A^2 + 2p_{AP_O}$ | n_A |
| B | $p_B^2 + 2p_{BP_O}$ | n_B |
| AB | $2p_{AP_B}$ | n_{AB} |
| 合計 | | n |

そこで、観測された頻度データは、目的とする血液遺伝子発生確率を推定するためには不十分な不完全データとみなし、表 1.2 のような血液遺伝子の発生確率と直接対応する頻度データの存在を想定してみると、この完全データの下では、目的パラメータの推定問題ははるかに容易になる。

Table 1.2: 観測不可能な完全データ

| 遺伝子型 | セル確率 | 完全データ |
|------|-------------|----------|
| O | p_O^2 | n_O |
| AA | p_A^2 | n_{AA} |
| AO | $2p_{AP_O}$ | n_{AO} |
| BB | p_B^2 | n_{BB} |
| BO | $2p_{BP_O}$ | n_{BO} |
| AB | $2p_{AP_B}$ | n_{AB} |
| 合計 | | n |

もちろん、想定した完全データ (表 1.2) は実際には観測されないので、これをどのように推定して置き換えるかの問題になる。EM アルゴリズムでは、観測される不完全データが与えられた下での完全データの対数尤度の条件付き期待値を介してこの置き換えを行う。具体的な計算例は後述するが、EM アルゴリズムの他にも、このようなロジックを背景にした方法論として、(a) multiple imputation (Rubin 1987a), (b) data augmentation (Taner and Wong 1987), (c) Gibbs sampler (Geman and Geman 1984), (d)

CONTENTS

Sampling/Importance Resampling Algorithm (Rubin 1987b) などがある。Rubin(1991) では、これらを拡張型 EM のフレームワークの下で解説しているが、その区別は一般に、ステップ 2 における \mathbf{y}_{mis} への値の埋め込み方法の違いで説明される。

1.3 EM アルゴリズムの理論と一般形

いま、完全データに対応する変数を \mathbf{Y} とし、 \mathbf{Y} は観測データ (不完全データ) に対する変数 \mathbf{Y}_{obs} と観測されていないデータ (欠測データ) に対応する変数 \mathbf{Y}_{mis} で構成されるとする。ここでは、2つの標本空間： $\Omega_{\mathbf{Y}}$ (完全データの空間) と $\Omega_{\mathbf{Y}_{obs}}$ (不完全データの空間) があり、これらは、 $\Omega_{\mathbf{Y}}$ から $\Omega_{\mathbf{Y}_{obs}}$ への多対 1 写像で関連づけられているとする。

このとき、 \mathbf{Y} の密度関数を $f(\mathbf{y} | \boldsymbol{\theta})$ とすると、観測変数 \mathbf{Y}_{obs} の密度関数 $g(\mathbf{y}_{obs} | \boldsymbol{\theta})$ は

$$g(\mathbf{y}_{obs} | \boldsymbol{\theta}) = \int_{\mathbf{y}_{mis}} f(\mathbf{y} | \boldsymbol{\theta}) d\mathbf{y}_{mis} \quad (1.1)$$

となる。

不完全データの発生過程に関しては、欠測がランダムである (missing at random), 即ち、欠測メカニズムを無視した下で、観測データ \mathbf{y}_{obs} の尤度に基づいてパラメータ $\boldsymbol{\theta}$ の推定を行うことが妥当であることを仮定しておく (Rubin(1976))。EM アルゴリズムの目的は、パラメータ $\boldsymbol{\theta}$ を最尤推定することである。

パラメータ $\boldsymbol{\theta}$ の定義域に対して完全データに基づく対数尤度が定義でき、その関数を $LL_c(\boldsymbol{\theta}) = \log f(\mathbf{y} | \boldsymbol{\theta})$, 不完全データに基づく対数尤度関数を $LL(\boldsymbol{\theta}) = \log g(\mathbf{y}_{obs} | \boldsymbol{\theta})$ とすれば、適当な条件の下で、 $LL(\boldsymbol{\theta})$ の最大化は $LL_c(\boldsymbol{\theta})$ の条件付き期待値の最大化を通して達成できる (Dempster, Laird and Rubin (1977)). つまり、一般に最適化が複雑になる不完全データ \mathbf{y}_{obs} の対数尤度 $LL(\boldsymbol{\theta})$ の直接的な導出を回避して、観測データ \mathbf{y}_{obs} と k 回目の $\boldsymbol{\theta}$ の暫定値 $\boldsymbol{\theta}^{(k)}$ が与えられた下で、完全データ \mathbf{y} の対数尤度 $LL_c(\boldsymbol{\theta})$ の条件付

CONTENTS

き期待値を最大化する値としてパラメータ θ の更新値を得る以下のプロセス

$$\theta^{(k+1)} = \max_{\theta \in \Theta} E[LL_c(\theta) | \mathbf{y}, \theta^{(k)}] \quad (1.2)$$

を繰り返すことによって、目的である $LL(\theta)$ の最大化を達成する θ の最尤推定値を導くことができる。

(1.2) 式は、具体的に、E-step(Expectation Step) と M-step(Maximization Step) の2つのステップに分解され、この2つのステップを反復することから EM アルゴリズムと呼ばれている。

E-step: 観測データ \mathbf{y}_{obs} と k 回目の θ の暫定値が与えられた下で、完全データの対数尤度の条件付期待値 (Q 関数) を計算するステップ

$$Q(\theta; \theta^{(k)}) = E[LL_c(\theta) | \mathbf{y}_{obs}, \theta^{(k)}] \quad (1.3)$$

M-step: E-step で求めた $Q(\theta; \theta^{(k)})$ を最大化する $\theta^{(k+1)}$ を求めるステップ

$$\theta^{(k+1)} = \max_{\theta \in \Theta} Q(\theta; \theta^{(k)}), \quad (1.4)$$

具体的なモデルの下で、推定値の系列 $\{\theta^{(k)}\}$ の変化率および不完全データの対数尤度の系列 $LL(\theta^{(k)})$ の変化量が十分に微少になるまで、上記の E-step と M-step を交互に繰り返すことで、一般に $LL(\theta^{(k)})$ の最尤推定値 $\hat{\theta}$ を導くことができる。Dempster, Laird and Rubin (1977) と Wu(1983) には、EM アルゴリズムが $LL(\theta)$ の定常点に収束するための条件が与えられている。

もし、 $f(\mathbf{y} | \theta)$ が、多項分布、ポアソン分布、正規分布などの指数分布族の範疇に入るモデルであれば、E-step と M-step は θ の十分統計量 $\mathbf{t} = \mathbf{t}(\mathbf{x})$ を媒介にしたステップに置き換えることができ、更に分かり易くなる。

E-step: 観測データ \mathbf{y}_{obs} と k 回目の θ の暫定値 $\theta^{(k)}$ が与えられた下で、十分統計量 \mathbf{t} の条件付期待値を計算するステップ

$$\mathbf{t}^{(k+1)} = E[\mathbf{t}(\mathbf{x}) | \mathbf{y}_{obs}, \theta^{(k)}] \quad (1.5)$$

CONTENTS

M-step: 次式をみたま $\theta^{(k+1)}$ を求めるステップ, すなわち, E-step で得られた $t^{(k+1)}$ を完全データから観測された値と見做して, 通常の指数分布族における良く知られた最尤解を与える式で $\theta^{(k+1)}$ を推定するステップ

$$E[t(\boldsymbol{x}) | \boldsymbol{\theta}] = \boldsymbol{t}^{(k+1)} \quad (1.6)$$

E-step を単に個々の欠測データではなく十分統計量を条件付き期待値で代挿する, と解釈すれば, EM は前節の Rubin(1991) のフレームワークに入る. とくに, 十分統計量が個々のデータ値の線形結合で表現できる場合は, 欠測データの値自身を条件付き期待値で代挿する, と考えても差し支えない.

1.4 EM アルゴリズムと適用例

1.4.1 血液型遺伝子に関する発生確率の推定

前節で示した血液遺伝子 (A,B,O) に関する発生確率 $\boldsymbol{\theta} = (p_A, p_B, p_O)$ の推定問題の場合, 表 2.1 の観測データ (不完全データ) \boldsymbol{y}_{obs} に基づく対数尤度は次式で与えられる.

$$LL(\boldsymbol{\theta}) = 2n_O \log p_O + n_A \log(p_A^2 + 2p_A p_O) + n_B \log(p_B^2 + 2p_B p_O) + n_{AB} \log(2p_A p_B) \quad (1.7)$$

(1.7) 式を $\boldsymbol{\theta} = (p_A, p_B, p_O)$ に関して最大化するための尤度方程式は, 明示的な式で解くことはできないが, 表 2.2 の遺伝子型に対応する頻度データに基づく以下の完全データ \boldsymbol{y} の対数尤度式は, 容易に解くことができる.

$$LL_c(\boldsymbol{\theta}) = 2n_A^+ \log p_A + 2n_B^+ \log p_B + 2n_O^+ \log p_O \quad (1.8)$$

ここに, n_A^+, n_B^+, n_O^+ は, 以下で与えられる頻度である.

$$n_A^+ = n_{AA} + \frac{1}{2}n_{AO} + \frac{1}{2}n_{AB} \quad (1.9)$$

$$n_B^+ = n_{BB} + \frac{1}{2}n_{BO} + \frac{1}{2}n_{AB} \quad (1.10)$$

$$n_O^+ = n_O + \frac{1}{2}n_{AO} + \frac{1}{2}n_{BO}. \quad (1.11)$$

CONTENTS

EM アルゴリズムをこの推定問題で具体的に構築する. E-step は, (1.3) 式の $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) = E[LL_c(\boldsymbol{\theta}) | \mathbf{y}_{obs}, \boldsymbol{\theta}^{(k)}]$ を具体的に評価していると考えられることもできるし, (2.7) 式が多項分布であることから, 十分統計量 $\{n_A^+, n_B^+, n_O^+\}$ を評価していると考えても良い.

E-step: 観測データとパラメータの k 回目の推定値が与えられた下で, 完全データの条件付き期待値を計算する.

$$n_{AA}^{(k)} \equiv E[n_{AA} | n_A; p_A^{(k)}, p_O^{(k)}] = \frac{n_A p_A^{(k)^2}}{p_A^{(k)^2} + 2p_A^{(k)} p_O^{(k)}} \quad (1.12)$$

$n_{AO}, n_{BB}, n_{BO}, n_{AB}$ の条件付き期待値も同様の式で計算する.

M-step: E-step で求めた値 (完全データの推定値) を使って, $LL_c(\boldsymbol{\theta})$ を最大化する $\boldsymbol{\theta}^{(k+1)}$ を求め, パラメータの推定値を更新する.

$$p_A^{(k+1)} = \frac{n_A^{+(k)}}{n} = \frac{n_{AA}^{(k)} + \frac{1}{2}n_{AO}^{(k)} + \frac{1}{2}n_{AB}^{(k)}}{n} \quad (1.13)$$

$$p_B^{(k+1)} = \frac{n_B^{+(k)}}{n}, \quad p_O^{(k+1)} = \frac{n_O^{+(k)}}{n}. \quad (1.14)$$

適当な初期値から始めて上記の E-step と M-step を反復することにより, とくに制約を設けることなく, 確率の定義域内で目的である血液遺伝子の発生確率を求めることができる. 例えば, $n_A = 20, n_B = 10, n_{AB} = 5, n_O = 10$ が観測されたとして, 初期値 $p_A = 0.3, p_B = 0.3, p_O = 0.4$ の下に EM アルゴリズムをスタートすると, 表 (1.3) のように値が収束していく.

1.4.2 多変量データにおける欠測：多変量正規モデル

この節では, p 次の多変量正規確率変数ベクトル $\mathbf{X} = (X_1, \dots, X_p)'$ の平均ベクトルが $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)'$, 分散共分散行列が $\boldsymbol{\Sigma} = (\sigma_{jk})$ のとき, 欠測値を含む大きさ n の標本から $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の最尤推定値を求める問題を考える.

CONTENTS

Table 1.3: 収束の様子

| 反復回数 | p_A | p_B | $p_O = 1 - p_A - p_B$ |
|------|--------|--------|-----------------------|
| 初期値 | 0.3000 | 0.3000 | 0.4000 |
| 1 | 0.3384 | 0.1970 | 0.4646 |
| 2 | 0.3371 | 0.1861 | 0.4768 |
| 3 | 0.3358 | 0.1848 | 0.4794 |
| 4 | 0.3354 | 0.1846 | 0.4800 |
| 5 | 0.3353 | 0.1846 | 0.4801 |
| 6 | 0.3353 | 0.1846 | 0.4801 |

第 i 番目のケースの観測値 \mathbf{x}_i は、欠測に応じて以下のように分解されているとする.

$$\mathbf{x}_i = \begin{bmatrix} \mathbf{x}_i^0 \\ \mathbf{x}_i^1 \end{bmatrix} \equiv \begin{bmatrix} \mathbf{x}_{i,mis} \\ \mathbf{x}_{i,obs} \end{bmatrix}$$

この欠測パターンに応じたパラメータ $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の分解を

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_0 \\ \boldsymbol{\mu}_1 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{00} & \boldsymbol{\Sigma}_{01} \\ \boldsymbol{\Sigma}_{10} & \boldsymbol{\Sigma}_{11} \end{bmatrix}$$

とする. このとき, $\{\mathbf{x}_{i,obs}\}, i = 1, 2, \dots, n$ に基づくパラメータ $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の最尤推定値を導く EM アルゴリズムは, 以下のように具体化される.

E-step : 十分統計量 $(\mathbf{x}_i, \mathbf{x}_i \mathbf{x}_i^T)$ の条件付期待値を計算するステップ

$$\begin{aligned} E[\mathbf{x}_{i,mis} \mid \mathbf{x}_{i,obs}; \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}] &= \boldsymbol{\mu}_0^{(k)} + \boldsymbol{\Sigma}_{01}^{(k)} \boldsymbol{\Sigma}_{11}^{(k)-1} (\mathbf{x}_{i,obs} - \boldsymbol{\mu}_1^{(k)}) \equiv \mathbf{x}_{i,mis}^{(k)} \\ E[\mathbf{x}_{i,mis} \mathbf{x}_{i,obs}^T \mid \mathbf{x}_{i,obs}; \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}] &= \mathbf{x}_{i,mis}^{(k)} \mathbf{x}_{i,obs}^T \\ E[\mathbf{x}_{i,mis} \mathbf{x}_{i,mis}^T \mid \mathbf{x}_{i,obs}; \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}] &= \mathbf{x}_{i,mis}^{(k)} \mathbf{x}_{i,mis}^{(k)T} + \boldsymbol{\Sigma}_{00}^{(k)} - \boldsymbol{\Sigma}_{01}^{(k)} \boldsymbol{\Sigma}_{11}^{(k)-1} \boldsymbol{\Sigma}_{10}^{(k)} \\ &\equiv [\mathbf{x}_{i,mis} \mathbf{x}_{i,mis}^T]^{(k)} \end{aligned} \quad (1.15)$$

CONTENTS

上記の結果を利用した十分統計量 $(\mathbf{x}_i, \mathbf{x}_i \mathbf{x}_i^T)$ の k 回目の反復値

$$\mathbf{x}_i^{(k)} \equiv \begin{bmatrix} \mathbf{x}_{i,mis}^{(k)} \\ \mathbf{x}_{i,obs} \end{bmatrix}, \quad [\mathbf{x}_i \mathbf{x}_i^T]^{(k)} \equiv \begin{bmatrix} [\mathbf{x}_{i,mis} \mathbf{x}_{i,mis}^T]^{(k)} & \mathbf{x}_{i,mis}^{(k)} \mathbf{x}_{i,obs}^T \\ \mathbf{x}_{i,obs} \mathbf{x}_{i,mis}^{(k)T} & \mathbf{x}_{i,obs} \mathbf{x}_{i,obs}^T \end{bmatrix}$$

に対して、次の M-step の計算が適用される。

M-step : E-step で求めた十分統計量の期待値を完全データからの観測結果と見做して、 $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の通常の最尤解を計算するステップ

$$\boldsymbol{\mu}^{(k+1)} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^{(k)} \quad (1.16)$$

$$\boldsymbol{\Sigma}^{(k+1)} = \frac{1}{N} \sum_{i=1}^n [\mathbf{x}_i \mathbf{x}_i^T]^{(k)} - \boldsymbol{\mu}^{(k+1)} \boldsymbol{\mu}^{(k+1)T} \quad (1.17)$$

E-step と M-step の反復により、とくに制約を設けることなく、 $\boldsymbol{\Sigma}$ に関して正定値性が保証された下で、最尤解を求めることができる。

表 1.4 を 2 変量正規分布からの標本と仮定し、EM アルゴリズムによって平均と分散、共分散の最尤推定値を求めてみる。ここで、表中の * は欠測を意味し、欠測は Rubin(1976) の意味でランダムであるとする。初期値として、データが完全に得られている 3 つのケース 1,2,6 から計算した値を使って E-step と M-step を反復させた収束の様子が、表 1.5 である。

Table 1.4: 欠測値を含む 2 変量データ

| No. | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|-------|----|----|----|----|---|----|---|
| X_1 | 5 | 8 | * | * | 4 | 10 | 6 |
| X_2 | 10 | 20 | 15 | 40 | * | 60 | * |

Table 1.5: 収束の様子

| 反復回数 | μ_1 | μ_2 | σ_1 | σ_2 | σ_{12} |
|----------|----------|----------|------------|------------|---------------|
| 初期値 | 7.667 | 30.000 | 6.333 | 700.000 | 40.000 |
| 1 | 6.864 | 24.474 | 4.791 | 436.150 | 29.996 |
| 2 | 6.735 | 24.372 | 4.350 | 381.408 | 30.286 |
| \vdots | \vdots | \vdots | \vdots | \vdots | \vdots |
| 34 | 6.767 | 22.861 | 3.995 | 385.525 | 34.695 |
| 35 | 6.767 | 22.861 | 3.995 | 385.525 | 34.695 |

1.4.3 重回帰モデル：目的変数に欠測がある場合

重回帰モデル $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, $\mathbf{e} \sim N(0, \sigma^2 \mathbf{I})$ で、目的変数 \mathbf{y} に欠測データが含まれる場合のパラメータの最尤推定値を求める問題を考える。この問題は、 \mathbf{X} が計画行列の場合には、意図された不完備実験計画および予期せぬ欠測などの理由でバランスの崩れた実験データの推測にも適用できる。

観測と欠測に応じて、変数 \mathbf{y} を観測変数 \mathbf{y}_{obs} , 欠測変数 \mathbf{y}_{mis} に分解する。ここでは、欠測はランダムであると仮定した下で、 \mathbf{y}_{obs} に基づく回帰パラメータ $\boldsymbol{\beta}$ と誤差分散 σ^2 の推定を EM アルゴリズムで以下に具体化する。ここで、 \mathbf{x}_i^T は説明変数行列 \mathbf{X} の第 i 行ベクトルとする。

E-step : 観測されている説明変数 \mathbf{x}_i の値と回帰パラメータ $\boldsymbol{\beta}$, 誤差分散 σ^2 の k 回目の推定値が与えられた下で、目的変数が欠測したケースに対して、十分統計量 (y_i, y_i^2) の期待値を計算する。

$$y_{i,mis}^{(k)} \equiv E[y_i \mid \mathbf{x}_i, \boldsymbol{\beta}^{(k)}, \sigma^{2(k)}] = \mathbf{x}_i^T \boldsymbol{\beta}^{(k)} \quad (1.18)$$

$$y_{i,mis}^{2(k)} \equiv E[y_i^2 \mid \mathbf{x}_i, \boldsymbol{\beta}^{(k)}, \sigma^{2(k)}] = (\mathbf{x}_i^T \boldsymbol{\beta}^{(k)})^2 + \sigma^{2(k)} \quad (1.19)$$

すべてのケースに対して十分統計量 (y_i, y_i^2) の k 回目の反復値を以下で定義

CONTENTS

する.

$$y_i^{(k)} \equiv \begin{cases} y_{i,obs}, & \text{if } y_i \text{が観測,} \\ y_{i,miss}^{(k)}, & \text{if } y_i \text{が欠測,} \end{cases} \quad (1.20)$$

$$y_i^{2(k)} \equiv \begin{cases} y_{i,obs}^2, & \text{if } y_i \text{が観測,} \\ y_{i,miss}^{2(k)}, & \text{if } y_i \text{が欠測.} \end{cases} \quad (1.21)$$

M-step : E-step で求めた値を使って, 通常の方法で目的パラメータの推定を行う.

$$\boldsymbol{\beta}^{(k+1)} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X} \mathbf{y}^{(k)} \quad (1.22)$$

$$\sigma^{2(k+1)} = \frac{\sum_{i=1}^n [y_i^{2(k)} - (\boldsymbol{\beta}^{(k)} \mathbf{x}_i)^2]}{n} \quad (1.23)$$

この場合も, σ^2 は, 正の値の範囲で推定される.

1.4.4 中途打ち切りデータに基づく例

$(n + m)$ 個の個体があり, n 個の個体はその生存時間 y_1, \dots, y_n が観測され, 残り m 個の個体については t 時間の観察の後, 生存か死亡のいずれかの結果 s_1, \dots, s_m だけが観測されている状況を考える. ここで, $s_i = 1$ は生存, $s_i = 0$ は死亡を意味する. また, ℓ は, 生存時間が観測されなかった m 個の個体のうち, 実験打ち切り時間 t において生存していた個体数, すなわち $\ell = \sum_{i=1}^m s_i$ とする.

Flury and Zoppe (2000) は, この設定で EM アルゴリズムが適用できる場合と適用できない場合の 2 つの例を与えている.

指数分布を仮定する場合 : EM が適用できる例

対象としている個体の生存時間が平均 θ を持つ指数分布に従っていると仮定すると, いま観測されている不完全データの対数尤度は,

$$LL(\theta) = -n(\log \theta + \bar{y}/\theta) - \ell t/\theta + (m - \ell) \log(1 - e^{-t/\theta}) \quad (1.24)$$

CONTENTS

となる. ここで, \bar{y} は y_1, \dots, y_n の平均を表す. $\ell = m$ のとき, (1.24) 式を最大化する θ の解は, 代数的に求めることができ, 最尤推定量は, $\hat{\theta} = (n\bar{y} + mt)/n$ となる. また, $\ell \neq m$ のときでも, (1.24) 式の最大化を通常最適化アルゴリズムで達成することは難しくはない. しかし, ここでは, この問題を EM アルゴリズムで考えてみる.

いま生存時間が正確に観測されなかった m 個の個体に対して, 本来観測されたであろう生存時間 (欠測データ) を x_1, \dots, x_m とし, 完全データを $y_1, \dots, y_n, x_1, \dots, x_m$ と想定すると, 完全データの対数尤度は

$$LL_c(\theta) = -n(\log \theta + \bar{y}/\theta) - \sum_{i=1}^m (\log \theta + x_i/\theta). \quad (1.25)$$

となる. (1.25) 式は x_i の線形関数なので, EM アルゴリズムは, 欠測データ x_i 自身の条件付期待値を介して構築される.

E-step: 不完全データとパラメータの k 回目の推定値 $\theta^{(k)}$ が与えられた下, 欠測データ x_i の条件付期待値を計算する.

$$x_i^{(k)} \equiv E[x_i | s_i, \theta^{(k)}] = \begin{cases} t + \theta^{(k)}, & \text{if } s_i = 1, \\ \theta^{(k)} - (te^{-t/\theta^{(k)}})/(1 - e^{-t/\theta^{(k)}}), & \text{if } s_i = 0. \end{cases} \quad (1.26)$$

(1.25) 式の x_i に (1.26) の $x_i^{(k)}$ を代入して,

$$\begin{aligned} Q(\theta; \theta^{(k)}) &= E[LL_c(\theta) | y_i, s_i, \theta^{(k)}] \\ &= -(n+m) \log \theta - \frac{1}{\theta} \{n\bar{y} + \ell(t + \theta^{(k)}) + (m-\ell)(\theta^{(k)} - tp^{(k)})\} \end{aligned} \quad (1.27)$$

を得る. $p^{(k)} = e^{-t/\theta^{(k)}}/(1 - e^{-t/\theta^{(k)}})$.

M-step: (1.27) 式を最大化する θ を求める.

$$\theta^{(k+1)} = \frac{n\bar{y} + \ell(t + \theta^{(k)}) + (m-\ell)(\theta^{(k)} - tp^{(k)})}{n+m} \quad (1.28)$$

適当な正の値を初期値とし, E-step と M-step を繰り返すことで, θ の最尤推定値を得ることができる. とくに, $\ell = m$ のとき, (1.28) 式の θ に関する自己一致解は, (1.24) 式から導いた最尤解 $\hat{\theta} = (n\bar{y} + mt)/n$ と一致する.

CONTENTS

(0, \theta] 上の一様分布を仮定する場合：EM が適用できない例

前節と同様な生存時間の観測状況に対して，生存時間が (0, \theta] 上の一様分布に従う場合のパラメータ \theta の最尤推定を考える。

仮に完全データ x_i が得られたとした場合の最尤推定値は $\max\{x_{max}, y_{max}\}$ である。 x_{max}, y_{max} は，それぞれ $\{x_i\}_{1 \leq i \leq m}, \{y_i\}_{1 \leq i \leq n}$ の最大値。

ここで，このまま先述の指数分布の場合のように， $E[x_i | s_i, \theta^{(k)}]$ を介した EM アルゴリズムを構築すると， $\ell \geq 1$ の場合は， $\theta \geq t$ で， E-step は，

$$E[x_i | s_i, \theta^{(k)}] = \begin{cases} (t + \theta^{(k)})/2, & \text{if } s_i = 1, \\ t/2, & \text{if } s_i = 0. \end{cases} \quad (1.29)$$

M-step は，

$$\theta^{(k+1)} = \max\{y_{max}, (t + \theta^{(k)})/2\} \quad (1.30)$$

となり，この E-step と M-step の反復結果は， $\max\{y_{max}, t\}$ に到達する。しかしこれは，このモデルの下での最尤推定値とはならない。

この場合の観測データ（不完全データ）に基づく尤度は，

$$L(\theta) = \theta^{-N} I_{[y_{max}, \infty)}(\theta) (t / \max(t, \theta))^{m-\ell} (1 - t / \max(t, \theta))^\ell \quad (1.31)$$

で，とくに $\ell = 0$ の場合，

$$L(\theta) = \theta^{-n} I_{[y_{max}, \infty)}(\theta) (t / \max(t, \theta))^m \quad (1.32)$$

となり， $\theta \geq y_{max}$ の範囲で減少関数なので，最尤推定値は y_{max} となる。

$\ell \geq 1$ の場合は， $\theta \geq t$ で， $\theta \geq t$ のとき，関数 $h(\theta) = \theta^{-(n+m)}(\theta - t)^\ell$ は， $\dot{\theta} = \frac{n+m}{n+m-\ell}t$ で最大値をとり， $\theta > \dot{\theta}$ の範囲で単調減少関数なので，(1.31) 式の尤度は， $\dot{\theta} > y_{max}$ のとき $\dot{\theta}$ で，また， $\dot{\theta} < y_{max}$ のとき y_{max} で最大となる。したがって，一様分布の下での θ の最尤推定値 $\hat{\theta}$ は，

$$\hat{\theta} = \begin{cases} \dot{\theta}, & \text{if } \dot{\theta} > y_{max} \text{ and } \ell \geq 1, \\ y_{max}, & \text{その他,} \end{cases} \quad (1.33)$$

となる。

CONTENTS

ここで、前述の (1.29) 式と (1.30) 式による EM アルゴリズムが正しく最尤推定値を導かなかった理由は、すべての $\theta > 0$ について対数尤度関数が存在するわけではないからである。例えば、生存時間が観測されなかった 1 つの個体の生存時間 x_i が t を超えている場合、この x_i の確率密度は

$$f_x(x_i; \theta) = \begin{cases} 1/\theta, & \text{if } 0 \leq x_i \leq \theta, \\ 0, & \text{その他,} \end{cases} \quad (1.34)$$

となる。ここで、 $x_i > t$ の範囲での条件付分布が $[t, \theta^{(k)}]$ の一様分布となるため、 $\theta < \theta^{(k)}$ となる θ においては、密度関数の値は 0 となる。したがって、 $\theta^{(k)}$ が与えられたときの x_i の条件付期待値は存在しないことになる。

この例のように、ある範囲において尤度の値が 0 となる場合、対数尤度の期待値を求めることができず、EM アルゴリズムが適用できない。EM アルゴリズムにおける E-step は、不完全データの補完ではなく、あくまで対数尤度の期待値を求めるステップであることに注意しなければならない。

1.4.5 混合分布モデル

潜在クラスモデルをはじめとする有限混合分布モデルでは、最尤推定に EM アルゴリズムが利用されることが多い。いま、一般に p 個のコンポーネントに対して、各コンポーネントの分布が $f_j(x; \theta_j)$ 、構成率が π_j の混合分布モデル

$$\sum_{j=1}^p \pi_j f_j(x; \theta_j), \quad \text{ここに, } \sum_{j=1}^p \pi_j = 1 \quad (1.35)$$

を想定する。

実際に n 個の対象に対して観測されるデータは x_1, \dots, x_n であるが、更に各対象のコンポーネントへの所属を示す 0-1 型のデータ $\mathbf{z}_i = (z_{i1}, \dots, z_{ip})$ の存在を仮定し、 $(x_1, \mathbf{z}_1), \dots, (x_n, \mathbf{z}_n)$ を観測されるべき完全データと見做す。ここで、 z_{ij} は、第 i 個体が第 j コンポーネントに所属する場合に 1 をとり、それ以外には 0 をとる。ただし、すべての \mathbf{z}_i は欠測して観測されない、いわゆる潜在データと考える。

CONTENTS

すなわち, (1.35) 式のモデルは, 潜在データ z_{ij} を導入することで,

$$x_i | (z_{ij} = 1) \sim f_j(x_i; \boldsymbol{\theta}_j), \quad \text{ただし, } Pr(z_{ij} = 1) = \pi_j \quad (1.36)$$

と書き直すことができる.

ここで, 完全データ $(x_i, \mathbf{z}_i), i = 1, \dots, n$ に基づく対数尤度は,

$$LL_c = \sum_{i=1}^n \sum_{j=1}^p z_{ij} \log \pi_j + \sum_{i=1}^n \sum_{j=1}^p z_{ij} \log f_j(x_i; \boldsymbol{\theta}_j) \quad (1.37)$$

となる. E-step では対数尤度 (1.37) の条件付期待値を求めることになるが, z_{ij} の線形関数であるため, z_{ij} の条件付期待値だけを求めればよいことが分かる. ベイズの定理による事後確率の評価から, 具体的に, E-step の一般形は以下となる.

E-step

$$\begin{aligned} w_{ij}^{(k)} &\equiv E[z_{ij} | x_i; \pi_1^{(k)}, \dots, \pi_p^{(k)}, \boldsymbol{\theta}^{(k)}] \\ &= Pr(z_{ij} = 1 | x_i; \pi_1^{(k)}, \dots, \pi_p^{(k)}, \boldsymbol{\theta}^{(k)}) \\ &= \pi_j f_j(x_i; \boldsymbol{\theta}_j) / \left\{ \sum_{\ell=1}^p \pi_\ell f_\ell(x_i; \boldsymbol{\theta}_\ell) \right\} \end{aligned} \quad (1.38)$$

M-step では, (1.37) 式の z_{ij} を E-step で求めた条件付期待値 $w_{ij}^{(k)}$ に置き換え最大化を達成するパラメータの値を求めるが, 混合割合 π_j を示すパラメータと第 j コンポーネントの分布を規定するパラメータ $\boldsymbol{\theta}_j$ は, それぞれ (1.37) 式右辺の第 1 項と第 2 項に分かれているので, それぞれ別々に最大化を行えばよいことになる.

M-step

$$\pi_j^{(k+1)} = \sum_{i=1}^n w_{ij}^{(k)} / n \quad (1.39)$$

パラメータ $\boldsymbol{\theta}_j$ に関する最大化は, コンポーネントごとに, 分布 $f_j(X; \boldsymbol{\theta}_j)$ のパラメータ $\boldsymbol{\theta}_j$ の $w_{ij}^{(k)}$ を重みとした重み付き推定に帰着する.

CONTENTS

混合正規分布のパラメータの推定

$$x \sim \sum_{j=1}^p \pi_j N(\mu_j, \sigma_j^2) \quad (1.40)$$

に対して、 μ_j, σ_j^2 の最尤推定値を求めるための EM アルゴリズムの具体的計算は以下ようになる：

E-step

$$w_{ij}^{(k)} = \frac{\pi_j^{(k)} \phi(x_i; \mu_j^{(k)}, \sigma_j^{(k)2})}{\sum_{\ell=1}^p \pi_{\ell}^{(k)} \phi(x_i; \mu_{\ell}^{(k)}, \sigma_{\ell}^{(k)2})}, \quad \phi(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) \quad (1.41)$$

M-step

$$\mu_j^{(k+1)} = \frac{\sum_{i=1}^n w_{ij}^{(k)} x_i}{\sum_{i=1}^n w_{ji}^{(k)}}, \quad (1.42)$$

$$\sigma_j^{2(k+1)} = \frac{\sum_{i=1}^n w_{ij}^{(k)} (x_i - \mu_j^{(k)})^2}{\sum_{i=1}^n w_{ji}^{(k)}}. \quad (1.43)$$

E-step と M-step の反復により、目的とするパラメータの収束値を得る。

混合正規分布の場合、平均 μ を観測されたデータの中のある 1 つの値に固定し、分散 σ^2 を 0 に近づけると、尤度が発散するという問題の存在が知られているが、これは、分散を一定以上の値にしておけば、回避ができる。上記の EM アルゴリズムの場合、分散が 0 より大きいという制約が自然と導入されており、尤度が発散するという状況は避けられている。

表 1.6 は、2005 年のシーズン中に松坂 (当時西武ライオンズ) が投げたスライダーに分類された球の球速の度数分布である。図 1.4.5 は、2 つの正規分布の混合モデルをあてはめた結果を表している。最尤推定の結果、全体の球速の分布は、93% が平均が時速 128.75km、分散 3.31、残りの 7% は平均が時速 136.86km、分散 1.74 の 2 つの正規分布の混合に従うと考えられる。

CONTENTS

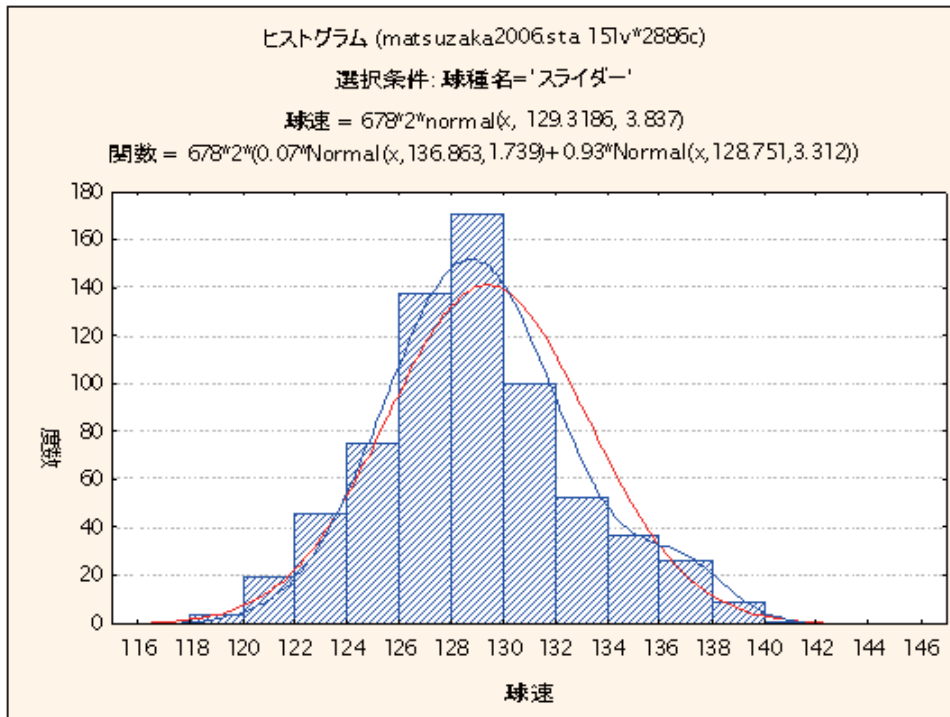


Figure 1.1: 混合正規分布のあてはめ

CONTENTS

Table 1.6: 松坂投手の球速分布 (2005年, 球種:スライダー)

| 球速 (km/h) | 球数 | 球速 (km/h) | 球数 | 球速 (km/h) | 球数 |
|-----------|----|-----------|-----|-----------|-----|
| 119 | 2 | 126 | 55 | 133 | 34 |
| 120 | 1 | 127 | 95 | 134 | 25 |
| 121 | 6 | 128 | 101 | 135 | 30 |
| 122 | 8 | 129 | 89 | 136 | 12 |
| 123 | 15 | 130 | 71 | 137 | 22 |
| 124 | 42 | 131 | 71 | 138 | 14 |
| 125 | 64 | 132 | 51 | 139 | 6 |
| 合計 | | | | | 814 |

1.5 EM を利用した漸近分散共分散行列の評価

EM アルゴリズムでは Newton-Raphson 法と異なりアルゴリズム自身の副産物として推定値の漸近分散共分散行列の推定値を与えない。したがって、直接的には、不完全データの対数尤度 $LL(\boldsymbol{\theta} | \mathbf{y})$ から導かれる観測情報行列の逆行列を数値的に評価する方法で、漸近分散共分散行列を推定しなければならないが、これはモデルによっては導出が困難になる。

しかし、不完全データの対数尤度 $LL(\boldsymbol{\theta} | \mathbf{y})$ を直接意識しないという EM アルゴリズムのメリットを保持したもとの漸近分散共分散行列を近似する手法もいくつか提唱されている。

1.5.1 Louis の方法

Louis(1982) は、以下のような不完全データの観測情報行列 $\mathbf{I}(\hat{\boldsymbol{\theta}}; \mathbf{y})$ を完全データのフレームで評価する式：

$$\mathbf{I}(\hat{\boldsymbol{\theta}}; \mathbf{y}) = E[\mathbf{B}_c(\mathbf{x}; \boldsymbol{\theta}) | \mathbf{y}; \boldsymbol{\theta}]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} - E[\mathbf{S}_c(\mathbf{x} | \boldsymbol{\theta}) \mathbf{S}_c^T(\mathbf{x} | \boldsymbol{\theta}) | \mathbf{y}; \boldsymbol{\theta}]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \quad (1.44)$$

CONTENTS

を示した. ここに, $S_c(x; \theta)$ は完全データ \mathbf{x} の対数尤度の1次導関数ベクトル, $B_c(x; \theta)$ は2次導関数行列の負値である.

(1.44) 式は, いわゆる Orchard and Woodbury (1972) による missing information principle に対応するもので,

Observed information = Complete information - Missing information
を表している. EM アルゴリズムにより $\hat{\theta}$ を得た後, (1.44) を評価すれば観測情報量を得ることができ, 漸近分散共分散行列はその逆行列の計算により導くことができる.

数値例として, Dempster, Laird and Rubin(1977) が取り上げている遺伝連鎖モデル (Genetic Linkage Model) を取り上げる. パラメータ θ を含むセル確率

$$\left(\frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1 - \theta), \frac{1}{4}(1 - \theta), \frac{\theta}{4}\right)$$

をもつ4つのカテゴリに対して, 度数が $\mathbf{y} = (y_1, y_2, y_3, y_4)$ のように観測される多項モデルがある. これを不完全データと見做し, 実際は, セル確率

$$\left(\frac{1}{2}, \frac{\theta}{4}, \frac{1}{4}(1 - \theta), \frac{1}{4}(1 - \theta), \frac{\theta}{4}\right)$$

を有する5つのカテゴリがあるとして, これに対応する度数 $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$ を完全データとして想定する. ここで, $x_1 + x_2 = y_1, x_3 = y_2, x_4 = y_3, x_5 = y_4$ と考えると, θ の推定は容易に EM アルゴリズムのフレームで構築できる.

E-step: 完全データの対数尤度の条件付期待値 $E[LL_c(\theta) | \mathbf{y}, \theta^{(k)}]$ の導出

$$\begin{aligned} Q(\theta; \theta^{(k)}) &= E[LL_c(\theta) | \mathbf{y}, \theta^{(k)}] = E[(x_2 + x_5) \log(\theta) + (x_3 + x_4) \log(1 - \theta) | \mathbf{y}, \theta^{(k)}] \\ &= \{E(x_2 | y_1, \theta^{(k)}) + y_4\} \log(\theta) + (y_2 + y_3) \log(1 - \theta) \end{aligned} \quad (1.45)$$

M-step: E-step で求めた $Q(\theta; \theta^{(k)})$ を最大化する $\theta^{(k+1)}$ を求める

$$\theta^{(k+1)} = \frac{E[x_2 | y_1, \theta^{(k)}] + y_4}{E[x_2 | y_1, \theta^{(k)}] + y_2 + y_3 + y_4} \quad (1.46)$$

Rao(1965) による観測データ $\mathbf{y} = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$ に対して, 上記の E-step と M-step の反復により収束値 0.6268 が最尤推定値 $\hat{\theta}$ として求められる (初期値 0.5).

CONTENTS

このとき, (1.44) 式の右辺の第 1 項と第 2 項はそれぞれ

$$\begin{aligned} E[\mathbf{B}_c(\mathbf{x}; \boldsymbol{\theta}) | \mathbf{y}; \boldsymbol{\theta}]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} &= \frac{E[x_2 | y_1, \hat{\theta}] + y_4}{\hat{\theta}^2} + \frac{y_2 + y_3}{(1 - \hat{\theta})^2} \\ &= \frac{(29.83 + 34)}{0.6268^2} + \frac{38}{(1 - 0.6268)^2} = 435.3, \end{aligned} \quad (1.47)$$

$$\begin{aligned} E[\mathbf{S}_c(\mathbf{x} | \boldsymbol{\theta}) \mathbf{S}_c^T(\mathbf{x} | \boldsymbol{\theta}) | \mathbf{y}; \boldsymbol{\theta}]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} &= \text{Var}[\mathbf{S}_c(\mathbf{x} | \boldsymbol{\theta}) | \mathbf{y}, \boldsymbol{\theta}]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\ &= \text{Var}\left[\left(\frac{E[x_2 | y_1, \boldsymbol{\theta}] + y_4}{\boldsymbol{\theta}} - \frac{y_2 + y_3}{1 - \boldsymbol{\theta}}\right) | \mathbf{y}, \boldsymbol{\theta}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\ &= \frac{\text{Var}[E[x_2 | y_1, \hat{\theta}]]}{\hat{\theta}^2} \\ &= \frac{y_1 \frac{\hat{\theta}}{(\frac{1}{2} + \frac{\hat{\theta}}{4})^2}}{\hat{\theta}^2} = 57.8 \end{aligned} \quad (1.48)$$

となり, $\mathbf{I}(\hat{\boldsymbol{\theta}}; \mathbf{y}) = 435.3 - 57.8 = 377.5$ を得る. したがって, $\hat{\theta}$ の標準誤差は, $\sqrt{\frac{1}{377.5}} \approx 0.05$ となる.

1.5.2 Oakes の方法

Oakes(1999) は, 不完全データに基づく観測情報行列 $\mathbf{I}(\boldsymbol{\theta}; \mathbf{y})$ を $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$ の 2 次導関数で表現する以下の式を与えた.

$$\mathbf{I}(\boldsymbol{\theta}^{(k)}; \mathbf{y}) = - \left[\left(\frac{\partial^2 Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} + \frac{\partial^2 Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{(k)T}} \right)_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}} \right] \quad (1.49)$$

ここで右辺の第 2 項は, k 回目の $\boldsymbol{\theta}$ の推定値 $\boldsymbol{\theta}^{(k)}$ で評価した missing information に対応している.

先の遺伝連鎖モデルの例では,

$$\frac{\partial^2 Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{(k)T}} = \frac{2y_1}{\theta (2 + \theta^{(k)})^2} \quad (1.50)$$

かつ, (1.47) 式より,

$$\frac{\partial^2 Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \frac{y_1 \frac{\theta}{2+\theta} + y_4}{\theta^2} + \frac{y_2 + y_3}{(1 - \theta)^2} \quad (1.51)$$

CONTENTS

となり, これらを $\theta^{(k)}, \theta$ とも $\hat{\theta} = 0.6268$ で評価すると, 前節でも求めた $I(\hat{\theta}; \mathbf{y}) = 435.3 - 57.8 = 377.5$ を得る.

1.5.3 SEM(Supplemented EM)

Meng and Rubin(1991) は, 不完全データの漸近分散共分散行列 $I^{-1}(\hat{\theta}; \mathbf{y})$ が完全データの分散共分散行列に欠測データが存在することによる増分を加えることで評価できるとして次式を導いた.

$$I^{-1}(\hat{\theta}; \mathbf{y}) = I_c^{-1}(\hat{\theta}; \mathbf{y}) + \Delta V, \quad (1.52)$$

ここに, $I_c(\hat{\theta}; \mathbf{y})$ は完全データの情報行列の条件付き期待値 $E[B_c(\mathbf{x}; \theta) | \mathbf{y}; \theta]$ を表し,

$$\Delta V = [I - J(\hat{\theta})]^{-1} J(\hat{\theta})(\theta; \mathbf{y}) I_c^{-1}(\hat{\theta}; \mathbf{y}), \quad (1.53)$$

ここで, $J(\hat{\theta}) = I_c(\hat{\theta}; \mathbf{y}) I_m^{-1}(\hat{\theta}; \mathbf{y})$ である.

(1.52) 式は, $I_c(\hat{\theta}; \mathbf{y})$ とヤコビアン $J(\hat{\theta})$ を計算することで求められる値である. $I_c(\hat{\theta}; \mathbf{y})$ は先述のように, 完全データにに基づく尤度式から比較的容易に求めることができる. また, ヤコビアン行列 $J(\hat{\theta})$ はパラメータの成分毎の収束率を各要素としているので, EM アルゴリズム自身の出力コードを収束後, 再評価することでこれも容易に計算できる量である.

1.6 EM の特性

ここで, EM アルゴリズムの特徴や収束に関しこれまでに得られている結果をまとめておく.

1. 反復の各過程において $LL(\theta | y)$ の値は増加 (正確には, 非減少) する. これは実際にプログラミングを行う場合, デバッキングに有効である. また, 初期値に関して頑健である.

CONTENTS

2. Newton-Raphson 法が 2 次収束, 準 Newton 法が超 1 次収束するのに対して, EM アルゴリズムは 1 次収束であり, 収束率 γ は, 完全データの情報行列に対する不完全データの情報行列の比 $\tilde{I}_c^{-1}(\boldsymbol{\theta}; \mathbf{y})\tilde{I}_m(\boldsymbol{\theta}; \mathbf{y})$ の最大固有値 (完全データの情報量に占める不完全データの情報量の割合) となる. このとき, 収束のスピード (global speed of convergence) は, $1 - \gamma$ に比例する.
3. EM アルゴリズムでは, とくに完全データが指数分布族の範疇であれば, 各ステップにおける出力が統計的に意味のあるものになっている. すなわち, 自然にパラメータの定義域や制約条件が充たされ, かつ容易に制約条件を組み込んで解を得ることができる. いわゆる不適解が出ない. 例えば, 欠測値のある多次元データ行列から分散共分散行列を推定する場合も初期値を正定値行列から出発すれば各ステップで得られる $\hat{\Sigma}$ も正定値であることが保証される. また, 生起確率の推定では自然に確率の定義域に納まる解が得られる.
4. $LL(\boldsymbol{\theta} | y)$ が有界であれば, 反復の過程での対数尤度の値 $\{LL(\boldsymbol{\theta}^{(k)} | y)\}$ は $LL(\boldsymbol{\theta} | y)$ の定常点 (stationary value) に収束する.
5. かなり一般的な条件の下で, $\boldsymbol{\theta}^{(k)}$ が収束すればその収束値は $LL(\boldsymbol{\theta} | y)$ の局所的最大値 (local maximum) か鞍点 (saddle point) であることが証明できる (Boyle 1983, Wu 1983). したがって, 尤度関数が単峰で, Q 関数の 1 次導関数が $\boldsymbol{\theta}^{(k)}$ と $\boldsymbol{\theta}$ に関して連続であれば, EM は唯一の極大値 (最大値) に収束する. しかし, 一般に不完全データの尤度関数が単峰であるとは限らない. したがって, 多くの初期値を試みる必要がある.
6. EM の収束が極端に遅くなる場合は, その方向に尤度関数がフラットになっていることを示唆している.
7. EM アルゴリズムでは, 一般に計算量が大きい不完全データの対数尤度の 1 次および 2 次の導関数の評価が不要である. このことが多くの場

CONTENTS

合、他の最大化アルゴリズムと比較して各反復における CPU タイムの節約につながる。即ち、EM アルゴリズムの欠点である収束の遅さは、収束に要する反復回数を基準として見ているので、全体的な CPU-time を基準に採れば、各反復における高次元の逆行列演算の回避を考慮するとき、Newton-Raphson 法等との実際的な収束の速さに関する優劣は一概には定まらない。準 Newton 法も、不完全データの情報行列の計算を含まないので、EM アルゴリズム同様 1 回の反復における計算時間は、Newton-Raphson 法より短く、かつ、収束に要する反復回数は、超 1 次収束する準 Newton 法の方が 1 次収束する EM アルゴリズムよりはるかに少なく済む。従って問題によっては、反復の初期の段階は EM アルゴリズム、その後、準 Newton 法に切り替える等の使い方も有効である。

1.7 GEM アルゴリズムとその他の拡張型 EM

1.7.1 GEM(Generalized EM)

EM アルゴリズムの利点は、適用モデルの汎用性にあるが、対象とするモデルの構造が複雑になれば、E-step でモデルから発生させた乱数による評価が必要になったり、M-step の中に更に Newton 法などの反復アルゴリズムを含まざるを得なくなる場合も出てくる。その際、EM を拡張した GEM アルゴリズムやその他の EM の拡張型が有効になることがある。

GEM アルゴリズムとは、M-step を、次式をみたす $\theta^{(k+1)}$ を見つける作業に置き換えたアルゴリズムである：

$$Q(\theta^{(k+1)} | \theta^{(k)}) \geq Q(\theta^{(k)} | \theta^{(k)}) \quad (1.54)$$

これは、M-step において、必ずしも Q 関数の最大化を達成する必要がなく、より大きな値に更新する $\theta^{(k+1)}$ を見つければ良いことを示している。したがって、M-step で代数的な解が得られず Newton-Raphson 法など利用して数値的

CONTENTS

に最大化を図らなければならないような場合でも、1回の反復だけで止めることも可能となる。Lange(1995)では、a gradient algorithmとしてこの方法を提唱している。

1.7.2 ECM アルゴリズム

ECM アルゴリズムは、E-step と CM(Conditional Maximization)-step からなり、名前の通り、従来の M-step を条件付の最大化ステップ (CM-step) に置き換えたものである (Meng and Rubin 1993)。ここで、条件付の最大化ステップでは、最大化する際に動かすパラメータの次元は、従来の M-step に比べ小さくなる。このため、モデルに含まれるパラメータの次元が大きな場合、すべてのパラメータに対して最大化を同時に行う M-step に場合に比べ、CM-step の方が、単純で速く、かつ安定している場合が多い。

具体的な ECM アルゴリズムでは、従来の M-step を、計算が簡素化された複数の (S 個の) CM-step で置き換える。 $(k+1)$ 回目の反復における s 番目の CM-step では、 $c_s(\boldsymbol{\theta}) = c_s(\boldsymbol{\theta}^{(k+(s-1)/S)})$ の制約の下で $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$ を最大化する $\boldsymbol{\theta}^{(k+(s-1)/S)}$ を求める。ここで、 $c_s(\boldsymbol{\theta}), s = 1, \dots, S$ はあらかじめ設定しておいた既知の関数である。

CM-step を進めるに従い、

$$\begin{aligned} Q(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}) &\leq Q(\boldsymbol{\theta}^{(k+1/S)}; \boldsymbol{\theta}^{(k)}) \\ &\leq Q(\boldsymbol{\theta}^{(k+2/S)}; \boldsymbol{\theta}^{(k)}) \\ &\dots \\ &\leq Q(\boldsymbol{\theta}^{(k+(S-1)/S)}; \boldsymbol{\theta}^{(k)}) \\ &\leq Q(\boldsymbol{\theta}^{(k+1)}; \boldsymbol{\theta}^{(k)}) \end{aligned}$$

となるようなパラメータの更新が行われ、結果として一回の CM-step で必ず完全データの尤度の期待値が上昇 (厳密には非減少) する。すなわち、ECM アルゴリズムは GEM アルゴリズムになっている。

また、ECM アルゴリズムにおいて、各 CM-step または一部の CM-step の

CONTENTS

合間に E-step を挿入する場合がある。このアルゴリズムを Multicycle ECM アルゴリズムと呼んでいる。Multicycle ECM アルゴリズムでは、E-step を挿入するため、一回の反復に要する計算量は増大するが、一回の反復でより大きな尤度の改善が期待できる。Multicycle ECM アルゴリズムは、GEM アルゴリズムではないが、各反復における観測データ (不完全データ) の尤度の単調非減少性は成立する (Meng and Rubin(1993))。

1.7.3 ECME アルゴリズム – Contaminated Normal –

ECME アルゴリズムは ECM アルゴリズムの拡張として、Liu and Rubin(1994, 1995) により提唱された。ECME は Expectation-Conditional Maximization Either のことで、最後の "Either" は、CM-step において、場合によっては、不完全データに基づく尤度の最大化を行うことを意味している。

確率変数 \mathbf{y} が p 次の混淆正規分布 (contaminated normal)

$$(1 - \delta) \times N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \delta \times N(\boldsymbol{\mu}, \boldsymbol{\Sigma}/\lambda) \quad (1.55)$$

に従うとする。このとき、 \mathbf{y} とは独立な正の、1 か $\lambda (> 0)$ の値をとる 2 値型の確率変数 q を想定し、 q の確率関数 $M(q)$ を

$$M(q) = \begin{cases} 1 - \delta & \text{if } q = 1, \\ \delta & \text{if } q = \lambda \quad (0 < \delta < 1, \lambda > 0), \\ 0 & \text{その他,} \end{cases} \quad (1.56)$$

とすると、 q が与えられた下での \mathbf{y} の条件付分布は、

$$\mathbf{y} | q \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}/q) \quad (1.57)$$

となる。

ここで、各個体に対して、 (\mathbf{y}_i, q_i) , $i = 1, \dots, n$ を観測されるべき完全データと考え (実際は、 q_i はすべて欠測データとなる)、ECME アルゴリズムの枠組みで、(1.55) 式の混淆正規分布に含まれるパラメータ $\lambda, \delta, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ の最尤推定値を導く方法を与える (Yamaguchi(1990))。

CONTENTS

E-step:

$$\begin{aligned} w_i^{(k+1)} &= E(q_i | \mathbf{y}_i; \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}, \delta^{(k)}, \lambda^{(k)}) \\ &= \frac{1 - \delta^{(k)} + \delta^{(k)} \lambda^{(k)1+p/2} \exp\{(1 - \lambda^{(k)})d_i^{(k)2}/2\}}{1 - \delta^{(k)} + \delta^{(k)} \lambda^{(k)p/2} \exp\{(1 - \lambda^{(k)})d_i^{(k)2}/2\}} \end{aligned}$$

$$v_i^{(k+1)} = Pr(q_i = \lambda^{(k)} | \mathbf{y}_i; \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}, \delta^{(k)}) \quad (1.58)$$

$$= \frac{\delta^{(k)} \lambda^{(k)p/2} \exp\{(1 - \lambda^{(k)})d_i^{(k)2}/2\}}{1 - \delta^{(k)} + \delta^{(k)} \lambda^{(k)p/2} \exp\{(1 - \lambda^{(k)})d_i^{(k)2}/2\}} \quad (1.59)$$

ただし, $d_i^{(k)2} = (\mathbf{y}_i - \boldsymbol{\mu}^{(k)})' \boldsymbol{\Sigma}^{(k)-1} (\mathbf{y}_i - \boldsymbol{\mu}^{(k)})$.

CM1-step:

$$\boldsymbol{\mu}^{(k+1)} = \frac{\sum_{i=1}^n w_i^{(k+1)} \mathbf{y}_i}{\sum_{i=1}^n w_i^{(k+1)}}, \quad (1.60)$$

$$\boldsymbol{\Sigma}^{(k+1)} = \frac{\sum_{i=1}^n w_i^{(k+1)} (\mathbf{y}_i - \boldsymbol{\mu}^{(k+1)}) (\mathbf{y}_i - \boldsymbol{\mu}^{(k+1)})'}{n}, \quad (1.61)$$

$$\delta^{(k+1)} = \frac{\sum_{i=1}^n v_i^{(k+1)}}{n} \quad (1.62)$$

CM2-step: λ に関して, Q 関数ではなく, 不完全データ \mathbf{y} に基づく対数尤度 LL に関する最大化を行い, 値を更新する:

$$\begin{aligned} LL = & - \frac{n}{2} \log |\boldsymbol{\Sigma}^{(k+1)}| - \frac{1}{2} \sum_{i=1}^n d_i^{(k+1)2} \\ & + \sum_{i=1}^n \log [1 - \delta^{(k+1)} + \delta^{(k+1)} \lambda^{p/2} \exp\{(1 - \lambda)d_i^{(k+1)2}/2\}] \end{aligned} \quad (1.63)$$

具体的には, 以下の解を求めることになる.

$$\lambda = \frac{p \sum Pr(q_i = \lambda | \mathbf{y}_i; \boldsymbol{\mu}^{(k+1)}, \boldsymbol{\Sigma}^{(k+1)}, \delta^{(k+1)})}{\sum d_i^{(k+1)2} Pr(q_i = \lambda | \mathbf{y}_i; \boldsymbol{\mu}^{(k+1)}, \boldsymbol{\Sigma}^{(k+1)}, \delta^{(k+1)})} \quad (1.64)$$

ここで, $Pr(q_i = \lambda | \mathbf{y}_i; \boldsymbol{\mu}^{(k+1)}, \boldsymbol{\Sigma}^{(k+1)}, \delta^{(k+1)})$, また d_i^2 の値は, E-step の結果を使用する.

CONTENTS

ECME アルゴリズムと ECM アルゴリズムの双方が適用できる場合は、ECME アルゴリズムの方が、観測データ (不完全データ) に対する対数尤度 LL を直接的に最大化するので、必ずしもではないが、多くの場合、収束のスピードが速くなる。これらのアルゴリズムのモデルに応じた比較研究については、Liu and Rubin (1994) が参考になる。

E-step が閉じた式展開で単純に解けない場合の取り扱いに関しては、近似分布で置き換える (Laird, 1978) やラプラス展開を用いる (Steele, 1996) などの工夫がある。Tanner and Wong (1987) は EM のフレームワークの拡張的解釈に基づいて、欠測データがある場合の θ の事後密度の計算にモンテカルロ法や data augmentation 法を利用している。Wei and Tanner (1990) は、E-step でシュミレーションを行うモンテカルロ EM を提唱している。

1.7.4 加速化を意識した完全データの探索

－ optimal EM アルゴリズム－

EM アルゴリズムでは、パラメータの推定が容易になる程度までデータを拡張したものを完全データとするが、Meng and van Dyke (1997) は、加速化も意識した最適なデータ拡張 (Data Augmentation) という考え方を示した。具体的には、EM アルゴリズムの収束スピードは完全データの情報量に依存するので、完全データを変化させるワーキングパラメータ α を導入してその情報量を α の関数で表現し、それを最小化をする α によって最適な完全データを規定する。

加速化については、種々の研究がなされているが (Jamshidian and Jennrich, 1997), その多くは EM アルゴリズムが持つ単純性と安定性のある程度犠牲にするものである。しかし、Optimal EM アルゴリズムでは、本来の単純性と安定性は損なわれていない。

ここでは、自由度 ν を既知とした多変量 t 分布モデルを例に、Optimal EM アルゴリズムを示す。多変量 t 分布は前節の混淆正規分布同様、正規分布の尺度混合分布の一つで、EM アルゴリズムを適用することで比較的簡単に最

CONTENTS

尤推定値を求めることができる。

いま p 次の確率ベクトル \mathbf{y} と \mathbf{y} と独立な正の確率変数 q があり, q が与えられた下での \mathbf{y} の条件付分布は, $\mathbf{y} | q \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}/q)$ であつ, $q \sim \chi_\nu^2/\nu$ とすると, \mathbf{y} の周辺分布は自由度 ν の多変量 t 分布となる. 各個体に対して, (\mathbf{y}_i, q_i) , $i = 1, \dots, n$ を完全データ, q_i はすべて観測されない欠測データとすると, E-step と M-step は次のようになる;

E-step:

$$w_i^{(k)} = \frac{\nu + p}{\nu + d_i^{2(k)}}, \quad d_i^{2(k)} = (\mathbf{y}_i - \boldsymbol{\mu}^{(k)})^T \boldsymbol{\Sigma}^{(k)-1} (\mathbf{y}_i - \boldsymbol{\mu}^{(k)}),$$

M-step:

$$\boldsymbol{\mu}^{(k+1)} = \frac{\sum_{i=1}^n w_i^{(k)} \mathbf{y}_i}{\sum_{i=1}^n w_i^{(k)}}, \quad (1.65)$$

$$\boldsymbol{\Sigma}^{(k+1)} = \frac{\sum_{i=1}^n w_i^{(k)} (\mathbf{y}_i - \boldsymbol{\mu}^{(k)}) (\mathbf{y}_i - \boldsymbol{\mu}^{(k)})^T}{n}. \quad (1.66)$$

optimal-EM アルゴリズムでは, モデル $\mathbf{y}|q \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}/q)$ を $\mathbf{y} = \boldsymbol{\mu} + \mathbf{z}/q^{1/2}$, ただし $\mathbf{z} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ と表現し直し, 更に α (working parameter) を導入して

$$\begin{aligned} \mathbf{y} &= \boldsymbol{\mu} + |\boldsymbol{\Sigma}|^{-\alpha/2} \mathbf{z} / (|\boldsymbol{\Sigma}|^{-\alpha} q)^{1/2} \\ &= \boldsymbol{\mu} + |\boldsymbol{\Sigma}|^{-\alpha/2} \mathbf{z} / q(\alpha)^{1/2}, \end{aligned}$$

$q(\alpha) = |\boldsymbol{\Sigma}|^{-\alpha} q$ と一般化する. $\alpha = 0$ のとき, $q(0) = q$ でもとのモデルとなるが, ここで, α の関数として $I_{obs} I_{aug}^{-1}(\alpha)$ の最小固有値 $\gamma = \gamma(\alpha)$ を評価すると, $\alpha = 1/(\nu + p) \equiv \alpha_{opt}$ のとき, γ が最大となることがわかる. つまり, $q(\alpha_{opt})$ でデータを完全化 (拡大) すると, 収束のスピードが最適化されることになる.

$q(\alpha_{opt})$ を使って EM アルゴリズムを再構築すると, M-step の (1.66) 式が次のようになる;

$$\boldsymbol{\Sigma}^{(k+1)} = \frac{\sum_{i=1}^n w_i^{(k)} (\mathbf{y}_i - \boldsymbol{\mu}^{(k)}) (\mathbf{y}_i - \boldsymbol{\mu}^{(k)})^T}{\sum_{i=1}^n w_i^{(k)}}. \quad (1.67)$$

CONTENTS

これは、(1.66) 式の分母の n を重みの和でおきかえただけの修正である。

Meng and van Dyke(1977)の論文は、EMアルゴリズムの提唱からちょうど20年が経過したことを記念する総括的な論文でもある。他にも1990年代以降、拡張的なEMを系統化する論文や成書が刊行されている。Rubin(1991)は、シミュレーションをベースにした4つの代表的なアルゴリズム: Multiple Imputation, Data Augmentation Algorithm, Gibbs Sampler, Sampling/Importance Resampling Algorithm を乱数機構をもつ拡張型EMのフレームワークの下に解説している。McLachlan and Krishnan(1997)は、拡張型のEMまでひろくかつ具体的に解説したEMアルゴリズムの成書である。雑誌 *Statistica Sinica* では1995年に、続く1997年には、*Statistical Methods in Medical Research* で、EMアルゴリズムの特集が組まれている。Meng(1997)やAmari(1996)もそれぞれ、医学系や情報系の雑誌で組まれたEMアルゴリズム招待論文である。Tanner(1996), Schafer(1997), Sorensen and Gianola(2002), Watanabe and Yamaguchi(2004)も、EMアルゴリズムおよびMCMCを含めその他のアルゴリズムが解説され参考になる。

Bibliography

- [1] Amari, S.(1995). Information geometry of the EM and em algorithms for neural networks, *Neural Networks* **8**, 1379-1408.
- [2] Boyles, R.A.(1983). On the convergence of the EM algorithm, *Journal of The Royal Statistical Society* **B 45**, 47-50.
- [3] Dempster, A.P., Laird, N.M. and Rubin, D.B.(1977). Maximum likelihood from incomplete data via the EM algorithm(with Discussion), *Journal of The Royal Statistical Society* **B 39**, 1-38.
- [4] Flury, B. and Zoppe, A.(2000). Excercises in EM, *The American Statisticians*, **54**, 207-209.
- [5] Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721-741.
- [6] Jamshidian, M. and Jennrich (1997). Acceleration of the EM algorithm by using Quasi-Newton methods, *Journal of the Royal Statistical Society* **B 59**, 569-587.
- [7] Laird, N.W. (1978). Emprical Bayes methods for two-way tables, *Biometrika* **69**, 581-590.

BIBLIOGRAPHY

- [8] Lange, K. (1995). A gradient algorithm locally equivalent to the EM algorithm, *Journal of The Royal Statistical Society B* **57**, 425-437.
- [9] Liu, C. and Rubin, D.B. (1994). The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence, *Biometrika* **81**, 633-648.
- [10] Liu, C. and Rubin, D.B. (1995). ML estimation of the t distribution using EM and its extensions, ECM and ECME, *Statistica Sinica* **5**, 19-39.
- [11] Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm, *Journal of the Royal Statistical Society B* **44**, 226-233.
- [12] McLachlan, G.J. and Krishnan T. (1997). *The EM algorithm and extensions* Wiley-Interscience.
- [13] Meng, X.L. (1997). The EM algorithm and medical studies: a historical link. *Statistical Methods in Medical Research* **6**, 3-23.
- [14] Meng, X.L. and van Dyke, D.B. (1997). The EM algorithm-an old folk-song sung to a fast new tune, *Journal of the Royal Statistical Society B* **59**, 511-567.
- [15] Meng, X.L. and Rubin, D.B. (1991). Using EM to obtain asymptotic variance-covariance matrices - the SEM algorithm, *Journal of the American Statistical Association* **86**,899-909.
- [16] Meng, X.L. and Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework, *Biometrika* **80**, 267-278.
- [17] Oakes, D.(1999). Direct calculation of the information matrix via the EM algorithm, *J. R. Statist. Soc. B*, **61**, Part2, 479-482.

BIBLIOGRAPHY

- [18] Orchard, T. and Woodbury, M.A. (1972). A missing information principle: theory and applications. *Proc.of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, **Vol.1**, 697-715.
- [19] Rao.C.R. (1965). *Linear Statistical Inference and its Applications*,New York: Wiley.
- [20] Rubin, D.B. (1976). Inference with missing data. *Biometrika*,**63** 581-592.
- [21] Rubin, D.B. (1987a). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley
- [22] Rubin, D.B. (1987b). The SIR algorithm. *Journal of the American Statistical Association*, **82** 543-546.
- [23] Rubin, D.B. (1991). EM and beyond. *Psychometrika* **56**, 241-254.
- [24] Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data* Chapman and Hall/CRC.
- [25] Sorensen, D. and Gianola, D. (2002). *Likelihood, Bayesian, and MCMC Methods in Quantative Genetics* Springer.
- [26] Steel, B.M. (1996). A modified EM algorithm for estimation in generalized mixed models. *Biometrika* **52**, 1295-1310.
- [27] Tanner, M.A. (1996). *Tools for Statistical Inference*, Springer.
- [28] Tanner, M.A. and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation *Journal of the American Statistical Association*, **82**, 528-550.
- [29] Yamaguchi, K.(1990). Generalized EM algorithm for models with contaminated normal error terms, *Statistical Methods and Data Analysis*,(Niki,N. ed),Tokyo: Scientist Inc.

BIBLIOGRAPHY

- [30] Watanabe, M. and Yamaguchi, K.(2004). *The EM Algorithm and Related Statistical Models*,, Marcel Dekker, New York.
- [31] Wei, G.C.and Tanner, M.A. (1990). A Monte carlo implementation of the EM algorithm and the poorman's data augmentation algorithms. *Journal of the American Statistical Association* **85**, 699-704.
- [32] Wu,C.F.J.(1983). On the convergence properties of the EM algorithm. *Annals of Statistics*, **11**, 95-103.