統計モデルのパラメタの推定

前川眞一
東京工業大学 大学院社会理工学研究科

1　統計モデル

N 個の個体に関するデータ $(\boldsymbol{y_i}, \boldsymbol{x_i}), i = 1, 2, \ldots, N$ が与えられた場合、統計学では以下のようなモデルを用いてこのデータの分析を行うことが多い。

$$\boldsymbol{y}_i = f(\boldsymbol{x}_i|\boldsymbol{\beta}) + \varepsilon_i, i = 1, 2, \ldots, N \tag{1}$$

ただし、N は観測値（個体）の数、$\boldsymbol{x_i}$ は $q \times 1$ の独立変数（説明変数）、$\boldsymbol{y_i}$ は $p \times 1$ の従属変数（基準変数）、$\boldsymbol{\varepsilon_i}$ は $p \times 1$ の誤差項（残差）、また、$\boldsymbol{\beta}$ はパラメタベクトルである。この形は、観測されるデータ $\boldsymbol{y}$ から誤差を取り除いた部分は、$\boldsymbol{x}$ の関数 f として記述できることを示しているが、これは、$\boldsymbol{x}$ の値を変化させれば $\boldsymbol{y}$ の値も変化し、その両者の関係が f という関数で記述出来ることを意味している。また、通常は、f には自由に決めることの出来るパラメタ $\boldsymbol{\beta}$ が含まれている場合が多い。このモデルは $\boldsymbol{y}$ から誤差を取り除いた部分、すなわち、システマティックに変化する部分を $\hat{\boldsymbol{y}}$ とし、それを $\boldsymbol{y}$ の $\boldsymbol{x}$ を与えたときの条件付き期待値と考えれば、

$$\begin{aligned} \boldsymbol{y_i} &= \hat{\boldsymbol{y}}_i + \boldsymbol{\varepsilon_i} \quad \text{and} \quad \boldsymbol{\mathcal{E}}(\boldsymbol{y_i} \mid x_i) = \hat{\boldsymbol{y}}_i \quad \text{or} \quad \boldsymbol{\mathcal{E}}(\varepsilon_i) = 0 & (2) \\ \hat{\boldsymbol{y}}_i &= f(\boldsymbol{x_i}|\boldsymbol{\beta}) & (3) \end{aligned}$$

と書くこともできる。また、全データに関してこの関係が成り立つことを、

$$\boldsymbol{Y} = F(\boldsymbol{X}|\boldsymbol{\beta}) + \boldsymbol{\mathcal{E}} \quad \text{or} \quad \boldsymbol{Y} = \widehat{\boldsymbol{Y}} + \boldsymbol{\mathcal{E}} \tag{4}$$

とも書くことにする。ただし、$\boldsymbol{Y}$ と $\widehat{\boldsymbol{Y}}$ は $N \times p$、$\boldsymbol{X}$ は $N \times q$、$\boldsymbol{E}$ は $N \times p$、F は $N \times p$ の行列である。なお、従属変数の数が $p = 1$ の場合を１変数（１変量）モデル、それ以外を多変量モデルと呼ぶことがある。

例えば、この形のモデルの最も単純なものは、分散分析や線形回帰分析のモデルであり、非線形の回帰分析（例えばニューラルネットワークモデル）や因子分析、主成分分析、クラスター分析、多次元尺度法等のモデルもこの形で説明することが出来る。

本稿においては、データが与えられた場合に、上記のモデルのパラメタを推定する方法を説明する。なお、パラメタの推定方法は様々な方式があるが、ここでは一般的に利用が可能な最尤法と最小２乗法に付いての説明を行う。

なお、データ $\boldsymbol{y_i}$ がカテゴリカルな場合には、その期待値 $\hat{\boldsymbol{y}}_i$ は生起確率となり、通常は２項分布や多項分布を仮定した最尤解を用いることになる。

2　データの種類と適合度の基準

2.1　計量的データ

基準変数 $\boldsymbol{Y}$ が計量データの場合、誤差項の分布として正規分布を仮定する場合が多い。

$$\varepsilon_i \sim N(0, \sigma^2) \quad \text{or} \quad \boldsymbol{\varepsilon_i} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}) \tag{5}$$

この場合は $\boldsymbol{Y}$ の $\boldsymbol{X}$ を与えたときの条件付き密度関数 g が

$$g(y_i|\boldsymbol{X},\boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}(y_i - f(\boldsymbol{x_i}|\boldsymbol{\beta}))^2/\sigma^2\right) \quad (6)$$

$$g(\boldsymbol{y_i}|\boldsymbol{X},\boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}|}}\exp\left(-\frac{1}{2}(\boldsymbol{y_i} - f(\boldsymbol{x_i}|\boldsymbol{\beta}))'\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{y_i} - f(\boldsymbol{x_i}|\boldsymbol{\beta}))\right) \quad (7)$$

と書けるため、N 個の独立な観測値を得た場合、全てのデータの同時密度は

$$g(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta}) = \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}(y_i - f(\boldsymbol{x_i}|\boldsymbol{\beta}))^2/\sigma^2\right) \quad (8)$$

$$g(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta}) = \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}|}}\exp\left(-\frac{1}{2}(\boldsymbol{y_i} - f(\boldsymbol{x_i}|\boldsymbol{\beta}))'\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{y_i} - f(\boldsymbol{x_i}|\boldsymbol{\beta}))\right) \quad (9)$$

となる。このデータの同時密度関数をパラメタ $\boldsymbol{\beta}$ の関数と見たものを尤度関数 L、また、その対数を対数尤度関数 lnL と呼ぶ。

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}(y_i - f(\boldsymbol{x_i}|\boldsymbol{\beta}))^2/\sigma^2\right) \quad (10)$$

$$lnL(\boldsymbol{\beta}) = -\frac{1}{2}\left(\ln 2\pi\sigma^2 + \sum_{i=1}^{N}(y_i - f(\boldsymbol{x_i}|\boldsymbol{\beta}))^2/\sigma^2\right) \quad (11)$$

この尤度を最大とするようなパラメタの値をパラメタの最尤解と呼ぶ。また、対数尤度の形から、誤差分散が全て等しい場合には、この解は

$$\mathrm{RSS} = \sum_{i=1}^{N}(y_i - f(\boldsymbol{x_i}|\boldsymbol{\beta}))^2 \quad (12)$$

$$\mathrm{RSS} = \sum_{i=1}^{N}(\boldsymbol{y_i} - f(\boldsymbol{x_i}|\boldsymbol{\beta}))'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y_i} - f(\boldsymbol{x_i}|\boldsymbol{\beta})) \quad (13)$$

を最少とする最小２乗解と同じであると言える。（ただし、 $\boldsymbol{\Sigma}^{-1}$ 行列は既知の重み行列として取り扱う。）なお、多変量の場合で重み行列がない場合は、

$$\mathrm{RSS} = \sum_{j=1}^{p}\sum_{i=1}^{N}(y_{ij} - f_j(\boldsymbol{x_i}|\boldsymbol{\beta}))^2 \quad (14)$$

(15)

と書くことが出来る。

2.1.1 例

計量データ：正規分布

平均値の推定：	$\boldsymbol{X}$ は全て 1
多群の平均値の推定：	$\boldsymbol{X}$ は群を示すダミー変数（デザイン行列）
回帰分析：	$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{\beta}'$
ニューラルネット：	$\widehat{\boldsymbol{Y}} = \boldsymbol{\beta}$ の要素の非線形関数
因子分析：	$\boldsymbol{X}$ は計量的潜在変数（これも推定する）
クラスター分析：	$\boldsymbol{X}$ は計数的潜在変数（これも推定する）
多次元尺度法：	$\boldsymbol{X}$ は刺激対を示すダミー変数（デザイン行列）

2.2 計数的データ（度数データ、カテゴリカルデータ）

データが、ある個体が p 個のカテゴリの中でどのカテゴリに反応したか、というような形で得られる場合、それを計数的データ、度数データ、もしくはカテゴリカルデータと呼ぶ。この場合、$\boldsymbol{y_i}$はその個体がどのカテゴリに反応したかを示すインディケーターベクトルとなる。また、$\boldsymbol{y_i}$ の期待値 $\hat{\boldsymbol{y}}_i$ はその生起確立となる。以下に、そのことを明記するため、生起確立を

$$\hat{\boldsymbol{y}}_i = f(\boldsymbol{x_i}|\boldsymbol{\beta}) = \pi(\boldsymbol{x_i}|\boldsymbol{\beta}) \tag{16}$$

と記す。

基準変数 $\boldsymbol{Y}$ が計数データの場合、その分布としてベルヌーイ分布を仮定する場合が多い。この場合は $\boldsymbol{Y}$ の $\boldsymbol{X}$ を与えたときの条件付き確率 g が

$$g(y_i|\boldsymbol{X},\boldsymbol{\beta}) \quad = \quad \prod_{j=1}^{p} \pi_j(\boldsymbol{x_i}|\boldsymbol{\beta})^{y_{ij}}, \quad \sum_{j=1}^{p} \pi_j(\boldsymbol{x_i}|\boldsymbol{\beta}) = 1 \tag{17}$$

と書けるため、N 個の独立な観測値を得た場合、全てのデータの同時密度や対数尤度関数は

$$g(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta}) \quad = \quad \prod_{i=1}^{N}\prod_{j=1}^{p} \pi_j(\boldsymbol{x_i}|\boldsymbol{\beta})^{y_{ij}} \tag{18}$$

$$lnL(\boldsymbol{\beta}) \quad = \quad \sum_{i=1}^{N}\sum_{j=1}^{p} y_{ij} \ln \pi_j(\boldsymbol{x_i}|\boldsymbol{\beta}) + \boldsymbol{\beta}\text{を含まない部分} \tag{19}$$

となる。

なお、この形とは少し異なるが、データ $\boldsymbol{y}$ が、p 個の２値変数からなる場合、条件付きの独立性を仮定して、

$$g(y_i|\boldsymbol{X},\boldsymbol{\beta}) \quad = \quad \prod_{j=1}^{p} \pi_j(\boldsymbol{x_i}|\boldsymbol{\beta})^{y_{ij}}(1-\pi_j(\boldsymbol{x_i}|\boldsymbol{\beta})^{1-y_{ij}} \tag{20}$$

また、$(n_{ij}, f_{ij}), \quad , i = 1,2,\ldots, j = 1,2,\ldots,p$ を各個体の変数ごとの試行数と成功の度数として、度数の条件付き確率が

$$g(\boldsymbol{f}_i|\boldsymbol{X},\boldsymbol{\beta}) \propto \prod_{j=1}^{p} \pi_j(\boldsymbol{x_i}|\boldsymbol{\beta})^{f_{ij}}(1-\pi_j(\boldsymbol{x_i}|\boldsymbol{\beta})^{n_{ij}-f_{ij}} \tag{21}$$

と書き表されるモデルを考えることもある。この場合、全てのデータの同時密度は

$$g(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta}) \quad = \quad \prod_{i=1}^{N}\prod_{j=1}^{p} \pi_j(\boldsymbol{x_i}|\boldsymbol{\beta})^{y_{ij}}(1-\pi_j(\boldsymbol{x_i}|\boldsymbol{\beta})^{1-y_{ij}} \tag{22}$$

$$g(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{\beta}) \quad = \quad \prod_{i=1}^{N}\prod_{j=1}^{p} \pi_j(\boldsymbol{x_i}|\boldsymbol{\beta})^{f_{ij}}(1-\pi_j(\boldsymbol{x_i}|\boldsymbol{\beta})^{n_{ij}-f_{ij}} \tag{23}$$

となり、パラメタの対数尤度関数 lnL は

$$lnL(\boldsymbol{\beta}) \quad = \quad \sum_{i=1}^{N}\sum_{j=1}^{p} y_{ij} \ln \pi_j(\boldsymbol{x_i}|\boldsymbol{\beta}) + (1-y_{ij}) \ln \left(1-\pi_j(\boldsymbol{x_i}|\boldsymbol{\beta})\right) + \boldsymbol{\beta}\text{を含まない部分} \tag{24}$$

$$lnL(\boldsymbol{\beta}) \quad = \quad \sum_{i=1}^{N}\sum_{j=1}^{p} f_{ij} \ln \pi_j(\boldsymbol{x_i}|\boldsymbol{\beta}) + (n_{ij}-y_{ij}) \ln \left(1-\pi_j(\boldsymbol{x_i}|\boldsymbol{\beta})\right) + \boldsymbol{\beta}\text{を含まない部分} \tag{25}$$

となる。

2.2.1 例

計数データ：２項分布
　平均（成功確率）：　$\boldsymbol{x_i}$ は全て 1
　多群の平均値の推定：　$\boldsymbol{x_i}$ は群を示すダミー変数（デザイン行列）
　ロジスティク回帰：　成功確率の対数がパラメタに関して線形
　項目反応理論：　$\boldsymbol{X}$ は潜在変数（これも推定する）

3 ベイズ推定法

パラメタに関する事前の知識がその事前分布 $h(\boldsymbol{\beta})$ という形で知られている場合には、パラメタの事後分布

$$h(\boldsymbol{\beta}) \propto g(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}) \times h(\boldsymbol{\beta}) \tag{26}$$

となるため、その事後モード（同時モード、周辺モード）や事後平均を求めることが出来る。

4 最小２乗法

RSS は非負の関数であるから、その最小値が存在し、それを与えるパラメタの値は

$$\frac{\partial\,\mathrm{RSS}}{\partial\,\boldsymbol{\beta}} = \mathbf{0} \tag{27}$$

の根となっているはずである。

4.1 線形の場合の例

- 回帰分析
- 冗長性分析
- 主成分分析
- その他

スカラー値を返す行列の関数の微分を用いると便利である。

5 交互最小２乗法

複数のパラメタ $\boldsymbol{\xi} = \xi_1, \xi_2, \ldots, \xi_r$ を持つ最小２乗基準をすべてのパラメタに関して同時に最小とするためには

$$\frac{\partial\,\mathrm{RSS}}{\partial\,\boldsymbol{\xi}} = \mathbf{0}$$

という同時方程式を解かなければならないが，これは時として困難である．そのような場合には，パラメタを適当に分割し，その分割ごとに RSS の最小化を繰り返していけば，最終的には RSS の極小値が得られることが知られている．すなわち，パラメタの分割を

$$\boldsymbol{\xi} = \{\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_S\},$$

その初期値を $\boldsymbol{\xi}^0$ とすれば，

> $s = 1, 2, \ldots, S$ に関して， $\boldsymbol{\xi}_k^0, k \neq s$ を既知とした時に RSS を最小とするような $\boldsymbol{\xi}_s$ を求め，
> その値を新たに $\boldsymbol{\xi}_s^0$ とする．

という過程を繰り返していけば，最終的に得られる $\boldsymbol{\xi}^0$ において RSS の極小値が求められる．このような方法は交互最小２乗法 (Alternating Least Squares method: ALS) と呼ばれている．ここではその概要を説明する．

RSS は $\boldsymbol{\xi}$ の関数であるからそれを $\mathrm{RSS}(\boldsymbol{\xi})$ と書く．また，RSS を $\boldsymbol{\xi}_s$ のみの関数と見たものを $\mathrm{RSS}(\boldsymbol{\xi}_s)$ ，$\boldsymbol{\xi}_s$ 以外のパラメタを $\boldsymbol{\xi}_{(s)}$ と書けば， ALS の過程は，$\boldsymbol{\xi}_{(1)}$ を既知として $\mathrm{RSS}(\boldsymbol{\xi}_1)$ を最小にするような $\boldsymbol{\xi}_1$ をもとめる，求められた $\boldsymbol{\xi}_1$ の値で現在の $\boldsymbol{\xi}$ を更新する，$\boldsymbol{\xi}_{(2)}$ を既知として $\mathrm{RSS}(\boldsymbol{\xi}_2)$ を最小にするような $\boldsymbol{\xi}_2$ をもとめる，求められた $\boldsymbol{\xi}_2$ の値で現在の $\boldsymbol{\xi}$ を更新する， $\cdots$ という過程を繰り返していくことになる．$\mathrm{RSS}(\boldsymbol{\xi}_s)$ は，RSS と基本的には同じものであるから，各段階は RSS の値そのものを少しずつ改善している．したがって，各 $\boldsymbol{\xi}_s$ が更新された段階で $\mathrm{RSS}(\boldsymbol{\xi})$ を評価すれば，少なくともその値は一つ前の段階で得られた RSS の値よりも大きくはないから，最終的に，RSS の極小値が得られることになる．すなわち， ALS が収束した段階においては，$\frac{\partial\, \mathrm{RSS}}{\partial\, \boldsymbol{\xi}} = \mathbf{0}$ となっている．なお，収束の判定は， ALS においては各段階における RSS の値が決して大きくならないため，たとえば，RSS の減少分がゼロと見なせるか否かで行えばよい．したがって， ALS のアルゴリズムはおおよそ次のようになる．

```
Store the initial value in ξ;
rssp = RSS(ξ);
convergence = 0;
do iteration=1 to maxiter (until covnergence = 1);
   do s=1 to S; * for each subset of parameters;
      Get ξ_s from ξ;
      Find the value, ξ̂_s of ξ_s which mninizes RSS(ξ_s),
         or, at least, satisfies that RSS(ξ̂_s) ≤ RSS(ξ);
      Update the corresponding elements of ξ by ξ̂_s;
   end;
   rss = RSS(ξ);
   if rssp - rss ≤ ε then convergence = 1; else convergence = 0;
   rssp = rss;
  end;
```

なお，計算機の容量の増大や速度の改善により，多くの問題を非線形の最小２乗法 (section ?? を参照) として直接的にくことが可能となってきたが，そのような場合でも，条件付き最小２乗解を求めるのが簡単な場合には（多くの場合，条件付き最小２乗解は線形の問題となる），交互最小２乗法を用いると効果的な場合が多い．また，変数の最適変換を含む最小２乗法のように交互最小２乗法を用いざるを得ない問題も多い．一般に，非線形の最小２乗法による定式化よりも交互最小２乗法を用いた方が収束までの繰り返し計算の回数は増えるが，計算時間としてはさほど変わらない場合が多い．また、Ramsay(1975) の方法等を用いれば交互最小２乗法を加速することができる．

以下に，$\boldsymbol{v} = \boldsymbol{U}\boldsymbol{\xi} + \boldsymbol{e}$ というモデルのパラメタを ALS を用いて推定するプログラムの例を示す．ここでは， Q 個のパラメタのひとつひとつをサブセットとしている．この場合，モデルが，

$$\boldsymbol{v} = \boldsymbol{U}_{(s)}\boldsymbol{\xi}_{(s)} + \boldsymbol{u}_{(s)}\xi_s + \boldsymbol{\varepsilon}$$

ただし，$\boldsymbol{u}_{(s)}$ は $\boldsymbol{U}$ 行列の第 s 列，$\boldsymbol{U}_{(s)}$ は $\boldsymbol{U}$ 行列の第 s 列を除いた $N \times (Q-1)$ の行列，$\boldsymbol{\xi}_{(s)}$ は $\boldsymbol{\xi}$ ベクトルの第 s 要素を除いた $(Q-1) \times 1$ のベクトル，と書けることから，第 s サブセットのパラメタ ξ_s の条件付きの最小２乗推定値は，$(\boldsymbol{v} - \boldsymbol{U}_{(s)}\boldsymbol{\xi}_{(s)})$ を $\boldsymbol{u}_{(s)}$ に回帰することにより，

$$\widehat{\xi_s} = \left(\boldsymbol{u}_{(s)}{}'\boldsymbol{u}_{(s)}\right)^{-1}\boldsymbol{u}_{(s)}{}'(\boldsymbol{v} - \boldsymbol{U}_{(s)}\boldsymbol{\xi}_{(s)})$$

となる．

6　非線形モデルの最小２乗解

モデル $\boldsymbol{\theta}(\boldsymbol{\xi})$ がパラメタ $\boldsymbol{\xi}$ (Q $\times$ 1) に関して非線形の場合，最小２乗基準

$$\mathrm{RSS}(\boldsymbol{\xi}) = \|\boldsymbol{v} - \boldsymbol{\theta}(\boldsymbol{\xi})\|^2$$

を最小にするようなパラメタの推定値を解析的に求めることは困難である場合が多い．しかし，そのような場合にも，何らかの数値的最適化の方法を用いれば，最小２乗解を求めることが可能である．

一般に，ある関数 $\mathrm{r}(\boldsymbol{\xi})$ の極値（極小値または極大値のことであるが，本節では極小値とする）を与えるパラメタ $\boldsymbol{\xi}$ の値を数値的に求めることを関数の最適化と呼ぶが，その方法は，何らかの方法で求めた $\boldsymbol{\xi}$ の初期値から，$\mathrm{r}(\boldsymbol{\xi})$ が減少する方向 $\boldsymbol{d}$ へある分量 α だけ修正する，

$$\boldsymbol{\xi}_{(\ell+1)} = \boldsymbol{\xi}_{(\boldsymbol{\ell})} + \alpha \boldsymbol{d}_{(\boldsymbol{\ell})} \tag{28}$$

という形の繰り返し計算でパラメタの値を更新していくという形をとる．数値的最適化の方法には様々なものがあるが，それらの主な違いは，更新の方向 $\boldsymbol{d}_{(\boldsymbol{\ell})}$ をどの様にして定めるかによる．

最も単純なものは最急降下法と呼ばれる方法であり，修正の方向を $\mathrm{r}(\boldsymbol{\xi})$ のパラメタの各要素に関する一次偏微分係数行列の逆方向

$$\boldsymbol{d}_{(\boldsymbol{\ell})} = -\boldsymbol{g}(\boldsymbol{\xi}_{(\boldsymbol{\ell})}) = -\left.\frac{\partial\, \mathrm{r}(\boldsymbol{\xi})}{\partial\, \boldsymbol{\xi}}\right|_{\boldsymbol{\xi}=\boldsymbol{\xi}_{(\boldsymbol{\ell})}}$$

に取るものである．この， $Q \times 1$ のベクトル $\boldsymbol{g}$ はグレイディエントと呼ばれることがあるが，これは $\boldsymbol{\xi}_{(\boldsymbol{\ell})}$ における $\mathrm{r}(\boldsymbol{\xi})$ のパラメタの各要素に関する傾きの逆方向を示すもので，Eq.(28) を用いることにより，

$$\mathrm{r}(\boldsymbol{\xi}_{(\ell+1)}) \leq \mathrm{r}(\boldsymbol{\xi}_{(\boldsymbol{\ell})}) \tag{29}$$

を満たす α が存在することが知られている．この方法は単純であるが，関数の極値に到達するまでの繰り返しの数が多くかかることが欠点である．

一次微分のみならず，２次微分の情報をも用いて最適化を行う方法が，通常 Newton-Raphson 法と呼ばれる方法である．$\mathrm{r}(\boldsymbol{\xi})$ を $\boldsymbol{\xi}_{(\boldsymbol{\ell})}$ のまわりで２次の項まで Taylor 展開すると

$$\mathrm{r}(\boldsymbol{\xi}) \approx \mathrm{r}(\boldsymbol{\xi}_{(\boldsymbol{\ell})}) + \boldsymbol{g}(\boldsymbol{\xi}_{(\boldsymbol{\ell})})'(\boldsymbol{\xi} - \boldsymbol{\xi}_{(\boldsymbol{\ell})}) + \tfrac{1}{2}(\boldsymbol{\xi} - \boldsymbol{\xi}_{(\boldsymbol{\ell})})'\boldsymbol{H}(\boldsymbol{\xi}_{(\boldsymbol{\ell})})(\boldsymbol{\xi} - \boldsymbol{\xi}_{(\boldsymbol{\ell})})$$

となるが，この $Q \times Q$ の２次微分行列（ r をパラメタの各要素で２回偏微分したものを要素として持つ行列）

$$\boldsymbol{H}(\boldsymbol{\xi}_{(\boldsymbol{\ell})}) = \left[\left.\frac{\partial^2 r(\boldsymbol{\xi})}{\partial\boldsymbol{\xi}\ \partial\boldsymbol{\xi}'}\right|_{\boldsymbol{\xi}=\boldsymbol{\xi}_{(\boldsymbol{\ell})}}\right]$$

はヘシアン行列と呼ばれるものである．したがって，極値の条件は

$$\frac{\partial\, \mathrm{r}(\boldsymbol{\xi})}{\partial\, \boldsymbol{\xi}} \approx \boldsymbol{g}(\boldsymbol{\xi}_{(\boldsymbol{\ell})}) + \boldsymbol{H}(\boldsymbol{\xi}_{(\boldsymbol{\ell})})(\boldsymbol{\xi} - \boldsymbol{\xi}_{(\boldsymbol{\ell})}) = 0$$

となり，これより

$$\boldsymbol{\xi} = \boldsymbol{\xi}_{(\boldsymbol{\ell})} - \boldsymbol{H}(\boldsymbol{\xi}_{(\boldsymbol{\ell})})^{-1}\boldsymbol{g}(\boldsymbol{\xi}_{(\boldsymbol{\ell})})$$

となる．したがって，Eq.(28) において

$$\boldsymbol{d}_{(\boldsymbol{\ell})} = -\boldsymbol{H}(\boldsymbol{\xi}_{(\boldsymbol{\ell})})^{-1}\boldsymbol{g}(\boldsymbol{\xi}_{(\boldsymbol{\ell})})$$

とおいたものが Newton-Raphson 法の更新式として得られる．Newton-Raphson 法は一般的に効率の良い方法であるが，ヘシアン行列の計算が複雑な場合には利用が困難であること，また，ヘシアン行列が必ずしも非負定値行列とはならないために，まれに Eq.(29) を満たすような α を見いだすことが不可能な場合が存在する（負のグレイディエント方向から 90 度以上方向がそれる）等の欠点がある．このような欠点を補うために，1 次微分の情報のみからヘシアン行列の逆行列を近似的に計算する方法（いわゆる準ニュートン法）も考案されている．

以下には，最適化されるべき関数が最小２乗基準の場合（ $\mathrm{r}(\boldsymbol{\xi}) = \mathrm{RSS}(\boldsymbol{\xi})$ ）に適した Gauss-Newton 法に関して説明する．いま，求めるべきパラメタの仮の値を $\boldsymbol{\xi}_{(\ell)}$ とする．モデルの各要素を $\boldsymbol{\xi}_{(\ell)}$ のまわりで 1 次の項まで Taylor 展開したものをまとめれば，

$$\boldsymbol{\theta}(\boldsymbol{\xi}) \approx \boldsymbol{\theta}(\boldsymbol{\xi}_{(\ell)}) + \boldsymbol{F}(\boldsymbol{\xi}_{(\ell)})(\boldsymbol{\xi} - \boldsymbol{\xi}_{(\ell)})$$

となるが，ここに，

$$\boldsymbol{F}(\boldsymbol{\xi}) = \frac{\partial\, \boldsymbol{\theta}(\boldsymbol{\xi})}{\partial\, \boldsymbol{\xi}} = \left[\frac{\partial\, \theta_i(\boldsymbol{\xi})}{\partial\, \boldsymbol{\xi}_k}\right], \quad i = 1, 2, \ \ldots, N, \ k = 1, 2, \ \ldots, Q$$

は，ヤコビアン行列と呼ばれる $N \times\ Q$ の 1 次微分行列，また，$\boldsymbol{F}(\boldsymbol{\xi}_{(\ell)})$ はそれを $\boldsymbol{\xi}_{(\ell)}$ で評価した行列

$$\boldsymbol{F}(\boldsymbol{\xi}_{(\ell)}) = \left.\frac{\partial\, \boldsymbol{\theta}(\boldsymbol{\xi})}{\partial\, \boldsymbol{\xi}}\right|_{\boldsymbol{\xi}=\boldsymbol{\xi}_{(\ell)}}$$

である．この 1 次近似を RSS に代入すれば，

$$\begin{aligned} \mathrm{RSS}(\boldsymbol{\xi}) &\approx \left\|\boldsymbol{v} - \left(\boldsymbol{\theta}(\boldsymbol{\xi}_{(\ell)}) + \boldsymbol{F}(\boldsymbol{\xi}_{(\ell)})(\boldsymbol{\xi} - \boldsymbol{\xi}_{(\ell)})\right)\right\|^2 \\ &= \left\|\boldsymbol{v} - \boldsymbol{\theta}(\boldsymbol{\xi}_{(\ell)}) - \boldsymbol{F}(\boldsymbol{\xi}_{(\ell)})(\boldsymbol{\xi} - \boldsymbol{\xi}_{(\ell)})\right\|^2 \end{aligned}$$

となるから，

$$\frac{\partial\, \mathrm{RSS}(\boldsymbol{\xi})}{\partial\, \boldsymbol{\xi}} \approx -2\boldsymbol{F}(\boldsymbol{\xi}_{(\ell)})'(\boldsymbol{v} - \boldsymbol{\theta}(\boldsymbol{\xi}_{(\ell)})) + 2\boldsymbol{F}(\boldsymbol{\xi}_{(\ell)})'\boldsymbol{F}(\boldsymbol{\xi}_{(\ell)})(\boldsymbol{\xi} - \boldsymbol{\xi}_{(\ell)}) \ = \ 0$$

の解は

$$\boldsymbol{\xi} = \boldsymbol{\xi}_{(\ell)} + \left(\boldsymbol{F}(\boldsymbol{\xi}_{(\ell)})'\boldsymbol{F}(\boldsymbol{\xi}_{(\ell)})\right)^{-1}\boldsymbol{F}(\boldsymbol{\xi}_{(\ell)})'(\boldsymbol{v} - \boldsymbol{\theta}(\boldsymbol{\xi}_{(\ell)}))$$

となる．また，最適化されるべき関数が最小２乗基準の場合には，先に示したグレイディエントベクトルが

$$\boldsymbol{g} = \frac{\partial\, \mathrm{RSS}(\boldsymbol{\xi})}{\partial\, \boldsymbol{\xi}} = -2\boldsymbol{F}(\boldsymbol{\xi})'(\boldsymbol{v} - \boldsymbol{\theta}(\boldsymbol{\xi}))$$

となるため， Gauss-Newton 法では Eq.(28) において

$$\boldsymbol{d}_{(\ell)} = -\tfrac{1}{2}\left(\boldsymbol{F}(\boldsymbol{\xi}_{(\ell)})'\boldsymbol{F}(\boldsymbol{\xi}_{(\ell)})\right)^{-1}\boldsymbol{g}(\boldsymbol{\xi}_{(\ell)})$$

と置いたものになっていることがわかる．Gauss-Newton 法は１次の情報しか利用しないが効率はよく，Eq.(29) を満たす α も必ず存在することが知られている．

また，基準変数が多変量の場合は，

$$\mathrm{RSS} \ = \ \|\boldsymbol{V} - \boldsymbol{\Theta}(\boldsymbol{\xi})\|^2 \ = \ \sum_{j=1}^{p}\left\|\boldsymbol{v}_{(j)} - \boldsymbol{\theta}_{(j)}(\boldsymbol{\xi})\right\|^2$$

と書けるため，変数ごとのヤコビアン行列を

$$\boldsymbol{F}_j(\boldsymbol{\xi}) = \frac{\partial\, \boldsymbol{\theta}_{(j)}(\boldsymbol{\xi})}{\partial\, \boldsymbol{\xi}} = \left[\frac{\partial\, \theta_{ij}(\xi)}{\partial\, \boldsymbol{\xi}_k}\right], \quad i = 1, 2, \ \ldots, N, \ k = 1, 2, \ \ldots, Q$$

と定義すれば，グレイディエント並びに更新の方向は

$$\begin{aligned} \boldsymbol{g} &= \frac{\partial\,\mathrm{RSS}(\boldsymbol{\xi})}{\partial\,\boldsymbol{\xi}} = -2\sum_{j=1}^{p}\boldsymbol{F}_{(j)}(\boldsymbol{\xi})'(\boldsymbol{v}_{(j)}-\boldsymbol{\theta}_{(j)}(\boldsymbol{\xi})) \\ \boldsymbol{d}_{(\boldsymbol{\ell})} &= -\tfrac{1}{2}\Big(\textstyle\sum_{j=1}^{p}\boldsymbol{F}_{(j)}(\boldsymbol{\xi}_{(\boldsymbol{\ell})})'\boldsymbol{F}_{(j)}(\boldsymbol{\xi}_{(\boldsymbol{\ell})})\Big)^{-1}\boldsymbol{g}(\boldsymbol{\xi}_{(\boldsymbol{\ell})}) \end{aligned}$$

で与えられる．

パラメタを更新すべき方向が定まれば，ステップサイズ α の値を適当に決めることにより，Eq.(29) を満たすような $\boldsymbol{\xi}_{(\ell)}$ を求めることが出来る．ステップサイズを決める最も簡単な方法は，まず $\alpha=1$ から始めて Eq.(29) が満たされない場合には，それが満たされるまで α の値を半減していくという方法である．より高度な方法としては，RSS の変化を参考にしてステップサイズの減少幅を決める方法や，$\boldsymbol{d}_{(\boldsymbol{\ell})}$ を既知として RSS を α の関数とみなし，それを３次関数で近似して極値を求める方法がある．

Gauss-Newton 法の一般的な形は以下のようになる．

Gauss-Newton 法

$\boldsymbol{\xi}_1$ に初期値を設定する．
for $\ell = 1$ to maxiter until($\mathrm{RSS}(\boldsymbol{\xi}_{(\boldsymbol{\ell})})$ - $\mathrm{RSS}(\boldsymbol{\xi}_{(\ell+1)})$ が小さい);
  $\boldsymbol{F}(\boldsymbol{\xi}_{(\boldsymbol{\ell})})$, $\boldsymbol{g}(\boldsymbol{\xi}_{(\boldsymbol{\ell})})$, $\boldsymbol{d}_{(\boldsymbol{\ell})}$ を計算する．
  α=1;
  for iters = 1 to maxhalf;
    $\boldsymbol{\xi}_{(\ell+1)} = \boldsymbol{\xi}_{(\boldsymbol{\ell})} + \alpha\,\boldsymbol{d}_{(\boldsymbol{\ell})}$;
    if $\mathrm{RSS}(\boldsymbol{\xi}_{(\ell+1)}) \le \mathrm{RSS}(\boldsymbol{\xi}_{(\boldsymbol{\ell})})$ then break;
    α=α/2;
  end;
end;

6.1 非線形の場合の例

- MDS
- ニューラルネットワーク
- その他

7 周辺尤度を最大にする方法：EM アルゴリズム

8 その他の方法