尤度と最尤推定

植野真臣
電気通信大学
情報理工学研究科
情報数理工学プログラム

今後のスケジュール（予定）

4月８日　授業の概要とガイダンス

4月15日　ベイズの定理

4月22日　ベイズはどのように誕生したか？

5月13日　ベイズはコンピュータ、人工知能の父である！！

5月20日　ビリーフとベイズ

5月27日　6月3日　6月10日　　尤度と最尤推定

6月17日　6月24日　7月1日　ベイズ推定と事前分布、階層ベイズ、因果推論

7月8日　　自宅でオンデマンド授業

7月15日　自宅でオンデマンド授業

7月22日　7月29日　確率的グラフィカルモデル、ベイジアンネットワークと機械学習

8月 5日　　テストと総括

本日の目標

1. 古典的統計学の考え方を理解する
2. 尤度（Likelihood)の概念を理解する
3. 最尤推定法（Maximum Likelihood Estimation)を理解する

1. 確率変数

定義1

頻度論

これから試行する実験の結果、実験結果として取り得る値

主観確率

確率法則に従う不確かな変数すべて。

2. 同時確率分布

定義2

いま，m個の確率変数をもつ確率分布$p(x_1, x_2, \cdots, x_m)$を変数$x_1, x_2, \cdots, x_m$ **の同時確率分布**（joint probability distribution）と呼ぶ.

3. 離散周辺確率分布

定義3

x_i のみに興味がある場合，同時確率分布から x_i の確率分布は，離散型の場合，

$$p(x_i) = \sum_{x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_m} p(x_1, x_2, \cdots, x_m)$$

4. 連続周辺確率分布

定義4

連続型の場合,

$$p(x_i) = \int p(x_1, x_2, \cdots, x_m)\, dx_1, \cdots, dx_{i-1}, dx_{i+1}, \cdots, dx_m$$

で求められ, $p(xi)$ を離散型の場合, **周辺確率分布**（marginal probability distribution）, 連続型の場合, **周辺密度関数**（marginal probability density function）と呼ぶ.

5. 確率分布とパラメータ

定義 5. (パラメータ空間と確率分布)

k 次元パラメータ集合を$\boldsymbol{\Theta} = \{\theta_1, \theta_2, \cdots, \theta_k\}$ と書くとき，確率分布は以下のような関数で示される.

$$f(x|\boldsymbol{\Theta})$$

すなわち，確率分布$f(x|\boldsymbol{\Theta})$ の形状はパラメータ$\boldsymbol{\Theta}$ のみによって決定され，パラメータΘ のみが確率分布$f(x|\boldsymbol{\Theta})$ を決定する情報である.

注意

1. 頻度論の統計学では　パラメータは確率変数でない
2. ベイズ統計学では　パラメータも確率変数

6. 確率分布とパラメータ

例 コインをn 回投げたとき，表が出る回数を確率変数x とした確率分布は以下の二項分布に従う.

$$f(x \mid \theta, n) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

ここで，θ は，コインの表が出る確率のパラメータを示す.

7. 尤度原理(フィッシャー)

定義6 (尤度) $X=(X_1,\cdots,X_i,\cdots,X_n)$ が確率分布$f(X_i|\theta)$ に従うn個の確率変数とする.

n 個の確率変数に対応したデータ $\boldsymbol{x}=(x_1,\cdots,x_n)$ が得られたとき,

$$L(\theta|x)=\prod_{i=1}^{n} f(x_i|\theta)$$

を尤度関数(likelihood function)と定義する(Fisher, 1925).

尤度の例

例　コインをn 回投げたとき，表が出た回数がx 回であったときのコインの表が出るパラメータθ の尤度は

$$L(\theta|n,x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

もしくは，

$$L(\theta|n,x) \propto \theta^x(1-\theta)^{n-x}$$

でもよい.

尤度は，データパターンが観測される確率に比例するパラメータθ の関数である.

尤度は確率の定義を満たす保証がないために確率とは呼べないが，これを厳密に確率分布として扱うアプローチが後述するベイズアプローチである.

8. 最尤推定法

尤度を最大にするパラメータθを求めることは，
データを生じさせる確率を最大にするパラメータθ
を求めることになり，その方法を**最尤推定法**
（maxmimum likelihood estimation，**MLE**）と呼ぶ.

9. 最尤推定値

定義7 (最尤推定量)

データx を所与として，以下の尤度最大となるパラメータを求めるとき，

$$L(\theta|x) = \max\{L(\theta|x): \boldsymbol{\theta} \in C\}$$

$\hat{\theta}$を最尤推定量（maximum likelihood estimator）と呼ぶ（Fisher 1925）.

ただし，C はコンパクト集合を示す.

10. 対数尤度とスコア関数

$$l = \ln L(\theta|x)$$

実際には　対数尤度を最大化する

以下のθについてlを偏微分したスコア関数=0となるθを求める.

$$\frac{\partial}{\partial\theta} l = \frac{\partial}{\partial\theta} \ln L(\theta|x) = \frac{1}{L(\theta|x)} \frac{\partial L(\theta|x)}{\partial\theta}$$

例題1　スコア関数の期待値

$$\mathrm{E}\left(\frac{\partial}{\partial\theta}l\right)$$

を求めよ。

例題1　　回答

$$\mathrm{E}\left(\frac{\partial}{\partial\theta}l\right)=\mathrm{E}\left(\frac{1}{L(\theta|x)}\frac{\partial L(\theta|x)}{\partial\theta}\right)$$

$$=\int_{-\infty}^{\infty}\frac{1}{L(\theta|x)}\frac{\partial L(\theta|x)}{\partial\theta}L(\theta|x)\partial x=\frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}L(\theta|x)\partial x$$

ここで　$\int_{-\infty}^{\infty}L(\theta|x)\partial x=\text{const}$

より

$\frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}L(\theta|x)\partial x=0$　　■

例題2　スコア関数の分散を求めよ

$$\mathrm{Var}\left(\frac{\partial}{\partial\theta}l\right)$$

例題2　回答

$$\mathrm{Var}\left(\frac{\partial}{\partial\theta}l\right) = \mathrm{E}\left(\frac{\partial}{\partial\theta}l - \mathrm{E}\left(\frac{\partial}{\partial\theta}l\right)\right)^2 = \mathrm{E}\left(\frac{\partial}{\partial\theta}l - 0\right)^2$$

$$= \mathrm{E}\left(\frac{\partial}{\partial\theta}l\right)^2 = \mathrm{E}\left(\frac{1}{L(\theta|x)}\frac{\partial L(\theta|x)}{\partial\theta}\right)^2$$

これをフィッシャー情報量と呼ぶ.

対数尤度の二回偏微分

$$\frac{\partial^2}{\partial\theta^2} l = \frac{\partial}{\partial\theta}\left(\frac{1}{L(\theta|x)}\frac{\partial L(\theta|x)}{\partial\theta}\right)$$

$$= \frac{\frac{\partial^2}{\partial\theta^2}L(\theta|x)\cdot L(\theta|x) - \left(\frac{\partial}{\partial\theta}L(\theta|x)\right)^2}{L(\theta|x)^2}$$

$$= \frac{\frac{\partial^2}{\partial\theta^2}L(\theta|x)}{L(\theta|x)} - \left(\frac{\frac{\partial}{\partial\theta}L(\theta|x)}{L(\theta|x)}\right)^2$$

$$= \frac{\frac{\partial^2}{\partial\theta^2}L(\theta|x)}{L(\theta|x)} - \left(\frac{\partial}{\partial\theta} l\right)^2$$

フィッシャー情報量の変形

$$\mathrm{E}\left(\frac{\partial}{\partial\theta}l\right)^2 = \mathrm{E}\left[\frac{\frac{\partial^2}{\partial\theta^2}L(\theta|x)}{L(\theta|x)}\right] - \mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ l\right]$$

$$= \int_{-\infty}^{\infty} \frac{\frac{\partial^2}{\partial\theta^2}L(\theta|x)}{L(\theta|x)} L(\theta|x)\partial x - \mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ l\right]$$

$$= \frac{\partial^2}{\partial\theta^2}\int_{-\infty}^{\infty} L(\theta|x)\partial x - \mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ l\right]$$

$\int_{-\infty}^{\infty} L(\theta|x)\partial x = const$ より $\frac{\partial^2}{\partial\theta^2}\int_{-\infty}^{\infty} L(\theta|x)\partial x = 0$

$$\mathrm{E}\left(\frac{\partial}{\partial\theta}l\right)^2 = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ l\right]$$

フィッシャー情報量とは
対数尤度の二回偏微分の負の期待値

$$\mathrm{Var}\left(\frac{\partial}{\partial\theta}l\right) = \mathrm{E}\left(\frac{\partial}{\partial\theta}l\right)^2 = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ l\right]$$

複数のパラメータを持つ場合のフィッシャー情報量行列の計算法

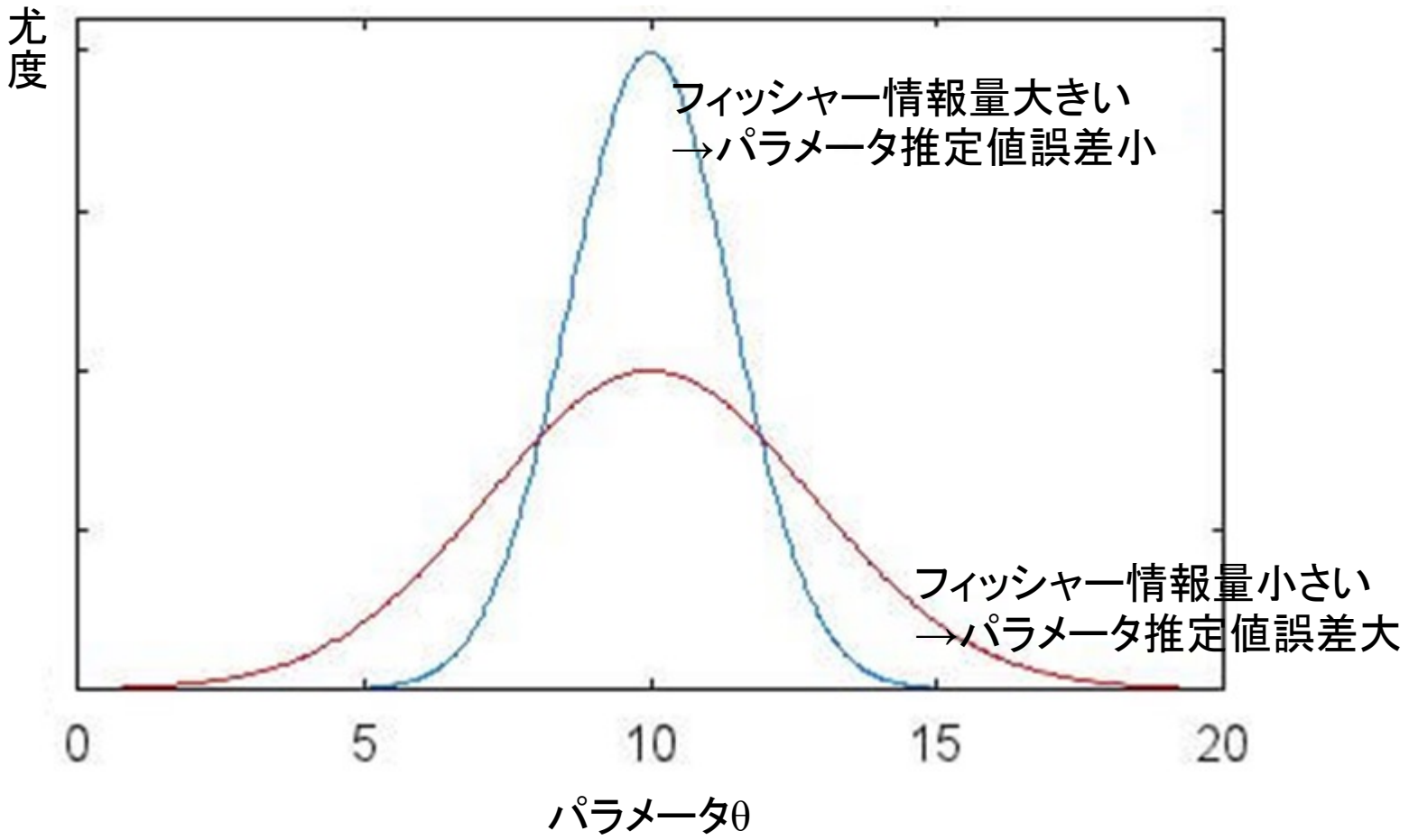
$\boldsymbol{\theta}^T = \{\theta_1, \theta_2, \cdots, \theta_k\}$とする

$$I(\boldsymbol{\theta}|x) = \mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\theta}} l \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} l\right)$$

$I(\boldsymbol{\theta}|x)$の(i,j)成分は

$$[I(\theta|x)]_{(i,j)} = \mathrm{E}\left(\frac{\partial}{\partial \theta_i} l \frac{\partial}{\partial {\theta_j}^{\mathrm{T}}} l\right)$$

フィッシャー情報量は推定値の信頼性



再掲:期待効用関数としてのフィッシャー情報量

事後分布$P(\theta|X)$のθを少しだけ変化させて$\theta + h$にする。

このときの情報量利得の効用関数は

$$u(a,\theta) = (-\log P(\theta|X)) - (-\log P(\theta + h|X))$$

期待効用関数は

$$\lim_{h\to 0} \int_{\theta} P(\theta + h|X) \log \frac{P(\theta + h|X)}{P(\theta|X)}$$

は　フィッシャー情報量に一致する。

証明は　上の期待効用関数を二次の項までテイラー展開すればよい

例題2

(二項分布の最尤推定)

コインを投げてn 回中x 回表が出たときの表の出る確率θ の最尤推定値を求めよ.

例題2　解答

$$L(\theta|n,x)=\binom{n}{x}\theta^{x}(1-\theta)^{n-x}\propto\theta^{x}(1-\theta)^{n-x}$$

$$l=x\log\theta+(n-x)\log(1-\theta)$$

$$\frac{\partial l}{\partial\theta}=\frac{x}{\theta}-\frac{n-x}{1-\theta}=\frac{n-x\theta-(n\theta-x\theta)}{\theta(1-\theta)}$$

$$=\frac{x-n\theta}{\theta(1-\theta)}$$

$\theta\neq 0,1$より

$\frac{\partial l}{\partial\theta}=0$となるのは　$\hat{\theta}=\frac{x}{n}$

例題2　二項分布のフィッシャー情報量

$$\mathrm{E}\left(\frac{\partial}{\partial\theta}l\right)^2 = \mathrm{E}\left(\frac{x-n\theta}{\theta(1-\theta)}\right)^2$$

$$= \left(\frac{\mathrm{E}((x-n\theta)^2)}{\theta^2(1-\theta)^2}\right) = \left(\frac{\mathrm{E}((x-\mathrm{E}(x))^2)}{\theta^2(1-\theta)^2}\right) = \frac{Var(x)}{\theta^2(1-\theta)^2} = \frac{n\theta(1-\theta)}{\theta^2(1-\theta)^2} = \frac{n}{\theta(1-\theta)}$$

例題3 (正規分布)

$$f(x_i|\mu,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{(x_i-\mu)^2}{2\sigma^2}\}$$

について，データ$(x_1,\cdots,x_n)$を得たときの平均値パラメータμ，および分散パラメータσ^2の最尤推定値を求めよ.

例題3　回答

データ$(x_1, x_2, \cdots, x_n)$を得たときの尤度は

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$l = n\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

$\frac{\partial l}{\partial \mu} = 0, \frac{\partial l}{\partial \sigma} = 0$のとき，$l$は最大となるので

$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i-\mu}{\sigma^2} = \frac{\sum_{i=1}^{n} x_i - n\mu}{\sigma^2} = 0 \quad \Rightarrow \quad \hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^4} = 0 \quad \Rightarrow \quad -n + \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^2} = 0$$

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i-\hat{\mu})^2}{n}$$

例題3 正規分布のフィッシャー情報量

$$\mathrm{E}\left(\frac{\partial}{\partial\theta}l\right)^2 = \mathrm{E}\left(\frac{\partial l}{\partial\mu}\right)^2 = \mathrm{E}(\frac{x_i-\mu}{\sigma^2})^2$$

$$= \frac{\mathrm{E}(x_i-\mu)^2}{\sigma^4} = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}$$

例題3 正規分布のフィッシャー情報量 σ^2について

$$\mathrm{Var}\left(\frac{\partial}{\partial\sigma^2}l\right) = \mathrm{E}\left(-\frac{\partial^2}{\partial^2\sigma^2}l\right)$$

$$= \mathrm{E}\left(-\frac{\partial}{\partial\sigma^2}\left(-\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^4}\right)\right)$$

$$= \mathrm{E}\left(-\frac{n}{2\sigma^4} + \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

$$= -\frac{n}{2\sigma^4} + \frac{n}{\sigma^6}\sigma^2$$

$$= \frac{n}{2\sigma^4}$$

$$I = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$$

例題4

母集団の確率分布がポアソン分布

$$p(x) = e^{-\lambda}\frac{\lambda^x}{x!} \quad (\lambda > 0, \quad x = 0, 1, \cdots)$$

についてn 回の観測を行ったところ

データ $\{x_1, x_2, \cdots, x_n\}$

を得た。λを最尤推定せよ。

例題4　回答

対数尤度は$l = log\left[\prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!}\right]$

$$= log\left[e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n}(x_i!)}\right]$$

$= -n\lambda+(\sum_{i=1}^{n} x_i)\log\lambda-\log(\prod_{i=1}^{n}(x_i!))$

$\frac{dl}{d\lambda} = -n+\frac{\sum_{i=1}^{n} x_i}{\lambda} = 0$

より

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n}$$

例題4 ポアソン分布のフィッシャー情報量

$$\mathrm{Var}\left(\frac{\partial}{\partial\lambda}l\right) = \mathrm{E}\left(-\frac{\partial^2}{\partial^2\lambda}l\right)$$

$$= \mathrm{E}\left(-\frac{\partial}{\partial\lambda}\left(\frac{1}{\lambda}\sum_{i=1}^{n}x_i - n\right)\right)$$

$$= \mathrm{E}\left(\frac{1}{\lambda^2}\sum_{i=1}^{n}x_i\right)$$

$$= \frac{n\lambda}{\lambda^2}$$

$$= \frac{n}{\lambda}$$

クラメール･ラオの不等式

$\hat{\theta}$をθの不偏推定値とする．つまり

$\theta = E(\hat{\theta}) = \int \hat{\theta} p(x|\theta) \mathrm{d}x$

両辺をθで微分する．

スコア関数を導入する工夫

$$1 = \frac{\partial}{\partial\theta}\int \hat{\theta} p(x|\theta)\mathrm{d}x = \int \hat{\theta}\frac{\partial}{\partial\theta} p(x|\theta)\mathrm{d}x$$

$$= \int \hat{\theta}\frac{\partial \ln p(x|\theta)}{\partial\theta} p(x|\theta)\mathrm{d}x = \mathrm{E}\left[\hat{\theta}\frac{\partial \ln p(x|\theta)}{\partial\theta}\right]$$

$$= \mathrm{E}\left[(\hat{\theta} - \mathrm{E}(\hat{\theta}))\left(\frac{\partial \ln p(x|\theta)}{\partial\theta} - \mathrm{E}\left(\frac{\partial \ln p(x|\theta)}{\partial\theta}\right)\right)\right]$$

$$\leq \left(V[\hat{\theta}]V\left[\frac{\partial \ln p(x|\theta)}{\partial\theta}\right]\right)^{\frac{1}{2}} = \left(V[\hat{\theta}]I(\theta)\right)^{\frac{1}{2}}$$

したがって

$$V[\hat{\theta}] \geq \frac{1}{I(\theta)}$$

不偏推定値の分散の下限値はフィッシャー情報量の逆数

強一致性

定義8　(強一致性)

推定値$\hat{\theta}$が真のパラメータθ^*に概収束するとき，$\hat{\theta}$は強一致推定値（strongly consistent estimator）であるという.

$$P(\lim_{n\to\infty} \hat{\theta} = \theta^*) = 1.0$$

つまり，データ数が大きくなると推定値が必ず真の値に近づいていくとき，その推定量を強一致推定値と呼ぶ.

最尤推定値の一致性

定理1 (最尤推定値の一致性)

最尤推定値$\hat{\theta}$は真のパラメータθ^*の強一致推定値である（Wald,1949）.

証明

以下を参照

https://qiita.com/PePrs/items/8d758e38df7a68004304

最尤推定値の漸近正規性

定義9

真の値θ^*の推定値 $\hat{\theta}$が**漸近正規推定量**（asymptotically normal estimator）であるとは，$\sqrt{n}(\hat{\theta} - \theta^*)$ の分布が正規分布に分布収束することをいう．すなわち，任意の$\theta^* \in \Theta^*$ と任意の実数に対して

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{as} N(0,\ \sigma^2(\theta^*))$$

$\sigma^2(\theta^*)$ を漸近分散（asymptotic variance）という．

最尤推定値の漸近正規性

定理2

確率密度関数が正則条件（regular condition）の下で，微分可能のとき，

最尤推定量は漸近分散$I(\theta^*)^{-1}$をもつ漸近正規推定量である.

$$I(\theta^*) = E_\theta\left[\left(\frac{\partial}{\partial\theta} lnL(\theta|\mathbf{x})\right)^2\right]$$

をフィッシャー（Fisher）の情報量と呼ぶ.

証明

以下を参照

https://qiita.com/PePrs/items/8d758e38df7a68004304

より複雑なモデル

$$y_i = w_0 + w_1 x_{i1} + w_2 x_{i2}^2 + \varepsilon_i,$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

入力　(x_{i1}, x_{i2}, y_i)(i=1, …n)

データファイル　の読み込み

パラメータw_0, w_1, w_2, σ^2を最尤推定せよ。

尤度は

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - w_0 - w_1 x_{i1} - w_2 x_{i2}^2)^2}{2\sigma^2}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\sum_{i=1}^{n} \frac{(y_i - w_0 - w_1 x_{i1} - w_2 x_{i2}^2)^2}{2\sigma^2}\right)$$

対数尤度は

$$l = n \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) \quad - \sum_{i=1}^{n} \frac{(y_i - w_0 - w_1 x_{i1} - w_2 x_{i2}^2)^2}{2\sigma^2}$$

非線形モデルは解析的に解けない

数値計算法

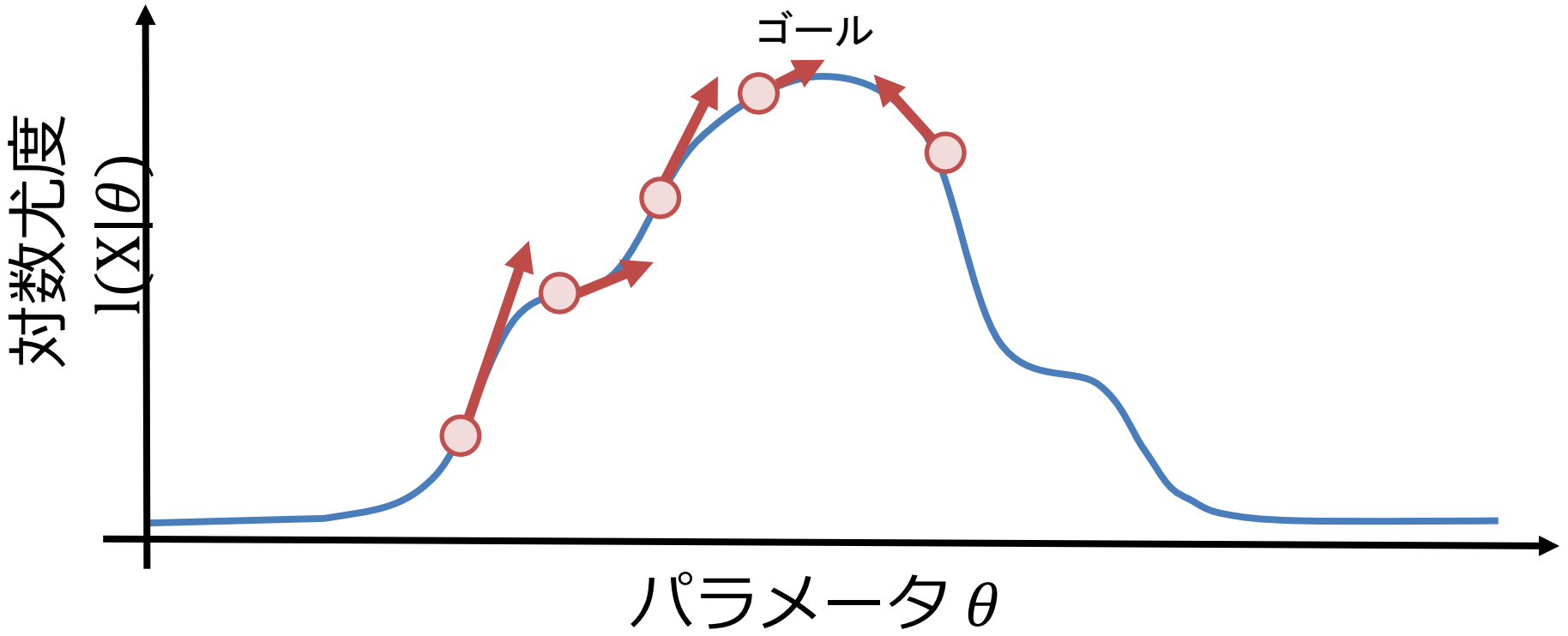
パラメータ推定値が解析的に求まらない場合には数値計算によって求める

代表的な手法

- 勾配上昇法
- ニュートン・ラフソン法

勾配上昇法（最急上昇法）

適当な初期値から、勾配方向にパラメータを更新することで極値（勾配0）を求める
傾きが正ならパラメータを正の方向へ、傾きが負ならば負の方向へ



最小値を求める問題の場合は

勾配降下法（最急降下法）と呼ばれる

勾配上昇法のアルゴリズム

パラメータθ　, 対数尤度関数 $l(X|\theta)$

<u>アルゴリズム</u>

1. パラメータθに適当な初期値を付与
2. 対数尤度関数の偏微分方向に微分値の η 倍更新

$$\theta_{n+1} = \theta_n + \eta \frac{\partial l(X|\theta)}{\partial \theta} : \forall n$$

3. 以下の収束条件を満たす（全てのパラメータ更新量が十分小さくなる＝ϵ 以下になる）まで2.を反復

$$\left| \frac{\partial l(X|\theta)}{\partial \theta} \right| \leq \epsilon : \forall n$$

一階偏微分

$$l = n \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) \quad - \sum_{i=1}^{n} \frac{(y_i - w_0 - w_1 x_{i1} - w_2 x_{i2}^2)^2}{2\sigma^2}$$

$$\frac{\partial l}{\partial w_0} = \sum_{i=1}^{n} \frac{(y_i - w_0 - w_1 x_{i1} - w_2 x_{i2}^2)}{\sigma^2}$$

$$\frac{\partial l}{\partial w_1} = \sum_{i=1}^{n} \frac{x_{i1}(y_i - w_0 - w_1 x_{i1} - w_2 x_{i2}^2)}{\sigma^2}$$

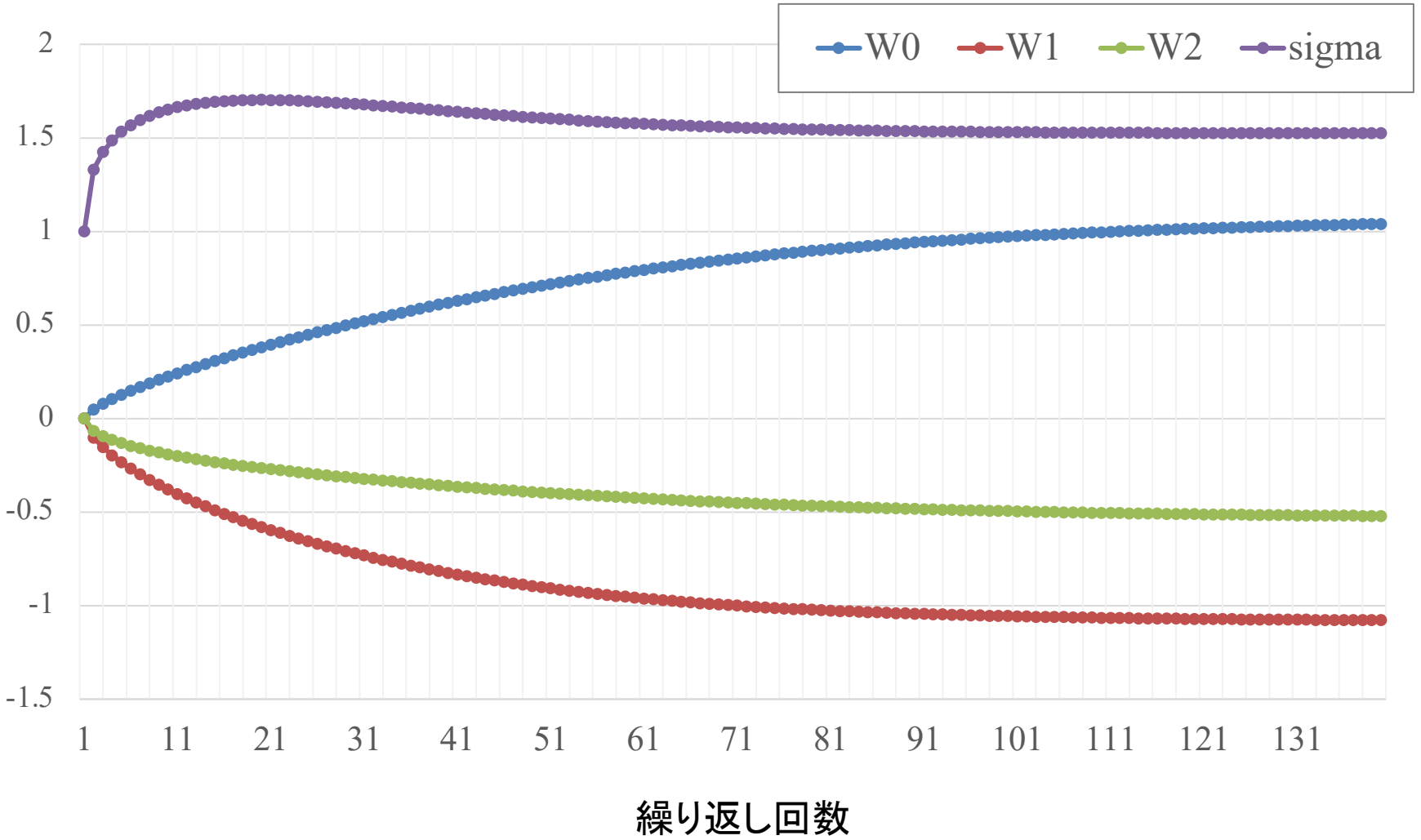
$$\frac{\partial l}{\partial w_2} = \sum_{i=1}^{n} \frac{x_{i2}^2(y_i - w_0 - w_1 x_{i1} - w_2 x_{i2}^2)}{\sigma^2}$$

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{\left(y_i - w_0 - w_1 x_{i1} - w_2 x_{i2}^2\right)^2}{\sigma^3}$$

推定例

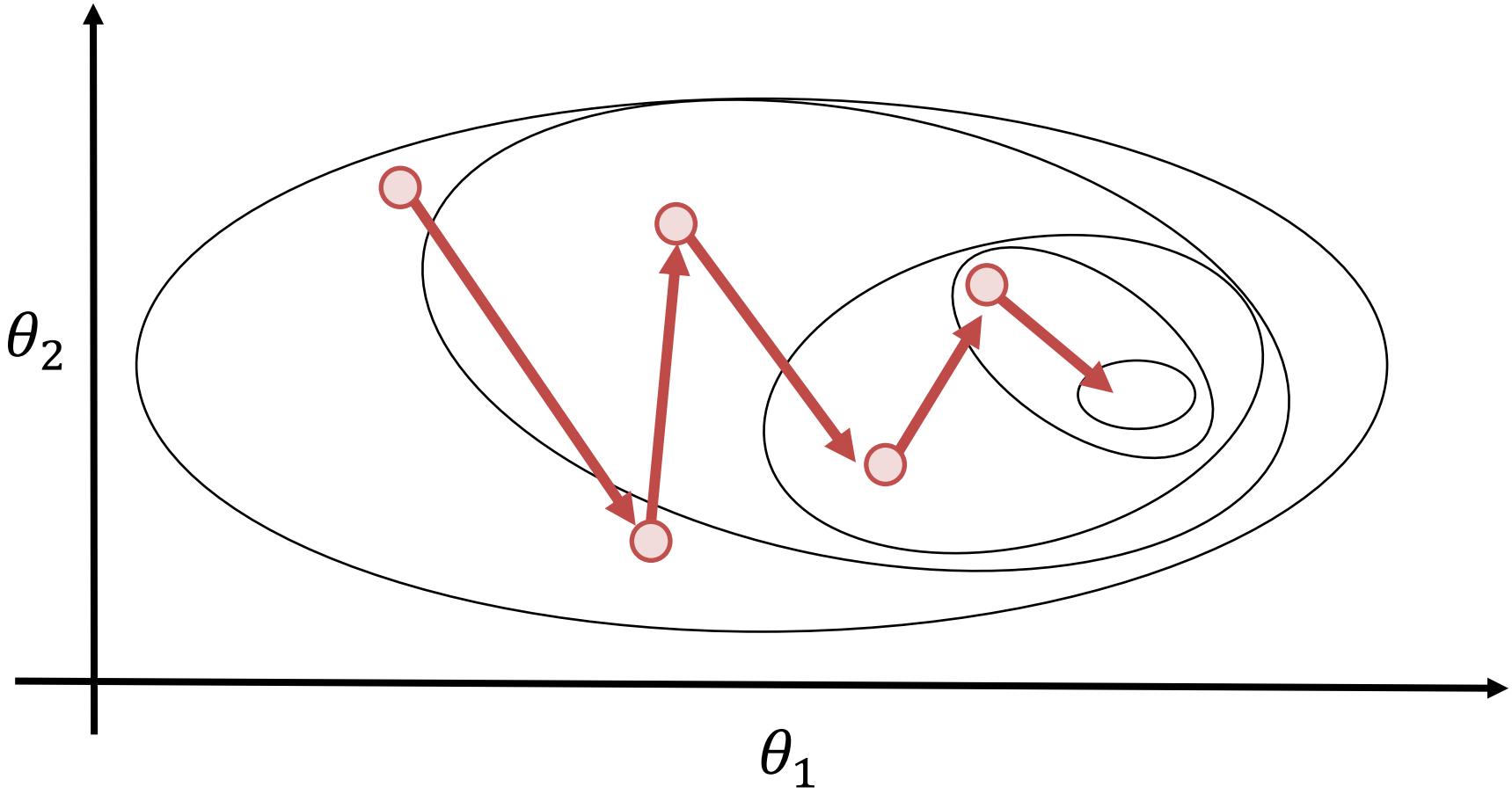
$\eta = 0.0001, \epsilon = 0.001$ サンプルサイズ1000

真値：$w_0 = 1.0, w_1 = -1.0, w_2 = -0.5, \sigma = 1.5,$



勾配上昇法の問題

勾配情報のみで更新方向を決定するため効率が悪い
勾配以外の情報を使用 ⇨ ニュートン･ラフソン法

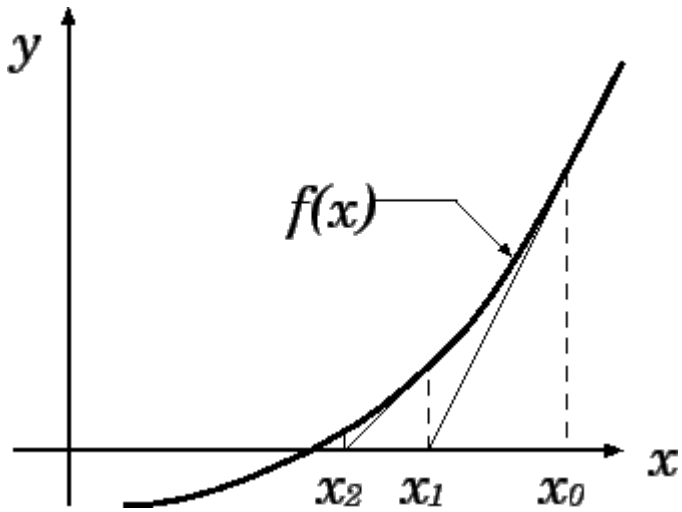


ニュートンラフソン法

方程式$f(x) = 0$を解く手法。

最大値問題の場合は、偏微分$f'(x) = 0$となるxを求める方程式を解けばよい。

ニュートン　ラフソン法



$f(x) = 0$を解く。
図のように適当な初期値 x_0においてf(x) に接線を引けば、接線の方程式は

$$y - f(x_0) = f'(x_0)(x - x_0)$$

X軸との交点は

$$x_1 = x_0 - f(x_0)/f'(x_0)$$

次にx_1での$f(x)$への接点とx軸との交点を求める。これを繰り返す。

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}$$

ニュートン法はテーラー近似

非線形関数の方程式$f(x_n) = 0$を解きたい。

$f(x_n)$をx_{n-1}のまわりでテーラー展開すると

$$f(x_n) = f(x_{n-1}) + f'(x_{n-1})(x_n - x_{n-1}) + O((x_n - x_{n-1})^2)$$

$f(x_n) = 0$より

$$f(x_{n-1}) + f'(x_{n-1})(x_n - x_{n-1}) = 0$$

これより、

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}$$

例 （1次元の場合）

$f(x) = x^2 - 2 = 0$ を解け（初期値 1.0とする）

ニュートンラフソン法を用いて 横軸に繰り返し数、縦軸にxの推定値を書け.

例

$f(x) = x^2 - 2 = 0$ を解け（初期値　1.0とする）

$$f'(x) = 2x$$

より

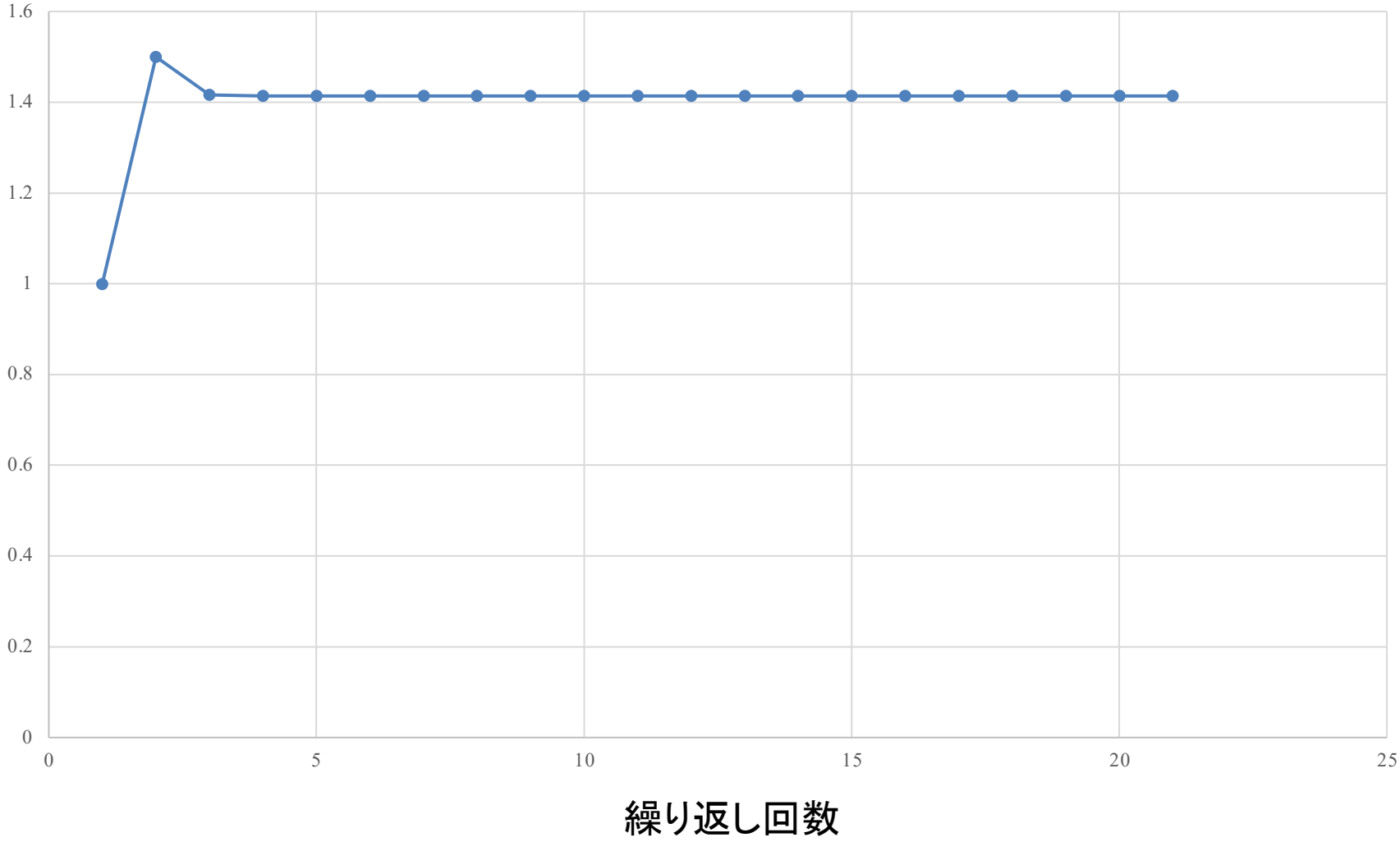
$$x_{n+1} = x_n - \frac{f(x)}{f'(x)}$$

$$x_{n+1} = x_n - \frac{x^2 - 2}{2x_n} = \frac{1}{2}\left(x_n + \frac{2}{x_n}\right)$$

初期値　1.0とする

数値例

推定値の遷移



多次元の場合の最尤法でのニュートン・ラフソン法

勾配（1階微分）に加えて、曲率（2階微分）を利用

パラメータ集合 $\boldsymbol{\theta} = \{\theta_1 \cdots \theta_N\}$, 対数尤度関数 $l(X|\boldsymbol{\theta})$とするとき、対数尤度関数の勾配行列 $g(\boldsymbol{\theta})$と2階微分行列:ヘッセ行列 $H(\boldsymbol{\theta})$をそれぞれ以下で表す

$$g(\boldsymbol{\theta}) = \begin{bmatrix} \dfrac{\partial l(X|\boldsymbol{\theta})}{\partial \theta_1} \\ \vdots \\ \dfrac{\partial l(X|\boldsymbol{\theta})}{\partial \theta_n} \end{bmatrix}, \qquad H(\boldsymbol{\theta}) = \begin{bmatrix} \dfrac{\partial l(X|\boldsymbol{\theta})^2}{\partial^2 \theta_1} & \cdots & \dfrac{\partial l(X|\boldsymbol{\theta})^2}{\partial \theta_1 \partial \theta_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial l(X|\boldsymbol{\theta})^2}{\partial \theta_n \partial \theta_1} & \cdots & \dfrac{\partial l(X|\boldsymbol{\theta})^2}{\partial^2 \theta_n} \end{bmatrix}$$

ニュートン・ラフソン法のアルゴリズム

パラメータ集合 $\boldsymbol{\theta} = \{\theta_1 \cdots \theta_N\}$, 対数尤度関数 $\mathrm{l}(X|\boldsymbol{\theta})$

<u>アルゴリズム</u>

1. 各パラメータ$\{\theta_1 \cdots \theta_N\}$に適当な初期値を付与
2. 対数尤度関数の偏微分方向に微分値の η倍更新

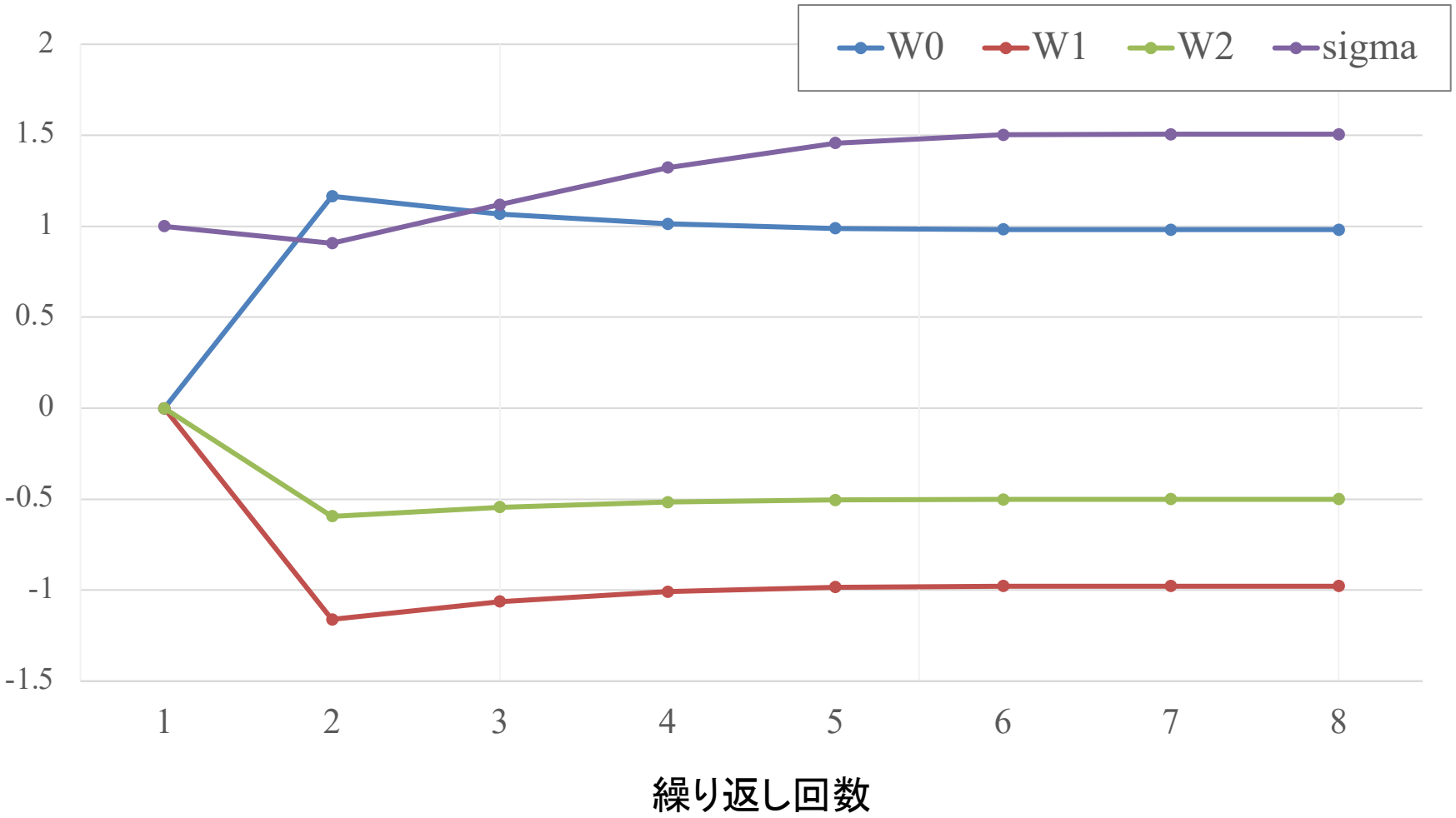
$$\boldsymbol{\theta} = \boldsymbol{\theta} - \eta\, H(\boldsymbol{\theta})^{-1} g(\boldsymbol{\theta})$$

3. 収束条件を満たす（全てのパラメータ更新量が十分小さくなる $= \epsilon$ 以下になる）まで2.を反復

推定例

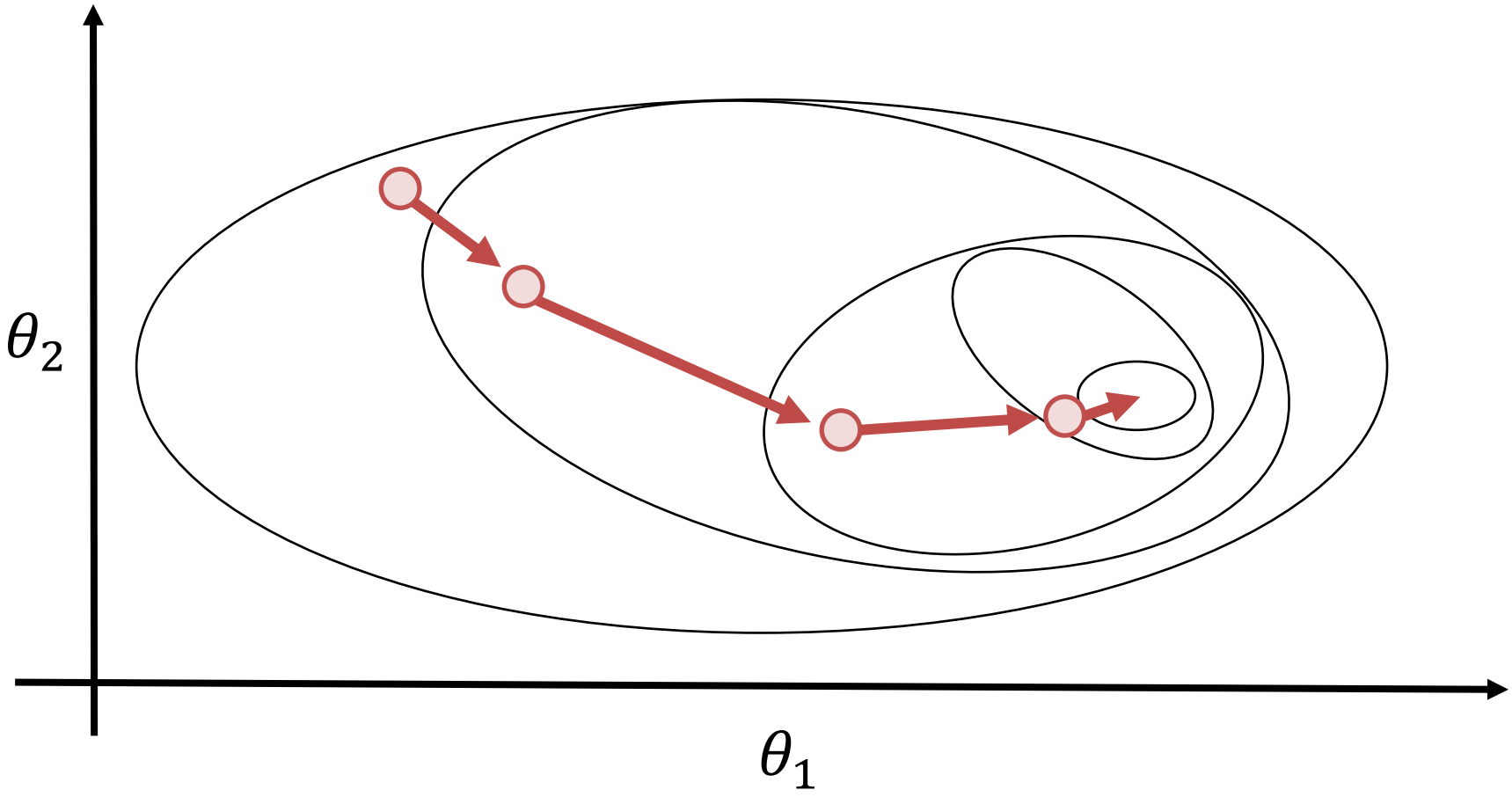
$\eta = 1.0, \epsilon = 0.001$ サンプルサイズ1000

真値:$w_0 = 1.0, w_1 = -1.0, w_2 = -0.5, \sigma = 1.5,$



ニュートン・ラフソン法のイメージ

曲率（勾配の変動）が大きい場所では更新幅を小さくし、
曲率が小さい場所では大きく更新



数値計算法の注意点

初期値依存

- 初期値によって推定値が発散することがある
- 発散したと判断される場合にはランダムに初期値を振り直して再スタートするなどの工夫が必要

学習率 η の設定

- 小さすぎると1ステップあたりの更新幅が小さくなり、収束に時間がかかる
- 大きすぎると極値を飛び越えてしまい収束しにくくなる。また、発散の可能性も高まる
- 適切な値を経験的に設定する必要がある

収束判定閾値 ϵ の設定

- 十分に小さく取るべき（例えば、0.001）だが、小さくするほど収束に時間がかかる

ニュートンラフソン法と最尤法はベイズ推定とは異なる

最尤法は$\frac{\partial l(X|\theta)}{\partial \theta} = 0$を解くためのアルゴリズム

$\rightarrow$

最大（最小）値しか解けない。

最尤法の尤度最大化には良いが　期待値を求めるベイズ推定では使えない場合が多い。

まとめ

１．古典的統計学の考え方では頻度のみが確率のよりどころ

２．対数尤度（Likelihood)とスコア関数（対数をとるのには意味がある）

３．最尤推定法（Maximum Likelihood Estimation)は　データ数が十分に大きいときに真の値に近づく

４．フィッシャー情報量は推定値の誤差の逆数を反映

５．複雑なモデルの最尤推定は　ニュートン・ラフソン法などの数値計算法を用いる.