画像工学特論講義ノート：
コンピュータビジョンにおけるパラメータ推定

玉木 徹

平成 21 年 11 月 25 日

まえがき

本書は、広島大学大学院工学研究科情報工学専攻における画像工学特論で講義した内容をもとに執筆した。コンピュータビジョンに関連する書籍は多数あるが、学生が自習するために詳しく書かれているテキストはほとんどないため、講義終了後に内容を整理し、web 公開のためにまとめなおした。

本書はまだ執筆途中であり、校正途中であり、つまりまだ書きかけである。しかし、締め切りのない執筆はいつまでも続くライフワークのようなものである。どこかで区切りをつけなければ、まとまるものもまとまらない。

内容に関するご意見や間違いの修正など、フィードバックをいただければ幸いである。

目 次

図 目 次

表 目 次

第2章　コンピュータビジョンにおけるパラメータ推定

Almost all problems in computer vision are related in one form or another to the problem of estimating parameters from noisy data.

Zhengyou Zhang

デジタルカメラで撮影した画像。ムービーデジカメで撮影した映像。これらを眺めるのは人間で、いまやそれを眺めて楽しむ人は世界中に何百万人といる。子供の顔が映っていたり、雄大な自然が映っていたり。小さな花をズームアップで撮った画像もあれば、大きなビルを画面いっぱいに切り取った映像もある。

画像に何が映っているのか？　人間はいとも簡単にそれを把握することができる。しかし、コンピュータに同じことをやらせるのは簡単ではない。それは、1960 年代にコンピュータに画像を見せて理解させるという研究が始まって以来、いまだに研究者が到達できていない目標である。コンピュータに画像や映像を理解させること。それがコンピュータビジョンという研究分野である。

それは画像処理という研究と同じではないのか？　たしかに、画像を処理するという意味では、コンピュータビジョンも画像処理も同じものかもしれない。しかし現代的な意味でのコンピュータビジョンとは、画像に映るものの性質を取り出して数学的なモデルを構築し、そのモデルを決める手法を指す。

画像に白い線が映っているとしよう。その線は直線だろうか？　曲線だろうか？　直線だとしたらそれを表す方程式はどのようなものだろうか？　曲線だったらどんな方程式が使えるのか？　直線の傾きや曲線の曲がり具合を決める係数（パラメータ）にはどのようなものがあるのか？　それらをどうやって求めたらいいのか？　どんな求め方がもっともよいのか？　よいというのはどんな意味でよいのか？

同じ物体をいくつものカメラで撮影しているとしよう。それらのカメラの関係はどのような数式で表わされるのだろうか？　そしてそれはどうやって求めるのか？　これはカメラキャリブレーション (camera calibration) と呼ばれる、コンピュータビジョンでは重要な問題設定の一つである。ある画像と別の画像が同じものかどうかを照合するには？　これは画像照合 (image matching) と呼ばれる問題で、画像と画像を一致させるための数式やパラメータを考えなければならない。物体の形状を復元するには？　これは 3 次元形状復元 (three-dimensional reconstruction) という問題で、3 次元幾何と 2 次元画像を結びつける数式を必要とする。他にも、人間がどのような姿勢をしているのか？　という姿勢推定 (pose estimation) や、物体がどのように移動しているのか？　という追跡 (tracking)、画像中の物体はどのように動いているのか？　という運動解析 (motion analysis) など、いくつもの問題がある。

これらのコンピュータビジョンの問題に共通しているのは、個々の問題に特化した数学的なモデル（つまり数式）を考え出すこと、そしてそのモデルのパラメータを求めること、である。すべての問題に共通した数学的モデルというものは存在しないが、すべての数学的モデルに共通したパラメータを求める手法というものは存在する。それがパラメータ推定 (parameter estimation) と呼ばれ、それ自体が独立したコンピュータビジョンの問題であるといってもいいものである。

何がそんなに難しいのだろうか？　直線の方程式は中学校で習ったし、傾きや切片はすぐに計算できるではないか。確かにその通りである。ある直線の方程式が

$$y = ax + b \tag{2.1}$$

で与えられたとしよう。ここで a は直線の傾きで、b は切片である。これは中学高校の数学でも習ったし、受験の問題にも出たかもしれない。この直線の 2 つのパラメータ a, b を求めるには、この直線が通る 2 つの点を与えればよかった。その二つの点を $(x_1, y_1), (x_2, y_2)$ としよう。すると、次のような連立方程式が得られる。

$$y_1 = ax_1 + b \tag{2.2}$$

$$y_2 = ax_2 + b \tag{2.3}$$

これを解けば、a, b はすぐに求められる。解き方はいろいろあるが、とりあえず両辺をそれぞれ引いて

$$y_1 - y_2 = a(x_1 - x_2) \tag{2.4}$$

より

$$a = \frac{y_1 - y_2}{x_1 - x_2} \tag{2.5}$$

が求まった。これを、たとえば最初の式に代入すれば、

$$b = y_1 - \frac{y_1 - y_2}{x_1 - x_2} x_1 \tag{2.6}$$

も求まった。いったい何が難しいというのだ？　高校で習った以上のことがあるのだろうか？

それがあるのだ。ちょっと考えてみてほしい。直線を通る点が 2 点よりも多かったら、たとえば 300 点ぐらいあったら、どうすればよいだろう？　直線上に乗っている点が与えられればいいけれども、測定誤差や人為ミスで、与えら得たデータが厳密に直線に乗っていないかもしれない。そのときにはどうやって方程式を解けばいいのだろう？　それからこの直線がまっすぐ上向きだったら、つまり $x_1 = x_2$ のときは、直線の傾きはどうやって求めるのだろう？そもそも傾きの定義はなんだろうか？　とても簡単な直線という数学的モデルでさえ、これだけいろいろなことを考えなければならないのだ。もっと複雑な数式を扱わなければならないときには、さらにややこしい複雑な状況がでてくるかもしれない。

ここでは、直線と、それよりもちょっとだけ複雑な円や楕円などの曲線 (conic) を題材にして、パラメータ推定の問題を説明していこうと思う。与えられたデータに誤差がある場合にはどのようにすればいいのか？曲線に適した数式とはどのようなものだろうか？　適したというのは、一体どういう意味で適しているのだろうか？　これらを考えていく過程で、適切な数学的なモデルを選ぶ必

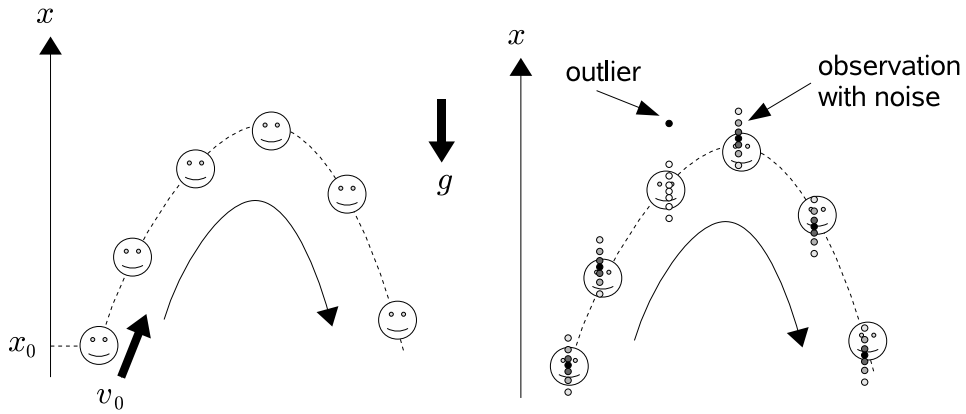

図 2.1: ボール落下の例。

要性、パラメータ推定の難しさ、そしてコンピュータビジョンの面白さを感じ取ってもらえればと思う。

なお、曲線のパラメータ推定に関しては Zhang[2] を、ロバスト推定はさらに Meer ら [1] を参考にしている。興味があれば、直接それらを参考にしてほしい。

2.1 パラメータ推定とは？

2.1.1 ボールが落下する例

まず、落下するボールを例にとって、その運動を記述するパラメータと推定における問題点を説明しよう。

図 2.1 に示すような、放り投げられて自由落下するボールの運動は、物理で習ったように以下のような運動方程式で記述される。

$$F = -mg \tag{2.7}$$

ここで F はボールにかかる力、m はボールの質量、g は重力加速度である。忘れていたら、高校の物理の教科書か古典力学の教科書を引っ張り出してこよう。地面からボールまでの高さ（つまりボールの位置）を x とすると、

$$m\frac{d^2x}{dt^2} = F = -mg \tag{2.8}$$

$$\frac{d^2x}{dt^2} = -g \tag{2.9}$$

となる。これを積分すれば速度 v となる。

$$v = \frac{dx}{dt} = \int -mgdt = -gt + v_0 \tag{2.10}$$

もう一度積分すれば、位置 x が出る。

$$x = \int -gt + v_0 dt = -\frac{1}{2}gt^2 + v_0 t + x_0 \tag{2.11}$$

ここで、x_0 は時刻 $t = 0$ での初期位置であり、v_0 は初期速度である。

明らかに、位置 x は時刻 t の二次関数である。一般的に書けば次のようになる。

$$x = at^2 + bt + c \tag{2.12}$$

ここで用いた a, b, c は、位置 x を決める（つまりボールがどのような放物線を描いて落下するのかを決定づける）パラメータである。先ほどとの対応を取れば、

$$a = -\frac{1}{2}g \tag{2.13}$$

$$b = v_0 \tag{2.14}$$

$$c = x_0 \tag{2.15}$$

となる。

さて、パラメータ a は重力加速度 g にしか依存していない。そして g は（教科書に載っているように）すでに値は求められている。このようなパラメータを「既知である」という。既知であるパラメータ (known parameters) には、すでに求めてある値を用いればよい。

問題は残りのパラメータ b, c である。初期位置や初期速度を変えれば、ボールが描く放物線も変化する。この放物線は物理の法則にのっとって導出したので、与えた初期位置や初期速度を用いて計算した放物線通りに軌跡を描くはずである。

しかし、キャッチボールをしているときには、どこに初期位置を定めてどのくらいの初期速度でボールを放り投げるのかは、当然のことながら分からない。そこで、これを「観測 (observation)」によって「計測 (measurement)」し、パラメータを「推定 (estimation)」して、ボールがどのような放物線を描いているのかを「予測 (prediction)」しよう。

推定するべきものは v_0, x_0 である。これらは未知パラメータ (unknown parameters) である。観測できるものは、各時刻 $t_1, t_2, \ldots$ におけるボールの位置 $x(t_1), x(t_2), \ldots$ である。これらの観測データ (observed data) から未知パラメータを推定する、という問題が「パラメータ推定」である。

いったい何個の位置を観測すれば推定できるのだろうか？　一般的には、方程式の数が未知パラメータの数以上になればよい。この場合の未知パラメータ (b, c) の数は 2 で、一つの観測 $x(t_i)$ につき一つの二次方程式が得られる。したがって、2 回観測すれば方程式が 2 つ得られて、それらを連立して解けば未知パラメータが求められる。やってみてほしい。

しかし普通はそれではうまくいかない。「観測」や「計測」には必ず「観測誤差」や「計測誤差」(error, noise) が入るからである。キャッチボールしている最中に、空中を飛んでいるボールの正確な高さを測ってほしいといわれて、あなたは何 mm の精度で計測できるだろうか？

このような誤差がない場合の、もともとのボールの挙動を表す二次方程式

$$x(t) = at^2 + bt + c \tag{2.16}$$

を、その現象を表現するモデル (model) という。観測では、モデルにはない誤差が含まれた値

$$x(t_1) + n_1,\ x(t_2) + n_2,\ \ldots \tag{2.17}$$

が計測される。どんな場合でも、すべての観測データには誤差が含まれていると考えるべきである。

この誤差は、正規分布に従うと仮定することが多い。つまり、すべての誤差 $n_1, n_2, \ldots$ は、平均 0、分散 σ の正規分布

$$n(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{2\sigma^2}} \tag{2.18}$$

から発生していると考えるのである。このように、誤差の挙動を表わす式を誤差モデル (noise model) という。誤差モデルは常に正規分布であるとは限らないが、たいていの場合には正規分布で十分であることが多い。応用によってはほかの分布を誤差モデルに採用した方がいい場合もあるが、ここでは省略しよう。

誤差モデルによって計測誤差がすべて表現できれば、それで問題ない。しかし、場合によってはこの誤差モデルで扱いきれないような誤差 α が観測データに含まれることがある。

$$x(t_1) + n_1,\ x(t_2) + n_2 + \alpha,\ \ldots \tag{2.19}$$

このような観測データは、もともとのモデルによっても表現できないし、誤差モデルを使っても表現することができない。このようなデータを外れ値 (outliers) と呼び、誤差モデルに含まれる観測データ (inliers) と区別する。

観測誤差がなければ、パラメータ推定の必要はない。モデルに基づいて連立方程式を解けば終わりである。観測誤差が誤差モデルに基づいている場合、パラメータ推定の一般的な手法を用いる必要がある。それがこれから説明していく内容である。観測データに外れ値が含まれている場合、その外れ値に影響されないような推定手法を考える必要がある。それが後の方で述べるロバスト推定 (robust estimation) である。

2.1.2 推定するときに考慮するべきこと

このような観測データからパラメータを推定する手法について、考えるべきことがいくつかある。

一つ目は、現象を記述するモデルと、誤差モデルである。これらはパラメータ推定をする前にあらかじめ検討しておかなければならない問題である。つまり、パラメータ推定においては、それらのモデルは既知であるとする。

二つ目は、パラメータを推定するための基準である。モデルと観測の間には誤差の分だけずれがある。そのため、どのような基準に基づいて元のモデルのパラメータを推定するのかによって、結果は大きく異なる。ここではその「ずれ」を、モデルと観測との「距離」と呼ぶ。距離を小さくすれば、モデルと観測のずれが小さくなる。後に出てくる説明では、大きく 2 種類の距離（代数的距離と幾何学的距離）がある。

三つ目は、推定に用いる手法である。与えられたモデルと、固定した基準を決めても、どのような方法で推定するのかによっても、結果は変わってくる。後に出てくる説明では、モデルを正規化するいくつかの異なる方法がある。

自分の問題においてパラメータを推定する必要が出てきたら、これらをいつも念頭において、どのような推定方法がもっともよいのか？ということを考えていてほしい。

2.1.3　一般的な定式化

パラメータ推定の一般的な形をここで示しておこう。直線でも円でもそのほかの複雑な場合でも、同じように扱うためには、パラメータとモデルを明確に定義しておく方が、後々になって楽である。

まずはパラメータから定義しよう。推定するべき未知パラメータが m 個あるとき、それをベクトルに並べて

$$\boldsymbol{p} = (p_1, \ldots, p_m)^T \tag{2.20}$$

と書くことにする。前述の例では、初期位置や初期速度、直線の傾きなどがこれにあたる。

次は現象を記述するモデルである。これを関数 $\boldsymbol{f}$ と書くことにする。観測にノイズが含まれない理想的な場合の、モデル $\boldsymbol{f}$ の出力を $\boldsymbol{Z}$ としよう。つまり $\boldsymbol{Z} = \boldsymbol{f}(\boldsymbol{p})$ と書きたいのである。しかし、どんなモデルでも $\boldsymbol{Z}$ が右辺にくくりだせるというわけではない。そこで、次のように定義する。

$$\boldsymbol{f}(\boldsymbol{p}, \boldsymbol{Z}) = \boldsymbol{0} \tag{2.21}$$

先ほど例として示した、ボールが落下する簡単な例の 2 次式を考えてみよう。ボールの位置 x は時刻 t の関数として以下のようにモデル化していた。

$$x(t) = at^2 + bt + c \tag{2.22}$$

ここで、パラメータは

$$\boldsymbol{p} = (b, c)^t \tag{2.23}$$

$$m = 2 \tag{2.24}$$

であり（a は既知であった）、モデルの出力は

$$\boldsymbol{Z} = (x, t)^T \tag{2.25}$$

である。この現象をモデル化する関数 f は、定義通りにしようとすれば次のようになる。

$$f(\boldsymbol{p}, \boldsymbol{Z}) = x - (at^2 + bt + c) = 0 \tag{2.26}$$

この場合は関数の出力はひとつ（スカラー）なので、ベクトル $\boldsymbol{f}$ ではなくスカラー f で書いている。

観測には必ずノイズが含まれる。そのノイズを ϵ と書くと、実際に観測されるデータは、モデルの理想的な出力にノイズが加わった

$$\boldsymbol{y} = \boldsymbol{Z} + \epsilon \tag{2.27}$$

になる。それぞれの観測データ $\boldsymbol{y}_i$ にはノイズが含まれているので、普通は $\boldsymbol{f}(\boldsymbol{p}, \boldsymbol{y}_i) = \boldsymbol{0}$ は成り立たない。

以上でパラメータ推定の登場人物は出そろった。パラメータ推定とは何をするのか？　それは、n 個の観測されたデータ $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n\}$ から、もっともよいパラメータ $\boldsymbol{p}$ を求めることである。

「もっともよい」とはどういうことだろうか？　それを決めるのが、目的関数 (objective function, cost function) と呼ばれるものである。これは、パラメータとすべての観測データの関数として次のように定義される。

$$F(\boldsymbol{p}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) \tag{2.28}$$

最適化問題 (optimization problem) とは、この目的関数を最適化することである。で、最適化とは、その関数の最大値や最小値を見つけて、そのときのパラメータを求めることである。そのため、最小化 (minimization) とか最大化 (maximization) とも呼ばれる。ただし、F を最大化したい場合には $-F$ を最小化する場合と同じなので、最小化なのか最大化なのかはあまり問題ではない。最適化＝最小化と言ってもよい理由はそれである。

最小値の数式での書き表わし方を覚えておこう。目的関数 F の最小値または最大値は、次のように書く。

$$\min_p F \tag{2.29}$$

$$\max_p F \tag{2.30}$$

ここで $\min_p$ とは、関数 F の引数であるパラメータ $\boldsymbol{p}$ をいろいろと変えたときの最小値、という意味である。一方、その最小値または最大値を与えるときのパラメータは、次のように書く。

$$\operatorname*{argmin}_p F \tag{2.31}$$

$$\operatorname*{argmax}_p F \tag{2.32}$$

これは混同しやすいので間違えないように。最小値を与えるパラメータを $\hat{\boldsymbol{p}}$ とすると、$\hat{\boldsymbol{p}} = \operatorname*{argmin}_p F$ は正しい数式であるが、$\hat{\boldsymbol{p}} = \min_p F$ は間違っている。

あるパラメータ $\boldsymbol{p}_1$ が目的関数の最小値であるための必要条件は、次の 2 つの式を満たすことである。

$$\frac{\partial F(\boldsymbol{p}_1, \boldsymbol{y}_1, \ldots)}{\partial \boldsymbol{p}} = 0 \quad \text{and} \quad \frac{\partial^2 F(\boldsymbol{p}_1, \boldsymbol{y}_1, \ldots)}{\partial \boldsymbol{p}^2} > 0 \tag{2.33}$$

この数式の意味は何だろうか？　ベクトル $\boldsymbol{p}$ の微分はどう定義するのか？　それを次に説明しよう。

2.1.4　ベクトルの微分

先ほどの数式は、それぞれヤコビ行列またはヤコビアン (Jacobian) とヘッセ行列またはヘシアン (Hessian) と呼ばれるものである。

ヤコビ行列

まずヤコビ行列から説明しよう。これはベクトル $\boldsymbol{p}$ での微分である。

微分の定義を思い出してみると、ある変数 x を少し動かしたときに、ほかの変数 y がどのくらい変化するのかを表すのが微分であった。x と y の関係が関数 f で表わされているとすると、

$$y = f(x) \tag{2.34}$$

のとき、x の微小変化分 dx と y の微小変化分 dy は、f の微分 $\frac{df}{dx}$ によった以下のように関係づけられていた。

$$dy = \frac{df}{dx}dx \tag{2.35}$$

これを多変数に拡張する。m 個の変数 $x_1, \ldots, x_m$ と n 個の変数 $y_1, \ldots, y_n$ が、n 個の関数 $f_1, \ldots, f_n$ で次のように関係づけられているとしよう。

$$y_1 = f_1(x_1, \ldots, x_m) \tag{2.36}$$
$$y_2 = f_n(x_1, \ldots, x_m) \tag{2.37}$$
$$\vdots \tag{2.38}$$
$$y_n = f_n(x_1, \ldots, x_m) \tag{2.39}$$

全微分の公式を思い出すと（微積分の教科書の中から探し出してみよう）、微小変化分同士の関係は次のようになる。

$$dy_1 = \frac{\partial f_1}{\partial x_1}dx_1 + \ldots + \frac{\partial f_1}{\partial x_m}dx_m \tag{2.40}$$
$$dy_2 = \frac{\partial f_2}{\partial x_1}dx_1 + \ldots + \frac{\partial f_2}{\partial x_m}dx_m \tag{2.41}$$
$$\vdots \tag{2.42}$$
$$dy_n = \frac{\partial f_n}{\partial x_1}dx_1 + \ldots + \frac{\partial f_n}{\partial x_m}dx_m \tag{2.43}$$

それでは、上の関係式をベクトルと行列で書き直してみよう。

$$d\boldsymbol{Y} = (dy_1, \ldots, dy_n)^T \tag{2.44}$$
$$d\boldsymbol{X} = (dx_1, \ldots, dx_m)^T \tag{2.45}$$

とすると、先ほどの式は次のように、係数行列を使った連立方程式の形で表わすことができる。

$$d\boldsymbol{Y} = \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{X}}d\boldsymbol{X} \tag{2.46}$$

ここで、

$$\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{X}} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_m} \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_m} \end{pmatrix} \tag{2.47}$$

を、ヤコビ行列という。これは微小変化分 $d\boldsymbol{X}, d\boldsymbol{Y}$ を行列計算で結びつけている。つまり、1 変数のときの微分係数に相当するもので、任意の非線形の関数 $f_1, \ldots, f_n$ を線形の行列計算で近似したものである。

パラメータ推定における目的関数 F の最小化では、パラメータ $\boldsymbol{p}$ でのヤコビ行列 $\frac{\partial F}{\partial \boldsymbol{p}}$ が必要である。目的関数 F はベクトルではなくスカラであるので、上の定義に従えば、ヤコビ行列は横ベクトル

$$\frac{\partial F}{\partial \boldsymbol{p}} = \left(\frac{\partial F}{\partial p_1}, \ldots, \frac{\partial F}{\partial p_m} \right) \tag{2.48}$$

となる。しかし、最適化問題ではこれを縦ベクトル

$$\frac{\partial F}{\partial \boldsymbol{p}} = \begin{pmatrix} \dfrac{\partial F}{\partial p_1} \\ \vdots \\ \dfrac{\partial F}{\partial p_m} \end{pmatrix} \tag{2.49}$$

にして用いることが多い。どちらの定義を用いてるのか間違えないように、常に確認しよう。

ヘッセ行列

ヤコビ行列は多変数関数の 1 微分に相当するものであったのに対して、ヘッセ行列は 2 階微分に相当する。

あるスカラ関数 F が多変数 $\boldsymbol{p}$ を引数とする時、そのヘッセ行列は次のように定義される。

$$\frac{\partial F}{\partial \boldsymbol{p}} = \begin{pmatrix} \dfrac{\partial F}{\partial p_1 p_1} & \cdots & \dfrac{\partial F}{\partial p_m p_1} \\ \vdots & & \vdots \\ \dfrac{\partial F}{\partial p_m p_1} & \cdots & \dfrac{\partial F}{\partial p_m p_m} \end{pmatrix} \tag{2.50}$$

パラメータ推定においては、$\frac{\partial^2 F}{\partial \boldsymbol{p}^2} > 0$ が必要条件であった。「行列が正である」とはどういうことだろうか？　これは、行列の固有値に関係している。このヘッセ行列のすべての固有値が正であるとき、その行列は正定値 (positive definit) であるといい、$\frac{\partial^2 F}{\partial \boldsymbol{p}^2} > 0$ と書く。もしすべての固有値が 0 以上の正であるとき、その行列は半正定値 (positive semi-definit) であるといい、$\frac{\partial^2 F}{\partial \boldsymbol{p}^2} \geq 0$ と書く。

行列が正定値であるかどうかは、数学的には非常に重要なことである。パラメータ推定においても、最小値が本当に最小値なのかどうかを判断する、重要な指標である。だが、これ以上のことはここでは必要ない。もし必要であれば、線形代数の教科書を参考にしていほしい。

ヤコビ行列とヘッセ行列の例

見慣れない数式になれるために、次の二つの関数のヤコビ行列とヘッセ行列を計算してみよう。

$$\boldsymbol{f}_1 = \begin{pmatrix} at^3 + bt + c \\ 3dt^2 + 2et + f \end{pmatrix}, \quad \boldsymbol{p}_1 = (a, b, c, d, e, f)^T \tag{2.51}$$

$$f_2 = abt^2 + b^2ct + d, \quad \boldsymbol{p}_2 = (a, b, c, d)^T \tag{2.52}$$

$\boldsymbol{f}_1$ のヤコビ行列は次の通り。

$$\frac{\partial \boldsymbol{f}_1}{\partial \boldsymbol{p}_1} = \begin{pmatrix} t^3 & t & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3t^2 & 2t & 1 \end{pmatrix} \tag{2.53}$$

f_2 のヤコビ行列とヘッセ行列は次の通り。

$$\frac{\partial f_2}{\partial \boldsymbol{p}_2} = (bt^2, at^2 + 2bct, b^2t, 1)^T \tag{2.54}$$

$$\frac{\partial^2 f_2}{\partial \boldsymbol{p}_2^2} = \begin{pmatrix} 0 & t^2 & 0 & 0 \\ t^2 & ct & 2bt & 0 \\ 0 & 2bt & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{2.55}$$

2.1.5　ベクトル微分の公式

ヤコビ行列のように、ベクトルで微分するような式がこの先よく出てくることになる。そのため、特殊な形の関数であれば、公式のように微分後の式を簡単に計算することができる。後の計算に必要となる式のいくつかを、ここで示しておこう。

ここではヤコビ行列の定義には以下の式を用いる。

$$\frac{\partial F}{\partial \boldsymbol{p}} = \left(\frac{\partial F}{\partial p_1}, \ldots, \frac{\partial F}{\partial p_m} \right) \tag{2.56}$$

またパラメータベクトルは以前に示してある。

$$\boldsymbol{p} = (p_1, \ldots, p_m)^T \tag{2.57}$$

係数ベクトル・行列としては次のものを使おう。

$$\boldsymbol{b} = (b_1, \ldots, b_m)^T \tag{2.58}$$

$$B = \begin{pmatrix} b_{11} & \cdots & b_{1m} \\ \vdots & & \vdots \\ b_{m1} & \cdots & b_{mm} \end{pmatrix} \tag{2.59}$$

$F = \boldsymbol{p}^T \boldsymbol{b}$ のとき

一つ目の例は、目的関数の形がパラメータの線形和になっているものである。

$$F = b_1 p_1 + \cdots + b_m p_m = \sum_{i=1}^{m} b_i p_i = \boldsymbol{p}^T \boldsymbol{b} \tag{2.60}$$

この微分は次のようになる。

$$\frac{\partial F}{\partial \boldsymbol{p}} = \frac{\partial \boldsymbol{p}^T \boldsymbol{b}}{\partial \boldsymbol{p}} = (b_1, \ldots, b_m) = \boldsymbol{b}^T \tag{2.61}$$

$F = \boldsymbol{b}^T \boldsymbol{p}$ のとき

二つ目の例は、一つ目と同じであるが、係数とパラメータの順序が逆になっているものである。

$$F = b_1 p_1 + \cdots + b_m p_m = \sum_{i=1}^{m} b_i p_i = \boldsymbol{b}^T \boldsymbol{p} \tag{2.62}$$

この微分は次のようになる。

$$\frac{\partial F}{\partial \boldsymbol{p}} = \frac{\partial \boldsymbol{b}^T \boldsymbol{p}}{\partial \boldsymbol{p}} = (b_1, \ldots, b_m) = \boldsymbol{b}^T \tag{2.63}$$

$F = \boldsymbol{p}^T B \boldsymbol{p}$ のとき

三つ目の例は、行列の前後にパラメータベクトルがある場合である。変形してみよう。

$$F = \boldsymbol{p}^T B \boldsymbol{p} = (p_1, \ldots, p_m) \begin{pmatrix} b_{11} & \cdots & b_{1m} \\ \vdots & & \vdots \\ b_{m1} & \cdots & b_{mm} \end{pmatrix} \begin{pmatrix} p_1 \\ \vdots \\ p_m \end{pmatrix} \tag{2.64}$$

$$= (p_1, \ldots, p_m) \begin{pmatrix} \sum_{j=1}^{m} b_{1j} p_j \\ \vdots \\ \sum_{j=1}^{m} b_{mj} p_j \end{pmatrix} \tag{2.65}$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{m} b_{ij} p_i p_j \tag{2.66}$$

これをあるパラメータ p_k で微分すると、次のようになる。

$$\frac{\partial F}{\partial p_k} = \sum_{j=1}^{m} b_{kj} p_j + \sum_{i=1}^{m} b_{ik} p_i \tag{2.67}$$

ヤコビ行列はこれを要素にしたものだから、以下のようになる。

$$\frac{\partial F}{\partial \boldsymbol{p}} = \left(\sum_{j=1}^{m} b_{1j}p_j + \sum_{i=1}^{m} b_{i1}p_i, \ldots, \sum_{j=1}^{m} b_{mj}p_j + \sum_{i=1}^{m} b_{im}p_i \right) \tag{2.68}$$

$$= \left(\sum_{j=1}^{m} b_{1j}p_j, \ldots, \sum_{j=1}^{m} b_{mj}p_j \right) + \left(\sum_{i=1}^{m} b_{i1}p_i, \ldots, \sum_{i=1}^{m} b_{im}p_i \right) \tag{2.69}$$

$$= (B\boldsymbol{p})^T + \boldsymbol{p}^T B \tag{2.70}$$

もし B が対称行列 $B^T = B$ であれば、$(B\boldsymbol{p})^T = \boldsymbol{p}^T B^T = \boldsymbol{p}^T B$ より、次のように簡単になる。

$$\frac{\partial F}{\partial \boldsymbol{p}} = 2(B\boldsymbol{p})^T = 2\boldsymbol{p}^T B \tag{2.71}$$

これらを覚えておけば、ベクトルでまとめて微分するコツがわかって計算が楽になる場合が多い。

2.2　最初の簡単な例：平均の推定に見る最尤推定と最小二乗推定

それではまず、いちばん簡単なパラメータ推定の問題から始めよう。これは先ほど説明した、ボール落下の問題設定よりも簡単である。

パラメータの数は 1 個だけ、観測されるデータもスカラーである、以下のような観測モデルを考えよう。

$$y_i = m + \epsilon \tag{2.72}$$

つまり、ある真の値 m に誤差 ϵ が含まれた観測データ y_i が得られるという問題である。ここでやるべきことは、たくさんの観測データ $y_1, y_2, \ldots$ が与えられたときに、もっとも「よい」パラメータ m を求めることである。

それは、平均を求めればよいのではないのか？　たしかにそうかもしれない。普通なら、$y_1, y_2, \ldots$ の平均を m とすれば済むことである。しかし、「なぜ」平均がよいのか？　どういう意味で平均が「もっともよい」のか？これを考察してみよう。なぜ平均がよいのかわからなければ、ほかの問題にどうやって応用したらよいかわからないからである。

2.2.1　尤度と最尤推定

まず、ノイズモデルを仮定しよう。よくつかわれるノイズモデルは、ϵ は正規分布 (Normal distribution, Gaussian distribution) に従うというものである。平均 m、標準偏差 σ の正規分布を $N(m, \sigma)$ と書くことにする。

観測データに含まれる微小なノイズ ϵ が、平均 0、標準偏差 σ の正規分布に従うことを、次のように表す。

$$\epsilon \sim N(0, \sigma) \tag{2.73}$$

すると観測データ y_i は、平均 m、標準偏差 σ の正規分布に従うことになる。

$$y_i \sim N(m, \sigma) \tag{2.74}$$

平均 m が未知であり推定するべきパラメータである。しかし、標準偏差 σ はどうだろうか？ もともとの観測モデルにはないパラメータである。これはノイズモデルを正規分布にしたために追加されたものであり、現象を記述するパラメータではなく、ノイズを記述するパラメータである。ここではまず、σ が既知であるとして話を進めよう。

図＊は、観測データ y_i が平均 m、標準偏差 σ の正規分布から生成されている様子を示している。m は未知であるが、そこを中心に観測データが生成されている。正規分布であるので、平均付近が最も確率か高く、平均から離れるに従って確率は低くなる。だから、観測データが平均付近に集中しているのは当然である。次に得られる観測データがまだ分からない場合でも、おそらく平均付近に現れるであろうことは、確率的にも直感的にも、すぐにわかるだろう。

では逆に、観測データが 1 点だけ与えられたとき、それを生成した正規分布はどのあたりにあるのだろうか？　その様子を図＊に示してある。正規分布の幅は標準偏差で決まり、それは今のところ既知であるとしている。問題は、その正規分布がどこくるのか、である。どこに正規分布が来るのかはわからないが、その正規分布の中心付近がもっとも確率が高いのは同じである。したがって、今与えられている 1 点の観測データの部分が正規分布の中心である、と思うしかない。

もういくつか観測データが得られたとしよう。そしてそのデータがある場所に集中しているとしよう。それならば、その集中している場所が最も確率が高そうであり、その付近を正規分布の中心にするべきであろう。

この考え方が、尤度 (likelihood) というものである。与えられた観測データを生成した確率分布（この場合は正規分布）がどこにありそうなのか、どこにあれば尤も（もっとも）らしいのか、ということを定量的に数字で表したものが尤度である。尤もらしくないところにある正規分布よりも、最も尤もらしいところにある正規分布のほうが、求める正規分布であろう。なんとまあ、文章で書くとややこしいことか。

パラメータ θ で表される、n 個の確率変数 $X_1, \ldots, X_n$ の同時確率密度関数 $f(X_1, \ldots, X_n \mid \theta)$ で表される確率分布があるとする。その尤度 $L(\theta \mid X_1, \ldots, X_n)$ は、次の式で定義される。

$$L(\theta \mid X_1, \ldots, X_n) = f(X_1, \ldots, X_n \mid \theta) \tag{2.75}$$

尤度といっても、つまりは、確率なのだ。ここで分かっておいてほしいのは、密度関数は確率変数（つまり観測データ）を生成する確率を表しているが、尤度は確率変数から密度関数を決めるためのものである。逆の関係にあるといってもいい。

例をあげよう。ある一つの確率変数 X_i が正規分布から生成される場合、つまり $X_i \sim N(m, \sigma)$ の場合の尤度はどうであろうか。確率密度関数は次のようになる。

$$f(X_i|\theta) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(X_i-m)^2}{2\sigma^2}} \tag{2.76}$$

ここで $\theta = m, \sigma$ である。尤度は、先ほどの定義から次のようになる。

$$L(\theta|X_i) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(X_i-m)^2}{2\sigma^2}} \tag{2.77}$$

なんのことはない。同じものである。

では n 個の確率変数 $X_1, \ldots, X_n$ が、統計的に独立に (statistically independentlly)、同じ正規分布から与えられたとしよう。この状況を i.i.d.(independent and identically distributed) と書くこともある。このとき、確率統計で習ったように、同時確率密度関数はここの確率密度関数の積で書き表すことができる。

$$L(\theta \mid X_1, \ldots, X_n) = f(X_1, \ldots, X_n \mid \theta) \tag{2.78}$$

$$= f(X_1 \mid \theta)\ f(X_2 \mid \theta)\ \cdots\ f(X_n \mid \theta) \tag{2.79}$$

$$= \prod_{i=1}^{n} f(X_i \mid \theta) \tag{2.80}$$

さらにこれを簡単にしたい。そこで、両辺の対数をとろう。なぜなら、積の対数をとると、和に変形できるからだ。

$$\log L(\theta \mid X_1, \ldots, X_n) = \sum_{i=1}^{n} \log f(X_i \mid \theta) \tag{2.81}$$

これを対数尤度 (log likelihood) という。

やりたいことは、最も尤もらしい、つまり尤度が大きい分布を探したいのである。ある分布の尤度がほかの分布の尤度よりも大きいかどうかという大小関係は、対数尤度で比較しても変わらない。そこで、対数尤度を最大化するパラメータを求める以下の最適化問題が登場する。

$$\max_{\theta} L = \max_{\theta} \log L = \min_{\theta} -\log L \tag{2.82}$$

この最適化問題を、最尤推定 (Maximum Likelihood Estimation, MLE) と呼ぶ。ただし次の例で示すように、対数尤度の符号を反転させて最小化問題を解く場合が多い。どちらにしても同じことである。

2.2.2　平均の最尤推定：分散既知の場合

では先ほどの平均を推定する簡単な例の答えを、最尤推定で求めてみよう。ノイズモデルは既知の標準偏差 σ をもつ正規分布、n 個の観測データが含むノイズは、この正規分布から統計的に独立に生成されたものであるとする。

正規分布の式は次の通り。

$$f(y_i|\theta) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(y_i-m)^2}{2\sigma^2}} \tag{2.83}$$

対数尤度を求めるためには、単にこれの対数をとればいいだけであるので、次のようになる。

$$-\log f(y_i|\theta) = \frac{1}{2}\log 2\pi\sigma + \frac{1}{2\sigma^2}(y_i - m)^2 \tag{2.84}$$

このようにマイナスの符号をつけておくと、右辺がすべて足し算になる。対数尤度の符号を反転させたのは、このためである。

では、その対数尤度を求めよう。先ほど示したように、f の対数をとったものの和をとればよい。

$$-\log L(\theta \mid y_1, \ldots, y_n) = -\sum_{i=1}^{n} \log f(y_i \mid \theta) \tag{2.85}$$

$$= \frac{n}{2} \log 2\pi\sigma + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - m)^2 \tag{2.86}$$

ここで、標準偏差 σ は分かっていることにしていたことを思い出してほしい。重要なのは正規分布の平均 m だけであり、そのほかのものは尤度の大小には影響しない。つまり、最小化には必要ないのである。したがって、第 1 項の $\frac{n}{2} \log 2\pi\sigma$ は使わないし、第 2 項の係数である $\frac{1}{2\sigma^2}$ も消すことができる。

結局、解くべき最適化問題は以下のようになる。

$$\min_m -\log L = \min_m \sum_{i=1}^{n} (y_i - m)^2 \tag{2.87}$$

この問題の最小化するべき目的関数 F は、次のものである。

$$F = \sum_{i=1}^{n} (y_i - m)^2 \tag{2.88}$$

これを微分して 0 になるパラメータを求めればよい。解いてみよう。

$$\frac{dF}{dm} = 0 \tag{2.89}$$

$$\frac{d}{dm} \sum_{i=1}^{n} (y_i - m)^2 = 0 \tag{2.90}$$

$$2 \sum_{i=1}^{n} (y_i - m) = 0 \tag{2.91}$$

$$\sum_{i=1}^{n} y_i - nm = 0 \tag{2.92}$$

$$m = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{2.93}$$

つまり、与えられた観測データ y_i の平均が、求めるパラメータ m の最尤推定の結果である。

やっぱり平均だったではないか。だから最初から平均を計算していればよかったのではないのか？　いや、そうではない。この平均であるという結果を導くまでに、どれだけの仮定を積み上げただろうか。まずノイズモデルは正規分布であると仮定した。その分散も既知であると仮定した。すべての観測データが、同じ正規分布から統計的に独立に生成されたノイズを含んでいると仮定した。これだけの仮定の上に、平均という結果が得られたのである。

逆に、平均をとって計算をするということの危うさを考えてほしい。もし得られた観測データのノイズが、正規分布に従っていなければ、平均は最尤推定結果にはならない。もし観測データによって違うノイズが加わるならば、平均は最尤推定にはならない。そして、そういう状況は結構多いのだ。どんな観測データが得られるのかということを、常に最初に考察してから、最適化問題を定式化するべきである。でなければ、やっていることとやるべきことが違ってしまう。

2.2.3　平均の最尤推定：分散未知の場合

標準偏差 σ が未知の場合は、結果はどうなるのだろう。先ほどは既知であると仮定して解いたが、やっぱり未知である場合も解きたい。

先ほどの対数尤度は次のようなものであった。これを目的関数 F とおこう。

$$F = \frac{n}{2}\log 2\pi\sigma + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - m)^2 \tag{2.94}$$

これを最小化する問題は、先ほどと同様に、微分して 0 とおけば解くことができる。

まず m で微分しよう。

$$\frac{\partial F}{\partial m} = 0 \tag{2.95}$$

$$\frac{-1}{\sigma^2}\sum_{i=1}^{n}(y_i - m) = 0 \tag{2.96}$$

$$\sum_{i=1}^{n}(y_i - m) = 0 \tag{2.97}$$

$$m = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2.98}$$

なんのことはない、結果は先ほどと同じである。つまり平均である。

次に σ^2 で微分しよう。σ^2 は分散であるが、こちらのほうが簡単である。

$$\frac{\partial F}{\partial \sigma^2} = 0 \tag{2.99}$$

$$\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - m)^2 = 0 \tag{2.100}$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - m)^2 \tag{2.101}$$

つまり、分散 σ^2 の最尤推定結果は、観測データの分散である。

2.2.4　任意関数の最尤推定：最尤推定と最小二乗誤差推定

平均という簡単な場合、最尤推定は比較的簡単に導出することができた。ではもっと複雑な、任意の関数という場合にはどうなるだろうか。

パラメータ θ をもつ、任意のスカラー関数 $g(x, \theta)$ が与えられたとき、以下のような観測モデルを考えよう。

$$y_i = g(x_i, \theta) + \epsilon \tag{2.102}$$

理想的な値は $g(x_i, \theta)$ であるが、それにノイズ ϵ が加わったものが観測データ y_i として得られる。

ノイズモデルは、やはり正規分布を仮定しよう。

$$\epsilon \sim N(0, \sigma) \tag{2.103}$$

すると観測データ y_i は、次のようになる。

$$y_i \sim N(g(x_i, \theta), \sigma) \tag{2.104}$$

先ほどの平均の例とそっくりである。こうなると、この後のやり方も全く同じようにできる。

観測データの確率分布は次の式である。

$$f(y_i|\theta) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(y_i - g(x_i,\theta))^2}{2\sigma^2}} \tag{2.105}$$

これがあれば、対数尤度を計算するのも簡単である。

$$-\log L = -\sum_{i=1}^{n} \log f(y_i \mid \theta) \tag{2.106}$$

$$= \frac{n}{2} \log 2\pi\sigma + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - g(x_i, \theta))^2 \tag{2.107}$$

さて、ノイズモデルの標準偏差は未知でも既知でも、結果は似たようなものなので、簡単のためにここでは既知であるとしよう。すると、標準偏差の部分は尤度の大小に影響しないので、解くべき最適化問題は結局以下のようになる。

$$\min_{\theta} -\log L = \min_{\theta} \sum_{i=1}^{n} (y_i - g(x_i, \theta))^2 \tag{2.108}$$

これを解くには、やはり微分して 0 とおけばよい。その結果は、任意の関数 g の形に依存するので、一般的な議論はここまでである。

この式の右辺に注目してほしい、右辺は、観測データ y_i と理想的な値 g との差だけが残っている。もっと正確にいえば、差の二乗和 (the sum of square of differences) である。つまり、目的関数

$$F = \sum_{i=1}^{n} (y_i - g(x_i, \theta))^2 \tag{2.109}$$

を最小化する問題になっている。そのため、このように差や距離の二乗和を最小にする最適化問題を、最小二乗推定 (Least Square Estimation, LSE) と呼ぶ。

最尤推定 (MSE) と最小二乗推定 (LSE) は、ノイズモデルを正規分布と仮定すれば、同じものである。差を二乗せずに絶対値をとった輪を最小化するような問題の定式化も可能であり、実際にそのような定式化もある。しかし、差の二乗和を最小化することは、ノイズモデルを正規分布と仮定すれば、理論的な裏付けがあるともいえる。しかし、正規分布に従うことが保証できない場合には、最小二乗推定はうまく働かないだろうということは、頭の片隅に覚えておいたほうがよい。もし正規分布よりも適切な確率分布があれば、それをノイズモデルに採用するべきであろう。そのような確率分布関数で表せないようなノイズも、当然存在する。そのようなときには、後述するロバスト推定をするべきである。最尤推定と最小二乗推定は、理解するべき基礎であるが、実際の問題に適用しなければならないのなら、そこにとどまっているべきではない。

2.3　直線の推定

前節の最後に説明した、任意の関数の最小二乗推定 (LSE) を使って、直線の推定をしてみよう。直線の推定は、平均の推定の次に簡単である。以下では、直線の場合の具体的な数式を導出してみよう。

このような直線の最小二乗推定は、線形回帰 (linear regression) とも呼ばれる。最小二乗推定それ自体は、差（距離）の二乗和を最小化するだけである。しかし、どのような距離を用いるのかということは、いろいろと変えることができる。前節の最後に説明したような、ある x_i のときの理想的な値 $g(x_i)$ と、その x_i のときの観測データ y_i の距離を使う場合、線形回帰と呼ばれるものになる。

2.3.1　直線の LSE：線形回帰の場合

直線の数式として、以下のものを考えよう。

$$g(x) = ax + b \tag{2.110}$$

ここでパラメータとなるのは、直線の傾き a と切片 b である。$x_1, x_2, \ldots$ に対応する観測データ $y_1, y_2, \ldots$ が得られた場合に、最小二乗推定によって直線を推定してみる。

最小化するべき目的関数 F は、前節のやり方に従えば次のようになる。

$$F = \frac{1}{2}\sum_{i=1}^{n}(y_i - (ax_i + b))^2 \tag{2.111}$$

これを最小化するパラメータを求める最適化問題

$$\min_{a,b} F \tag{2.112}$$

を解く。

この問題の解は、パラメータで微分すれば求まる。

$$\frac{\partial F}{\partial a} = 0 \tag{2.113}$$

$$\frac{\partial F}{\partial b} = 0 \tag{2.114}$$

具体的に微分してみよう。まずは a の微分。

$$\frac{\partial F}{\partial a} = 0 \tag{2.115}$$

$$\sum_{i=1}^{n}(-x_i)(y_i - ax_i - b) = 0 \tag{2.116}$$

$$a\sum_{i=1}^{n}x_i^2 + b\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}x_i y_i = 0 \tag{2.117}$$

(2.118)

次は b の微分。

$$\frac{\partial F}{\partial b} = 0 \tag{2.119}$$

$$\sum_{i=1}^{n} -(y_i - ax_i - b) = 0 \tag{2.120}$$

$$a\sum_{i=1}^{n} x_i + bn - \sum_{i=1}^{n} y_i = 0 \tag{2.121}$$

これで、二つの方程式が得られた。

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i = 0 \tag{2.122}$$

$$a\sum_{i=1}^{n} x_i + bn - \sum_{i=1}^{n} y_i = 0 \tag{2.123}$$

つまり連立方程式であり、求めるべき未知数はパラメータ a, b である。これは行列形式に直せばより分かりやすい。

$$\begin{pmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^2 \\ \sum_{i=1}^{n} x_i & n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i \end{pmatrix} \tag{2.124}$$

この連立方程式を、正規方程式 (normal equation) と呼ぶ。

この連立方程式を解くのは簡単である。行列部分と右辺は観測データから計算できるし、あとは 2×2 行列の逆行列を求めてかけるだけである。

2.3.2 線形回帰の問題点

線形回帰で直線が求められたが、これでなんの問題もなく直線の推定ができたことになるだろうか？　もちろんそうではない。線形回帰は便利で簡単であるが、いろいろと問題がある。

一つ目は、縦に伸びる直線を推定することはできないという問題である。直線を $y = ax + b$ という式で表現したが、垂直方向の直線は傾き a が無限大になってしまい、求めることはできない。もちろん切片も無限大になってしまう。直線が少しでも傾いていればよいではないか、と思うかもしれない。しかし、傾き a の値が非常に大きくなり、結局その値を扱うことができない。これは直線をどのようにパラメータで表現するかという問題であり、$y = ax + b$ という式が任意の直線を推定するという問題には適さないのである。

二つ目は、軸を変えたり回転したりすると結果が変わってしまうという問題である。線形回帰では $y = ax + b$ という式で直線を表していたため、最小化する差は、y_i と $g(x_i) = ax_i + b$ の距離であった。しかし、x 軸と y 軸を取り替えて $x = \frac{1}{a}y - \frac{b}{a}$ という式で直線を表してみよう。数式的にはまったく同じであるが、最小化する差には違いがある。今度の差は、x_i と $\frac{1}{a}y_i - \frac{b}{a}$ との差である。これは先ほどとはまったく違う距離の測り方であり、そのため結果も違ってくる。軸を取り替えるだけでなく、もっと一般的に xy 平面でデータを回転しても、同じ問題が生じる。このように、ある変換（この場合は軸の取り換えや回転）を施したときに結果が変わってしまうような性質は、好

ましくない。ある変換に対して結果が変わらないような性質を、その変換に不変である (invariant) という。この場合は、回転に不変ではない (not invariant) ので、問題が生じるのである。

三つ目は、直線以外の円や楕円などの一般的な形状には適用できないという問題である。直線の場合には、y_i と $g(x_i)$ の距離を使うことができた。しかし、円の場合には一つの x_i に二つの $g(x_i)$ が存在する。線形回帰で用いる距離の定義とは違うものを使わなければ、この問題は解決しないのだ。

これらの問題を解決するには、直線を表す式に別のものを使わなければならないし、最小化する距離に別のものを使わなければならない。

2.3.3　垂直距離の最小化

ここでは、直線の最小二乗推定のために、観測データから直線までの垂直距離を定義しよう。これは線形回帰で用いた距離とは異なり、先ほどの問題を解決するものである。

まずその垂直距離を導出し、その後で改めて垂直距離の二乗和を最小にする最適化問題を解くことにする。

2.3.4　直線までの垂直距離

ここでは以下の式で直線をパラメータ化しよう。

$$ax + by + c = 0 \tag{2.125}$$

線形回帰で用いた式とは異なり、x にも y にも係数がある。そのため、傾きが無限大になってしまうような問題は発生しない。ただし右辺が 0 であるので、両辺を何倍しても同じ直線を表してしまうので、このままではパラメータ a, b, c が一意に決まらない。そこで、とりあえずここでは $a^2 + b^2 = 1$ であるとしよう。このほかにも正規化の仕方はある。$c = 1$ にしてもよい。その比較は後の節で述べよう。

さて、ある観測データの座標を (x_0, y_0) としよう。この点から、直線までの垂直距離を求めたい。直線上のある点 (x, y) までの距離を d は、次のように表される。

$$d = \sqrt{(x - x_0)^2 + (y - y_0)^2} \tag{2.126}$$

点 (x_0, y_0) から直線までの垂直距離は、この距離 d の最小値である。つまり、次のような最適化問題を解かなければならない。

$$\min d \quad \text{subject to} \quad ax + by + c = 0 \tag{2.127}$$

ここで $a^2 + b^2 = 1$ である。勘違いしないように、確認しておこう。ここでは、与えられた直線までの垂直距離（つまり最短距離）を求めたい。だから、直線のパラメータ a, b, c は与えられており、既知である。観測点 (x_0, y_0) も既知である。求めるべきパラメータは、垂直距離を与える直線上の座標 (x, y) である。

このような問題は、制約条件付き最適化問題 (coditional optimization) と呼ばれる。単に目的関数である d を最小化するだけでなく、条件式 $ax+by+c=0$ を満たさなければならない。“subject to” とは、その条件式に従う (subject) という意味である。

このような制約条件付き最適化問題を解くための王道は、ラクランジュ乗数 (Lagrange multiplier) を用いる未定係数法である。つまり、ラクランジュ乗数 λ を用いて、制約条件付き最適化問題

$$\min f \quad \text{subject to} \quad g=0 \tag{2.128}$$

を変形して、次のような目的関数を最小化する問題を考えのだ。

$$F=f-\lambda g \tag{2.129}$$

こうすることで、制約条件のない最適化問題を解けばよいことになる。

先ほどの問題を、変形してみよう。距離 d の代わりに d^2 を用いても、値の大小関係は同じなので、これを最小化しよう。ラグランジュ乗数 λ を用いて、最小化するべき目的関数を次のように設定する。

$$F=(x-x_0)^2+(y-y_0)^2-\lambda(ax+by+c) \tag{2.130}$$

求めるべきパラメータは、直線上の点 (x,y) であった。これにラグランジュ乗数 λ も加わるので、次の三つの微分を計算しなければならない。

$$\frac{\partial F}{\partial x}=0 \tag{2.131}$$
$$\frac{\partial F}{\partial y}=0 \tag{2.132}$$
$$\frac{\partial F}{\partial \lambda}=0 \tag{2.133}$$

それぞれを計算してみよう。

$$\frac{\partial F}{\partial x}=2(x-x_0)-a\lambda=0 \tag{2.134}$$
$$\therefore\ x=\frac{a}{2}\lambda+x_0 \tag{2.135}$$
$$\frac{\partial F}{\partial y}=2(y-y_0)-b\lambda=0 \tag{2.136}$$
$$\therefore\ y=\frac{b}{2}\lambda+y_0 \tag{2.137}$$
$$\frac{\partial F}{\partial \lambda}=ax+by+c=0 \tag{2.138}$$

結局、次の三つの方程式を連立することになる。

$$x=\frac{a}{2}\lambda+x_0 \tag{2.139}$$
$$y=\frac{b}{2}\lambda+y_0 \tag{2.140}$$
$$ax+by+c=0 \tag{2.141}$$

ここからは、ややこしくなる。手順を追って説明しよう。まず、λ を求めよう。それには、パラメータ x, y の微分で得られた最初の二つの式を、λ の微分で得られた最後の式に代入する。

$$ax + by + c = 0 \tag{2.142}$$

$$a(\frac{a}{2}\lambda + x_0) + b(\frac{b}{2}\lambda + y_0) + c = 0 \tag{2.143}$$

$$a^2\lambda + 2ax_0 + b^2\lambda + 2by_0 + 2c = 0 \tag{2.144}$$

$$(a^2 + b^2)\lambda + 2(ax_0 + by_0 + c) = 0 \tag{2.145}$$

$$(a^2 + b^2)\lambda = -2(ax_0 + by_0 + c) \tag{2.146}$$

$$\lambda = \frac{-2(ax_0 + by_0 + c)}{a^2 + b^2} \tag{2.147}$$

次に、パラメータ x, y を求める。そのためには、求められた λ を、最初の二つの式に代入しなおす。

$$x = x_0 - \frac{a(ax_0 + by_0 + c)}{a^2 + b^2} \tag{2.148}$$

$$y = y_0 - \frac{b(ax_0 + by_0 + c)}{a^2 + b^2} \tag{2.149}$$

これで終わりではない。求めたいのは、距離 d である。いま求められた x, y を、d の式に代入しよう。

$$d^2 = \frac{a^2(ax_0 + by_0 + c)^2}{(a^2 + b^2)^2} + \frac{b^2(ax_0 + by_0 + c)^2}{(a^2 + b^2)^2} \tag{2.150}$$

$$= \frac{(ax_0 + by_0 + c)^2}{a^2 + b^2} \tag{2.151}$$

$$d = \frac{ax_0 + by_0 + c}{\sqrt{a^2 + b^2}} \tag{2.152}$$

$$= ax_0 + by_0 + c \quad (a^2 + b^2 = 1) \tag{2.153}$$

これで、ある点からの直線までの垂直距離が得られた。この距離は、高校の数学の教科書に、幾何学の公式として載っているものと同じである。

最初に $a^2 + b^2 = 1$ であると仮定していたのだが、結局最後の最後までその条件は使わなかった。だから、別に $a^2 + b^2 = 1$ でなくとも構わないわけだ。必要なのは、$a^2 + b^2$ で割れること。だから、$a^2 + b^2 \neq 0$ でありさえすればよい。もし $a^2 + b^2 = 0$ であったとしたら？　それはそもそも直線ではない。

次はこの垂直距離を使って直線を推定してみる。そのときに、正規化の仕方の違いで結果が違うのかをみてみるために、$c = 1$ という正規化と $a^2 + b^2 = 1$ という正規化の 2 種類で解いてみよう。

垂直距離の最小化：$c = 1$ の場合

さて、ある点から直線までの垂直距離の式が得られた。これを使って、直線の推定をしよう。つまり、与えられた観測データと直線の垂直距離の二乗和を最小にする最小二乗推定問題を解くのである。

n 個の観測データ (x_i, y_i) が与えられたときに、直線 $ax + by + c = 0$ までの垂直距離を d_i と書くことにする。ここでは $c = 1$ であるとしよう。今回は、この直線のパラメータ a, b, c が未知であり、求めるべきパラメータである。

最小化するべき目的関数は次のようになる。

$$J = \sum_{i=1}^{n} d_i^2 \tag{2.154}$$

$$= \sum_{i=1}^{n} (ax_i + by_i + c)^2 \tag{2.155}$$

$$= \sum_{i=1}^{n} (ax_i + by_i + 1)^2 \tag{2.156}$$

$$= \sum_{i=1}^{n} a^2 x_i^2 + b^2 y_i^2 + 2abx_i y_i + 2ax_i + 2by_i + 1 \tag{2.157}$$

これを解くためには、いつものように微分すればよい。

a, b で微分してみよう。

$$\frac{\partial J}{\partial a} = \sum_{i=1}^{n} 2ax_i^2 + 2bx_i y_i + 2x_i \tag{2.158}$$

$$= 2a \sum_{i=1}^{n} x_i^2 + 2b \sum_{i=1}^{n} x_i y_i + 2 \sum_{i=1}^{n} x_i = 0 \tag{2.159}$$

$$\therefore a \sum_{i=1}^{n} x_i^2 + b \sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} x_i = 0 \tag{2.160}$$

$$\frac{\partial J}{\partial b} = \sum_{i=1}^{n} 2by_i^2 + 2ax_i y_i + 2y_i \tag{2.161}$$

$$= 2b \sum_{i=1}^{n} y_i^2 + 2a \sum_{i=1}^{n} x_i y_i + 2 \sum_{i=1}^{n} y_i = 0 \tag{2.162}$$

$$\therefore b \sum_{i=1}^{n} y_i^2 + a \sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i = 0 \tag{2.163}$$

これを行列形式に直せば、以下の正規方程式が得られる。

$$\begin{pmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} x_i y_i & \sum_{i=1}^{n} y_i^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} y_i \end{pmatrix} \tag{2.164}$$

これを解くには、左辺の行列の逆行列をかければよい。

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} x_i y_i & \sum_{i=1}^{n} y_i^2 \end{pmatrix}^{-1} \begin{pmatrix} -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} y_i \end{pmatrix} \tag{2.165}$$

これで a, b が求められた。

だが、まってほしい。本当に、逆行列は存在するのか？　それが大問題なのだ。正規化のために $c = 1$ を仮定した。これが逆行列が存在しないことと関連している。直線の方程式が $ax+by+c = 0$ で

与えられているときに、$c = 1$ が満たされると仮定した。すると、任意の x, y に対して $ax+by+1 = 0$ が成り立たなければならないのだが、成り立たない場合があるのだ。$x = 0, y = 0$、つまり原点ではその式が成り立たない。

これはどういう意味だろうか？　直線の方程式が与えられているが、原点ではその式が成立しない。それは、直線が原点を絶対に通らないということを意味する。つまり、$c = 1$ という正規化をすると、原点を通るというよくある直線を表すことができないのだ。

逆に、直線が原点を通る場合を想定してみよう。傾きが α だとすると、直線の式は $y = \alpha x$ である。このとき、先ほどの行列は以下のようになる。

$$\begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n x_i y_i & \sum_{i=1}^n y_i^2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \alpha x_i \\ \sum_{i=1}^n x_i \alpha x_i & \sum_{i=1}^n (\alpha x_i)^2 \end{pmatrix} \tag{2.166}$$

$$= \begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n \alpha x_i^2 \\ \sum_{i=1}^n \alpha x_i^2 & \sum_{i=1}^n \alpha^2 x_i^2 \end{pmatrix} \tag{2.167}$$

$$= \begin{pmatrix} \sum_{i=1}^n x_i^2 & \alpha \sum_{i=1}^n x_i^2 \\ \alpha \sum_{i=1}^n x_i^2 & \alpha^2 \sum_{i=1}^n x_i^2 \end{pmatrix} \tag{2.168}$$

$$= \sum_{i=1}^n x_i^2 \begin{pmatrix} 1 & \alpha \\ \alpha & \alpha^2 \end{pmatrix} \tag{2.169}$$

つまりこの行列の 2 行目は、1 行目に α をかけたものになっている。

この行列の逆行列は存在するだろうか？　線形代数を思い出せば、ある列を定数倍したものが別の列になっている場合、その行列の行は独立ではなくなり、逆行列も存在しない。したがって、原点を通る直線にはこのやり方は使えない。原点ではないが原点に非常に近い点を通る場合はどうだろうか？　その場合もだめなのだ。逆行列は存在するとは言っても、行列式が非常に小さく（0 に近く）なるため、計算の精度が非常に悪くなる。

直線を求めるには、パラメータ化だけでなく正規化の仕方も重要なのである。パラメータ化には $y = ax + b$ と $ax + by + c = 0$ の二通りあり、さらに $ax + by + c = 0$ には正規化のやり方が 2 通りあるのだ。一つは $c = 1$ で、上に示したように使い物にはならない。二つ目が、$a^2 + b^2 = 1$ である。こちらが本命であり、次に説明しよう。

垂直距離の最小化：$a^2 + b^2 = 1$ の場合

ここでは $a^2 + b^2 = 1$ という正規化で直線を求めよう。直線 $ax + by + c = 0$ までの垂直距離が d_i である n 個の観測データ (x_i, y_i) が与えられている。今回も、この直線のパラメータ a, b, c が未知であり、求めるべきパラメータである。

最小化するべき目的関数は次のようになる。

$$J = \sum_{i=1}^n d_i^2 = \sum_{i=1}^n (ax_i + by_i + c)^2 \tag{2.170}$$

これを解くためには、いつものように微分すればよい、訳ではない。この式を直接微分せず、少し変形してから微分してみる。これは覚えておくべきテクニック、職人芸、試験問題に特有のトリックのようなもの、である。そのような方法は、今後も出てくるので、注意していてほしい。

まずはパラメータ c から求める。これは、素直に c で微分してみる。

$$\frac{\partial J}{\partial c} = 0 \tag{2.171}$$

$$\sum_{i=1}^{n} 2(ax_i + by_i + c) = 0 \tag{2.172}$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} c = 0 \tag{2.173}$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} y_i + nc = 0 \tag{2.174}$$

$$c = -\frac{a}{n}\sum_{i=1}^{n} x_i - \frac{b}{n}\sum_{i=1}^{n} y_i \tag{2.175}$$

$$\tag{2.176}$$

ここで、ちょっと注目してほしい、$\frac{1}{n}\sum_{i=1}^{n} x_i$ は x 座標の平均であり、$\frac{1}{n}\sum_{i=1}^{n} y_i$ は y 座標の平均である。だから、観測データの平均座標を (m_x, m_y) とおくと、次のように簡単になる。

$$c = -am_x - bm_y \tag{2.177}$$

a, b はまだ求まってはいないが、c は a, b でかけることが分かった。

ここで、この式を次のように書きなおしてみよう。

$$am_x + bm_y + c = 0 \tag{2.178}$$

この式が満たされなければならない。ということは、点 (m_x, m_y) が直線 $ax + by + c = 0$ 上にあるということである。つまり、求めるべき直線は（まだ求まってはいないが)、観測データの平均を通るのである。

さて、c が分かったので、目的関数に代入しなおして、変形してみよう。

$$J = \sum_{i=1}^{n}(ax_i + by_i - am_x - bm_y)^2 \tag{2.179}$$

$$= \sum_{i=1}^{n}(a(x_i - m_x) + b(y_i - m_y))^2 \tag{2.180}$$

$$= \sum_{i=1}^{n} a^2(x_i - m_x)^2 + 2ab(x_i - m_x)(y_i - m_y) + b^2(y_i - m_y)^2 \tag{2.181}$$

$$= \sum_{i=1}^{n} \begin{pmatrix} a & b \end{pmatrix} \begin{pmatrix} (x_i - m_x)^2 & (x_i - m_x)(y_i - m_y) \\ (x_i - m_x)(y_i - m_y) & (y_i - m_y)^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} \tag{2.182}$$

$$= \begin{pmatrix} a & b \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{n}(x_i - m_x)^2 & \sum_{i=1}^{n}(x_i - m_x)(y_i - m_y) \\ \sum_{i=1}^{n}(x_i - m_x)(y_i - m_y) & \sum_{i=1}^{n}(y_i - m_y)^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} \tag{2.183}$$

さて、目的関数を行列とベクトルで表したこの式の中身を、もう少し見てみよう。

2×2 行列の中身は、分散と共分散である。

$$\sigma_x^2 = \sum_{i=1}^{n}(x_i - m_x)^2 \tag{2.184}$$

$$\sigma_y^2 = \sum_{i=1}^{n}(y_i - m_y)^2 \tag{2.185}$$

$$\sigma_{xy}^2 = \sigma_{yx}^2 = \sum_{i=1}^{n}(x_i - m_x)(y_i - m_y) \tag{2.186}$$

これらを要素とする行列を次のように置く。

$$C = \begin{pmatrix} \sigma_x^2 & \sigma_{xy}^2 \\ \sigma_{yx}^2 & \sigma_y^2 \end{pmatrix} \tag{2.187}$$

この行列は、共分散行列 (covariance matrix) と呼ばれるものである。x と y の共分散が対称なので、共分散行列は対称行列になる。つまり $C^T = C$ である。

ここで、パラメータ a, b を要素に持つベクトルも、次のように $\boldsymbol{a}$ とおこう。

$$\boldsymbol{a} = \begin{pmatrix} a \\ b \end{pmatrix} \tag{2.188}$$

ここで、$a^2 + b^2 = 1$ としていたことを思い出してほしい。このベクトルで表現すれば、$||\boldsymbol{a}||^2 = 1$ である。

これらを用いると、先ほどの目的関数の式は非常にすっきりする。

$$J = \boldsymbol{a}^T C \boldsymbol{a} \quad \text{where} \quad ||\boldsymbol{a}||^2 = 1 \tag{2.189}$$

さて、ベクトルの微分の公式によれば、J を $\boldsymbol{a}$ で微分したら次のようになる。

$$\frac{\partial J}{\partial \boldsymbol{a}} = 2\boldsymbol{a}^T C \tag{2.190}$$

これが 0 になれば、そのときの $\boldsymbol{a}$ が求めるパラメータである。つまり、与えられた行列 C に対して次の式を満たすベクトル $\boldsymbol{a}$ を求めることになる。

$$\boldsymbol{a}^T C = \boldsymbol{0} \quad \text{or} \quad C\boldsymbol{a} = \boldsymbol{0} \tag{2.191}$$

ある行列にかけると 0 になってしまうようなベクトルを、その行列の零ベクトル (null vector) という。

これを求めれば、ようやく直線のパラメータがすべて求まることになる。零ベクトルの求め方は、ちょっと後で説明しよう。そのまえに、垂直距離とは別のアプローチで直線を推定する方法をみてみよう。

2.3.5　MEL:垂直距離の最小化とおなじ

前節では、観測データから直線までの垂直距離を最小化する問題を解いて直線を推定した。しかし、面倒くさい。最小二乗推定 (LSE) は、距離を定式化する必要があるが、距離を出すだけでも一苦労である。

では、最尤推定 (MLE) で直線を推定できないか？　ただし線形回帰ではない方法を使って。このやり方を説明しよう。

最尤推定では、まずノイズモデルを仮定しなければならない。ここでもやはり、標準偏差が既知の正規分布であるとしよう。ただし、今度は 2 次元平面上の点に、どの方向にも均等にノイズがのる（等方的)、と仮定する。つまり、平均 $\boldsymbol{m} = \mathbf{0}$ で共分散行列が

$$\Lambda_x = \sigma I \tag{2.192}$$

で与えられる 2 次元正規分布

$$f(\boldsymbol{x}\ |\theta) = \frac{1}{2\pi|\Lambda_x|} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})\Lambda_x^{-1}(\boldsymbol{x}-\boldsymbol{m})} \tag{2.193}$$

から生成されるノイズが、観測データに含まれているとする。

正規分布の等方性を仮定する理由は、簡単になるからである。観測データを $\boldsymbol{x}_i = (x_i, y_i)^T$、平均を $\boldsymbol{m} = (m_x, m_y)^T$ とすると、2 次元正規分布は二つの 1 次元正規分布の積であらわすことができる。

$$\begin{aligned} f(\boldsymbol{x}\ |\theta) &= \frac{1}{2\pi|\Lambda_x|} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})\Lambda_x^{-1}(\boldsymbol{x}-\boldsymbol{m})} && (2.194)\\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-m_x)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-m_y)^2}{2\sigma^2}} && (2.195)\\ &= \frac{1}{2\pi\sigma^2} e^{-\frac{(x_i-m_x)^2+(y_i-m_y)^2}{2\sigma^2}} && (2.196) \end{aligned}$$

こうなれば尤度を計算するのは簡単である。ノイズモデルが分散が σ（既知）で平均 $\mathbf{0}$ の 2 次元正規分布なので、観測データは真の（ただし未知の）点 $\boldsymbol{x}_{ti} = (x_{ti}, y_{ti})^T$ を平均とする分散が σ（既知）の 2 次元正規分布に従う。

対数尤度は、以下のように計算しよう。

$$\begin{aligned} -\log L &= -\log f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) && (2.197)\\ &= -\log \prod_{i=1}^{n} f(\boldsymbol{x}_i) && (2.198)\\ &= -\log \prod_{i=1}^{n} f(x_i) f(y_i) && (2.199)\\ &= -\log \prod_{i=1}^{n} \frac{1}{2\pi\sigma^2} e^{-\frac{(x_i-x_{ti})^2+(y_i-y_{ti})^2}{2\sigma^2}} && (2.200)\\ &= n \log 2\pi\sigma^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left\{(x_i - x_{ti})^2 + (y_i - y_{ti})^2\right\} && (2.201) \end{aligned}$$

ここで、分散 σ^2 は既知であるので、最小化には関係ない。したがって、最小化するべきものは第二項の和の部分だけである。ただし、真の（未知の）点 $\boldsymbol{x}_{ti} = (x_{ti}, y_{ti})^T$ は直線上にあるとする。

$$\min_{\{x_i, y_i\}} \sum_{i=1}^{n} \left\{(x_i - x_{ti})^2 + (y_i - y_{ti})^2\right\} \quad \text{subject to} \quad ax_{ti} + by_{ti} + c = 0 \tag{2.202}$$

この制約条件付き最適化問題は、やはりラグランジュ乗数 λ を用いて制約条件のない最適化問題に変形する。

$$J = \sum_{i=1}^{n} \left\{ (x_i - x_{ti})^2 + (y_i - y_{ti})^2 \right\} - \sum_{i=1}^{n} \lambda_i (a x_{ti} + b y_{ti} + c) \tag{2.203}$$

ここではまず、真の点 $\boldsymbol{x}_{ti} = (x_{ti}, y_{ti})^T$ を求めなければならない。直線のパラメータを求めるのはそのあとである。

それでは、目的関数 J を $x_{ti}, y_{ti}, \lambda_i$ で微分しよう。

$$\frac{\partial J}{\partial x_{ti}} = -2(x_i - x_{ti}) - \lambda_i a = 0 \tag{2.204}$$

$$\therefore\ x_{ti} = x_i + \frac{a}{2}\lambda_i \tag{2.205}$$

$$\frac{\partial J}{\partial y_{ti}} = -2(y_i - y_{ti}) - \lambda_i b = 0 \tag{2.206}$$

$$\therefore\ y_{ti} = y_i + \frac{b}{2}\lambda_i \tag{2.207}$$

$$\frac{\partial J}{\partial \lambda_i} = a x_{ti} + b y_{ti} + c = 0 \tag{2.208}$$

i の和をとる式を、ある i だけの x_{ti}, y_{ti} などで微分するようなやり方には慣れておいたほうがよいだろう。一見複雑に見えても、このように非常に簡単になる場合がある。

さて、λ_i で微分して得た式に x_{ti}, y_{ti} を代入しよう。

$$a x_{ti} + b y_{ti} + c = 0 \tag{2.209}$$

$$a\left(x_i + \frac{a}{2}\lambda_i\right) + b\left(y_i + \frac{b}{2}\lambda_i\right) + c = 0 \tag{2.210}$$

$$a x_i + b y_i + c = -\frac{\lambda_i}{2}(a^2 + b^2) \tag{2.211}$$

$$\therefore\ \lambda_i = \frac{-2(a x_i + b y_i + c)}{a^2 + b^2} \tag{2.212}$$

これで λ_i が得られた。

最後は、もともと最小化したい式に x_{ti}, y_{ti} を代入し、さらに λ_i も代入する。こうすると、余計なものがなくなってくる。

$$J_1 = \sum_{i=1}^{n} \left\{ (x_i - x_{ti})^2 + (y_i - y_{ti})^2 \right\} \tag{2.213}$$

$$= \sum_{i=1}^{n} \left\{ \left(\frac{a}{2}\lambda_i\right)^2 + \left(\frac{b}{2}\lambda_i\right)^2 \right\} \tag{2.214}$$

$$= \sum_{i=1}^{n} \left(\frac{a^2 + b^2}{4} \lambda_i^2 \right) \tag{2.215}$$

$$= \sum_{i=1}^{n} (a x_i + b y_i + c)^2 \tag{2.216}$$

この目的関数 J_1 を最小化したいのだから、最小化するために求めるパラメータは a, b, c しかない。つまり、次のような制約条件のない最適化問題に落ち着く。

$$\min_{a,b,c} \sum_{i=1}^{n} (ax_i + by_i + c)^2 \tag{2.217}$$

さて、これはどこかで見た式ではないだろうか？　これは垂直距離を最小化する問題と同じなのだ。その解は、共分散行列の零ベクトルであった。それは次に説明しよう。

一つだけ注意しておきたい。最尤推定による直線の推定が、垂直距離を最小化して得られる直線と同じになった。しかしこれは、非常に簡単な問題設定にしていたことを思い出してほしい。つまり、観測データに含まれるノイズのモデルを、等方的な 2 次元の正規分布であると仮定していたのだ。そうでなければ、二つの最適化問題は同じものにはならない。もしノイズモデルに正規分布以外のものを用いたら、もし非等方的な正規分布を仮定したら、上のような簡単な目的関数には変形することはできない。

最尤推定は、適切なノイズモデルを与えれば、その威力を発揮する。しかし、ある特定の問題に対してどんなノイズモデルが適切か？　それが分からない場合には（等方的な）正規分布が一番簡単で、まあまあ間違ってはいない。ただしそれは、大体において非常に単純な最小二乗推定と同じになることを覚えておこう。逆に、最小二乗推定は非常に簡単であり、単純な正規分布を（暗黙のうちに）仮定してしまっている、ということも覚えておこう。

2.3.6　零ベクトルの求め方

さて、直線を推定するための零ベクトルの求め方を説明しよう。ただし、このような問題は今後も出てくるので、一般的な条件で説明することにする。

求めるべき m 個のパラメータをもつベクトルを $\boldsymbol{p}$ としよう。正規化のために、ノルムを 1 とする。つまり $||\boldsymbol{p}|| = 1$ である。そして最小化するべき目的関数を次のように置く。

$$J = \boldsymbol{p}^T B \boldsymbol{p} \tag{2.218}$$

ここで $m \times m$ 行列 B は対称行列、つまり $B = B^T$ とする。そして解くべき最適化問題は次のものである。

$$\min_{\boldsymbol{p}} J \quad \text{with} \quad ||\boldsymbol{p}|| = 1 \tag{2.219}$$

J を $\boldsymbol{p}$ で微分すると、以下のようになる。

$$\frac{\partial J}{\partial \boldsymbol{p}} = B\boldsymbol{p} = \boldsymbol{0} \tag{2.220}$$

$\boldsymbol{p}$ は行列 B の零ベクトルである。しかしノイズの影響により、実際の問題では厳密に右辺が $\boldsymbol{0}$ になることはない。それではどうやって J を最小化すればよいだろうか？　それは固有値問題 (Eigenvalue problem) を解くのである。ということで固有値問題をおさらいしよう。忘れていれば線形代数の教科書の後ろのほうを参照してほしい。

行列 B の固有ベクトル (eigenvector) $\boldsymbol{e}_i$ とは、次式を満たすベクトルである。

$$B\boldsymbol{e}_i = v_i\boldsymbol{e}_i \tag{2.221}$$

ここで v_i は、固有ベクトル $\boldsymbol{e}_i$ の固有値 (eigenvalue) である。実対称 $m \times m$ 行列 B は m 個の固有ベクトルと固有値を持つ。詳しくは線形代数の教科書を眺めてほしい。さらに、固有ベクトルのノルムは 1 であり、異なる固有ベクトル同士は直交している。これを式で書くと次のようになる。

$$\boldsymbol{e}_i^T\boldsymbol{e}_j = \delta_{ij} \tag{2.222}$$

上のことを行列形式で書こう。行列 B は、以下のように行列の積で書くことができる。

$$B = UDU^T \tag{2.223}$$

ここで D は固有値を小さい順に並べた対角行列、U は固有ベクトルを並べた直交行列である。

$$D = \begin{pmatrix} v_1 & & & \\ & v_2 & & \\ & & \ddots & \\ & & & v_m \end{pmatrix}, \quad v_1 \le v_2 \le \cdots \le v_m \tag{2.224}$$

$$U = (\boldsymbol{e}_1\ \boldsymbol{e}_2\ \cdots\ \boldsymbol{e}_m), \quad UU^T = U^U = I \tag{2.225}$$

このような問題を対角化 (diagonalization) ともいう。非対角行列 B の左右から U をかけると対角行列 D になるため、そのような名前が付いている。

さて、m 個の固有ベクトルはノルムが 1 で互いに直交しているため、m 次元空間の正規直交基底である。つまり、任意の m 次元ベクトルは m 個の固有ベクトルの線形和で書くことができる。$\boldsymbol{p}$ をそうしていけない理由はない。ではそうしてみよう。

$$\boldsymbol{p} = a_1\boldsymbol{e}_1 + \cdots + a_m\boldsymbol{e}_m = \sum_{i=1}^{m} a_i\boldsymbol{e}_i \tag{2.226}$$

線形代数をよく知っているなら、係数 a_i はすぐに求められると思うだろう。つまり、$a_i = \boldsymbol{p}^T\boldsymbol{e}_i$ という内積を使うのである。しかし、思い出してほしい。$\boldsymbol{p}$ はこれから求めるべきパラメータであり、未知である。だからそんなに簡単な話ではない。

$\boldsymbol{p}$ のノルムを 1 に正規化していたので、固有ベクトルで表した式でもそれが成り立っていなければならない。では計算してみよう。簡単のためのノルムの二乗を計算するが、同じことなので気に

しないように。

$$||\boldsymbol{p}||^2 = \boldsymbol{p}^T\boldsymbol{p} \tag{2.227}$$

$$= (\sum_{i=1}^{m} a_i\boldsymbol{e}_i)^T \sum_{i=1}^{m} a_i\boldsymbol{e}_i \tag{2.228}$$

$$= (\sum_{i=1}^{m} a_i\boldsymbol{e}_i^T) \sum_{j=1}^{m} a_j\boldsymbol{e}_j \tag{2.229}$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j \boldsymbol{e}_i^T\boldsymbol{e}_j \tag{2.230}$$

$$= \sum_{i=1}^{m} a_i a_i \tag{2.231}$$

$$= \sum_{i=1}^{m} a_i^2 = 1 \tag{2.232}$$

つまり、$\boldsymbol{p}$ のノルムが 1 であるという制約が、係数の二乗和が 1 であるという制約に置き換わったのである。

では、何をどう解けばいいだろうか？　もともとは、与えられた行列 B があって、目的関数 J を最小化するベクトル $\boldsymbol{p}$ を求めたいのであった。しかし今は、B は対角化され、$\boldsymbol{p}$ はその固有ベクトルで表されている。ここで分からないのは係数 $\{a_i\}$ である。つまり、以下のような最適化問題に変形されている。

$$\min_{\{a_i\}} \boldsymbol{p}^T B\boldsymbol{p} \quad \text{with} \quad B = B^T,\ \boldsymbol{p} = \sum_{i=1}^{m} a_i\boldsymbol{e}_i \quad \text{subject to} \quad \sum_{i=1}^{m} a_i^2 = 1 \tag{2.233}$$

ではラグランジュ乗数 λ を用いて、最小化するべき目的関数を以下のように設定しよう。

$$J = \boldsymbol{p}^T B\boldsymbol{p} + \lambda\left(\sum_{i=1}^{m} a_i^2 - 1\right) \tag{2.234}$$

ここで、$\boldsymbol{p}^T B\boldsymbol{p}$ を変形しておこう。そのまえに $U^T\boldsymbol{p}$ を以下のように変形しよう。

$$U^T\boldsymbol{p} = \begin{pmatrix} \boldsymbol{e}_1^T \\ \vdots \\ \boldsymbol{e}_m^T \end{pmatrix} (a_1\boldsymbol{e}_1 + \cdots + a_m\boldsymbol{e}_m) = \begin{pmatrix} a_1 T \\ \vdots \\ a_m^T \end{pmatrix} \tag{2.235}$$

すると、$\boldsymbol{p}^T B\boldsymbol{p}$ が次のようになる。

$$\boldsymbol{p}^T B\boldsymbol{p} = \boldsymbol{p}^T U D U^T\boldsymbol{p} \tag{2.236}$$

$$= (a_1, \ldots, a_m) \begin{pmatrix} v_1 & & \\ & \ddots & \\ & & v_m \end{pmatrix} \begin{pmatrix} a_1 T \\ \vdots \\ a_m^T \end{pmatrix} \tag{2.237}$$

$$= \sum_{i=1}^{m} a_i^2 v_i \tag{2.238}$$

これで、行列形式だった目的関数が全部ばらばらになった。

$$J = \sum_{i=1}^{m} a_i^2 v_i + \lambda \left(\sum_{i=1}^{m} a_i^2 - 1 \right) \tag{2.239}$$

では a_i, λ で微分しよう。

$$\frac{\partial J}{\partial a_i} = 2a_i v_i + 2a_i \lambda = 0 \tag{2.240}$$

$$\therefore\ a_i(v_i + \lambda) = 0 \tag{2.241}$$

$$\frac{\partial J}{\partial \lambda} = \sum_{i=1}^{m} a_i^2 - 1 = 0 \tag{2.242}$$

$$\therefore\ \sum_{i=1}^{m} a_i^2 = 1 \tag{2.243}$$

さて、これはいったいどうなっているのだ？　これを解くには少々トリックが必要である。

まず、$\sum_{i=1}^{m} a_i^2 = 1$ という式から、a_i がすべて 0 であってはまずい。つまり、どれかの a_i は 0 ではないのだ。次に、$a_i(v_i + \lambda) = 0$ という式から、$a_i = 0$ もしくは $\lambda = -v_i$（もしくは両方成立）である。これをよく考えてみよう。

仮にもし、ある i で $\lambda = -v_i$ だったとしよう。このとき、

$$a_i(v_i + \lambda) = a_i(v_i - v_i) = 0 \tag{2.244}$$

を満たすには、a_i は 0 でなくてもよい（0 であってもいいが）。しかし、その他のすべての $j(\neq i)$ において、

$$a_j(v_j + \lambda) = a_j(v_j - v_i) = 0 \tag{2.245}$$

を満たすには、$a_j = 0$ が必要である。なぜなら（一般的に）固有値は異なる、つまり $v_j \neq v_i$ だからである。すると、ある i で $a_i \neq 0$、かつすべての $j(\neq i)$ で $a_j = 0$ のときにも $\sum_{i=1}^{m} a_i^2 = 1$ を満たすためには、$a_i = 1$ でなければならない。

こうして、一つの解が得られた。

$$a_i = 1 \tag{2.246}$$

$$a_j = 0 \quad (j \neq i) \tag{2.247}$$

$$\lambda = -v_i \tag{2.248}$$

$$\min J = v_i \tag{2.249}$$

しかし i は $i = 1, \ldots, m$ である。つまり、m 個の解が得られるのである。

ではどれがいいのだろうか？　もちろん、その m 個の解の中で一番目的関数を小さくするものがよい。つまり、$i = 1$ である。

$$a_1 = 1 \tag{2.250}$$

$$a_j = 0 \quad (j \neq 1) \tag{2.251}$$

$$\lambda = -v_1 \tag{2.252}$$

$$\min J = v_1 \tag{2.253}$$

なぜなら、固有値は小さい順に $v_1 \leq v_2 \leq \cdots \leq v_m$ と並べてあったからである。

さて、何がしたかったのか？　パラメータベクトル $\boldsymbol{p}$ を求めたかったのだ。その解は、もう簡単である。

$$\boldsymbol{p} = \sum_{i=1}^{m} a_i \boldsymbol{e}_i = \boldsymbol{e}_1 \tag{2.254}$$

つまり、最小固有値 v_1 に対応する固有ベクトル $\boldsymbol{e}_1$ である。検算しよう。

$$J = \boldsymbol{p}^T B \boldsymbol{p} = \boldsymbol{e}_1^T U D U^T \boldsymbol{e}_1 = \boldsymbol{e}_1^T U D U^T \boldsymbol{e}_1 = (1, 0, \ldots, 0) D (1, 0, \ldots, 0)^T = v_1 \tag{2.255}$$

たしかに先ほどの解が与える最小値に一致した。

最後にもう一度まとめておこう。$\boldsymbol{p}^B \boldsymbol{p}$ を最小にする $\boldsymbol{p}$ は、B の最小固有値に対応する固有ベクトルである。

線形代数を勉強したときには、固有ベクトルなんか何に役に立つのか分からなかったかもしれない。しかしそれは、いろいろな問題を解くときに便利な強力なツールなのである。

2.4　さらなる話題

直線の推定ができたら、次の問題は曲線の推定である。曲線の中でも楕円などの 2 次曲線は、コンピュータビジョンにおける中心的な問題である。2 次曲線の推定 (conic fitting) については、今後執筆予定である。

関連図書

[1] Peter Meer, Doron Mintz, Azriel Rosenfeld, and Dong Yoon Kim. Robust regression methods for computer vision: A review. *International Journal of Computer Vision*, Vol. 6, No. 1, pp. 59–70, 1991.

[2] Zhengyou Zhang. Parameter estimation techniques: a tutorial with application to conic fitting. *Image and Vision Computing*, Vol. 15, No. 1, pp. 59–76, 1997.