第3章　最小二乗法

3.1　目的

最小二乗法とは，与えられたデータに合う数理モデルあるいはそのパラメータを決定するために，二乗誤差を目的関数とした最小化問題を解いて推定値を得る手法のことである．この手法は最小二乗誤差推定とも呼ばれ，原理は単純ながらも強力で，実験データを扱う多くの場面で標準的に用いられる手法の 1 つである．最小二乗法を単に最適化問題の 1 つとみなすのではなく，最適化問題を構成するデータの取得方法や用いるデータ数，得られた最適化解の評価の仕方を工夫することも重要であり，課題を通して簡単なデータ分析ができるようになってほしい．本実験では，最小二乗法の基礎から逐次的な処理方法，さらに交互最小二乗法までを使えるようになり，データ処理の初歩を学ぶことを目的とする．

本稿では，2 回前期配当科目の全学共通科目「確率論基礎」の知識があることが望ましいが，内容や課題の多くは高校数学の範囲内の確率・統計学の知識でカバーできるように記述した．そのため，回りくどい部分や天下り的な箇所もあるいが，それらを含む最小二乗法の統計的性質や発展的な内容は，3 回前期の「確率と統計」を受講するか，あるいは章末に挙げた参考文献 [1–5] などで自習してもらいたい．

- 課題ではデータをダウンロードする必要がある．PandA にファイルを置くので，適宜アクセスしてダウンロードすること．
- 課題を取り組むためのプログラミング言語は特に指定しないが，C 言語は補助資料を PandA に置いておく．
- 「課題 X」がレポートの対象で,「問題」は演習（必要に応じて行う課題）.

記法

$\mathbb{R}$ は実数全体を表し，n 次の数ベクトル空間を $\mathbb{R}^n$ で表す．また，$m \times n$ の実行列全体を $\mathbb{R}^{m\times n}$ で表す．とくに I_m は m 次の単位行列を表す．行列 A の転置は，$A^\top$ で表す．正方行列 A に対して $A \geq 0$ で半正定値対称行列を，$A > 0$ で正定値対称行列を表す．$\|\bullet\|$ は Euclid ノルムを表す．ただしノルムに下添字をつけることで，別のノルムを表すことがある．$\mathbb{E}$ は適当な確率分布による期待値演算であり，確率分布あるいは確率密度関数を陽に表すときには $\mathbb{E}_P$ と書く．確率変数 $w \in \mathbb{R}^m$ に対し $\mathcal{N}(\mu, V)$ は平均 $\mu \in \mathbb{R}^m$, 共分散行列 $V \in \mathbb{R}^{m\times m}$ の正規分布を表す．

3.2　最小二乗法とその評価方法

次の関係式を満たす方程式が与えられているとする．

$$z = f(\theta, x) \tag{3.1}$$

ここで $z \in \mathbb{R}^m$, $\theta \in \mathbb{R}^n$, $x \in \mathbb{R}^p$ とし，$f : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^m$ とする.「適当に固定されたパラメータ θ の下で，x を入れると z を返す」という意味で，これはパラメータを θ とし，x を入力，z を出力とするシステム f であるという．一般に観測には誤差を伴うため，入力 x や出力 z の値には誤差が乗る．最近の応用で多く見られるように，ここでは出力にのみ観測誤差のある場合を考えよう．入力に観測誤差がある場合も考えられるが，基本的には同じように議論できるため，本項では考えない．

誤差を $w \in \mathbb{R}^m$ とおき，加法的な誤差を考えると，観測値は次で与えられる．

$$y = z + w. \tag{3.2}$$

観測誤差は偶然誤差と系統誤差の 2 種類に分類されるが，本稿では偶然誤差のみを考える．特に断らない限り，本稿では偶然誤差は次の性質を持つものと仮定する．

1. 各試行ごとに観測誤差は同じ確率分布 P から確率的に発生する．

2. 入力 x やパラメータ θ とは独立で，かつ観測ごとに独立に定まる．

3. i 回目の観測に対する誤差を $w_i \in \mathbb{R}^m$ で表す．このとき，次の 2 つが成り立つものとする．

$$\mathbb{E}_P[w_i] = 0,\ \mathbb{E}_P[w_i w_j^\top] = \delta_{i,j} V. \tag{3.3}$$

ここで 0 は m 次のゼロベクトルを表し，V は m 次の正定値対称行列である．$\delta_{i,j}$ は Kronecker のデルタである．

同じ計測機器で計測することを考えるのならば，同じ条件で測定する限り，観測誤差は独立同一分布に従うものとみてもよいであろう．ただし，このような確率分布を正確に知ることは難しい．以降断らない限り，我々はこの観測誤差に関する確率分布は正確に分からないものとし，観測誤差は平均ゼロ，共分散行列が有界に存在することのみを知っているものとする．例として，図 3.1 に $f(\theta, x) = \theta x$，$\theta, x \in \mathbb{R}$ の場合の例を表す．

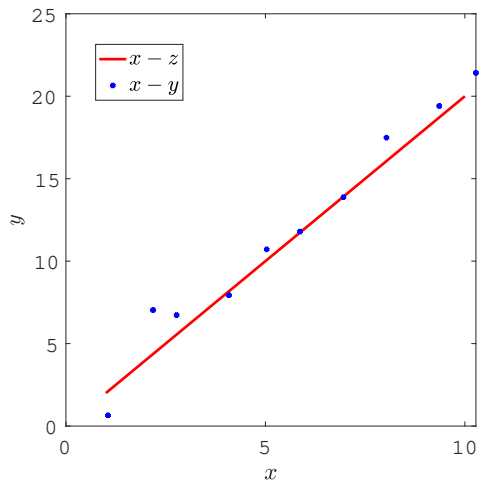

図 3.1: $f(\theta, x) = \theta x$ の場合．青色の点は観測値 y_i で，赤色の線が z_i の値．

ここで関数 f が既知であるが，パラメータ θ が未知である場合を考えよう．このとき，入力と観測値のデータから θ を決定したい．例えば，図 3.1 の青点から赤線を求める問題を考えたい．θ さえ適切に決められれば，未知のデータに対する出力の予測ができるようになり，様々な場面で恩恵を受けることができるからである．N 組の入力および観測値のデータ $\{(x_i, y_i)\}_{i=1}^N$ が与えられているとしよう．このとき，

$$\min_{\theta \in \mathbb{R}^n} \sum_{i=1}^N \|y_i - f(\theta, x_i)\|^2 \tag{3.4}$$

を解くことで，真のパラメータ θ^* を得ることを考える．ここで $\|\bullet\|$ は Euclid ノルムである．このパラメータを求めることを，で f が θ に関して線形である場合を**線形最小二乗法（linear least squares）**あるいは単に**最小二乗法（least mean squares）**，非線形である場合を**非線形最小二乗法（non-linear least squares）**と呼ぶ（least の代わりに minimum が用いられることもある）．一般に最小二乗法は，適切な最適化の手法を使って解けばよい．しかし，以下で見るように，特別な場合は解を陽に書き下せたり，後からデータが追加されても逐次的に解を得られる場合があり，データ容量を抑えられるなどメリットがある．

注意 3.2.1 最小二乗法において誤差の確率「解釈」は，理解のしやすさのために導入されるもので，いつも必ず必要というわけではない．例えば，A 君は四条河原町に，B さんは吉田キャンパスに，C 君は御所にいるとき，それぞれが最短距離で落ち合う場所を求める問題も最小二乗法として解くことができるが，最適な落合場所からそれぞれの現在地が確率的に決まっているかどうかがアルゴリズムや解を決定するわけではない．

注意 3.2.2 実際のデータ分析の現場では，関数 f が既知であるというのも強い仮定であることに注意が必要である．関数 f はデータの統計的性質や物理的考察から決めるか，線形関数などの簡単なものから試していけばよい．深層ニューラルネットワークでは，関数を柔軟に表現できる基底の組み合わせを使うことで f を表現するし，最近では Gauss 過程回帰 [3] と呼ばれる手法も使われる．

3.2.1 回帰問題

まずは θ が f に対して線形である場合を扱おう．このとき，適当な行列値関数 $\varphi : \mathbb{R}^p \to \mathbb{R}^{m \times n}$ が

存在して，

$$f(\theta, x) = \varphi(x)\theta$$

となる場合を考えていることになる．ここで，$\sum_{i=1}^{N} \varphi(x_i)^\top \varphi(x_i)$ が正則ならば，最適化問題 (3.4) の最適解は次で求められる．

$$\hat{\theta}_N = \left(\sum_{i=1}^{N} \varphi(x_i)^\top \varphi(x_i) \right)^{-1} \sum_{j=1}^{N} \varphi(x_j)^\top y_j. \tag{3.5}$$

この $\hat{\theta}_N$ を，**最小二乗誤差推定値（Minimum square error estimate）**あるいは**最小二乗推定値**という[1]．このデータをもらったときの処理の仕方（推定規則）のことを，**最小二乗誤差推定量（minimum square error estimator）**という．また，このとき観測誤差 $w_i, i = 1, \ldots, N$ の共分散行列の推定値は，

$$\hat{V}_N = \frac{1}{N-n} \sum_{i=1}^{N} \left(y_i - \varphi(x_i)\hat{\theta}_N \right) \left(y_i - \varphi(x_i)\hat{\theta}_N \right)^\top \tag{3.6}$$

で与えられる[2]．以降，表記の簡単のために，

$$\Phi_N := \left(\sum_{i=1}^{N} \varphi_i^\top \varphi_i \right)^{-1}, \quad \varphi_i := \varphi(x_i)$$

とおく．

式 (3.5) より推定値 $\hat{\theta}_N$ が得られたからといって，喜んでばかりいてはいけない．推定値が得られることと，その推定精度が良いことは別問題だからである．そこで得られた推定値がどのくらいよいか，検討しよう．真のパラメータ θ と $\hat{\theta}_N$ の間には

$$\begin{aligned} \theta - \hat{\theta}_N =& \Phi_N \left(\Phi_N^{-1}\theta - \sum_{j=1}^{N} \varphi_j^\top y_j \right) \\ =& \Phi_N \sum_{j=1}^{N} \varphi_j^\top \left(\varphi_j \theta - y_j \right) \\ =& \Phi_N \sum_{j=1}^{N} \varphi_j^\top w_j \end{aligned}$$

の関係がある．入力のデータは既知の数であるので，確率的な変動は w_j のみに依存することに注意されたい．したがって，w_j が平均ゼロであることは既知なので，推定が**不偏的（unbiased）**であるという性質

$$\mathbb{E}[\theta - \hat{\theta}_N] = 0$$

が成り立つ．また，もし観測誤差の共分散行列 V が既知であるならば，推定誤差の共分散行列は

$$\mathbb{E}[(\theta - \hat{\theta}_N)(\theta - \hat{\theta}_N)^\top] = \Phi_N \sum_{i=1}^{N} \varphi_i^\top V \varphi_i \Phi_N \tag{3.7}$$

を得る．V の最大特異値 $\sigma_{\max}(V)$ を用いると，

$$\mathbb{E}[(\theta - \hat{\theta}_N)(\theta - \hat{\theta}_N)^\top] \leq \sigma_{\max}(V)\Phi_N \tag{3.8}$$

が成り立つ（不等号の意味は半正定値行列の意味）．とくに $V = \sigma^2 I_n$ ならば，

$$\mathbb{E}[(\theta - \hat{\theta}_N)(\theta - \hat{\theta}_N)^\top] = \sigma^2 \Phi_N$$

となり，Φ_N は N に対して（正定値行列の意味で）単調減少なので，漸近的に推定誤差がゼロに近づくことが示唆される．観測誤差の推定値が分からない場合は，式 (3.7) 右辺の V を $\hat{V}_N$ に置き換えればよい．

問題 1 式 (3.8) を示せ．

回帰推定誤差も求めておこう．元の最小化問題 (3.4) のコスト関数を正規化するため，データ数 N で割ると，

$$R_N := \frac{1}{N} \sum_{i=1}^{N} \left\| y_i - \varphi_i \hat{\theta}_N \right\|^2 \tag{3.9}$$

[1] $\hat{\theta}_N$ を求める問題は，本質的には連立 1 次方程式の解を求める問題と同じになるため，逆行列の計算は Gauss の消去法などを利用して解けばよい．

[2] パラメータ数分をデータから引いて $N - n$ としているのは不偏推定値にするため．$N \gg n$ ならあまり気にしなくてもよい．

とおくと，

$$R_N = \frac{1}{N}\boldsymbol{y}^\top \left(I_{mN} - \boldsymbol{\varphi}(\boldsymbol{\varphi}^\top\boldsymbol{\varphi})^{-1}\boldsymbol{\varphi}^\top\right)\boldsymbol{y} \quad (3.10)$$

$$= \frac{1}{N}\left\|\left(I_{mN} - \boldsymbol{\varphi}(\boldsymbol{\varphi}^\top\boldsymbol{\varphi})^{-1}\boldsymbol{\varphi}^\top\right)\boldsymbol{y}\right\|^2 \quad (3.11)$$

$$= \frac{1}{N}\left\|\left(I_{mN} - \boldsymbol{\varphi}(\boldsymbol{\varphi}^\top\boldsymbol{\varphi})^{-1}\boldsymbol{\varphi}^\top\right)\boldsymbol{w}\right\|^2 \quad (3.12)$$

となる．ただし，

$$\boldsymbol{y} := \begin{bmatrix} y_1^\top & \dots & y_N^\top \end{bmatrix}^\top \in \mathbb{R}^{mN},$$

$$\boldsymbol{w} := \begin{bmatrix} w_1^\top & \dots & w_N^\top \end{bmatrix}^\top \in \mathbb{R}^{mN},$$

$$\boldsymbol{\varphi} := \begin{bmatrix} \varphi_1^\top & \dots & \varphi_N^\top \end{bmatrix}^\top \in \mathbb{R}^{mN\times n}$$

とおいた．Φ_N の正則性は，$\boldsymbol{\varphi}$ が列フルランクであることを意味することに注意されたい．また，式 (3.10) から式 (3.11) への変形には $(I_{mN} - \boldsymbol{\varphi}(\boldsymbol{\varphi}^\top\boldsymbol{\varphi})^{-1}\boldsymbol{\varphi}^\top)$ が射影行列になることを用いた．式 (3.12) から，誤差の二乗平均を表す R_N は観測誤差と入力から定まる行列 $\boldsymbol{\varphi}$ で定まる．もし各ベクトルやデータの数が $mN = n$ を満たす場合，$\boldsymbol{\varphi}$ そのものも逆行列をもつため，$R_N = 0$ になる．R_N の最小値を達成していることになるが，これはよい推定と言えるだろうか？一方，データを増やすと $mN > n$ となり，$R_N > 0$ と一般にはなってしまうが，これは悪い推定値を得ていることになるのだろうか？

推定問題が「データが与えられたときの単なる最適化問題である」と言えないのは，推定問題の本来の目的は入出力関係を適切に定める関数 φ とパラメータ θ を決定することであり，最適化問題はそのための手段だからである．そのため，取得するデータ数や方法を含め，どのように最適化問題を定めれば適切に推定できるかを検討する必要がある．上に挙げた例は，観測誤差に対してパラメータを合わせてしまった**過学習（over learning）**あるいは**過適合（over fitting）**と呼ばれる現象が起こりうる例であり，一般的に歓迎するものではない．それでは，推定性能はどのように評価すればよいだろうか？θ が何かの分布に従って発生しており，雑音の分布も分かっているのならば，R_N の期待値，あるいは $\|\theta - \hat{\theta}_N\|^2$ の期待値を取って評価することができる．しかし，ここで考えている条件下では，確率分布が分からないため，（観測誤差を含む）観測値と答え合わせをするしかない．1 つのやり方が，**交差検証（cross validation）**である（和名よりもクロスバリデーションと呼ばれることが多い）．簡単なやり方は，データを学習用と検証用の 2 つに分け，学習用データでパラメータ推定を行い，検証用データで誤差を評価する方法である．検証用のデータを用いた誤差のことを**テスト誤差（testing error）**という．学習用のデータで得られた推定値が検証用のデータでも良い推定誤差をもつのならば，良い推定ができていると言えるだろう．学習用のデータはできるだけ多くとることは重要であるが，そのために検証用のデータがなくなるのは困るので，多くの場合はそれらの兼ね合いを考えて行う必要がある．学習データがどの程度あれば十分そうかを見積もるには，$\hat{\theta}_N$ がどのくらいのデータ数 N で収束しているかを確認すればよい．例えば，$y = 2x + w$, $w \sim \mathcal{N}(0, 4)$ の場合の推定値のデータ数に対する収束の様子を図 3.2 に示す．また，区間 $[0, 10]$ 内から一様乱数を用いて x_i を発生させデータセットを 10000 組用意し，学習用データ N を増やしたときの観測誤差 R_N の様子と，次で定義される学習用データを除いた残りの $M = 10000 - N$ 個のデータを使った誤差を図 3.3 に示す．

$$R_N^M := \frac{1}{M}\sum_{i=N+1}^{10000} \|y_i - \varphi_i\hat{\theta}_N\|^2,\ M > 0$$

これによると大体 $N = 200$ 程度から $\hat{\theta}_N$ はほとんど変化しなくなるので，この例においては学習用のデータは 200 くらいあればよさそうである．また，この例では $R_1 = 0$ となるが，$\hat{\theta}_1 \simeq 1.2$ なので真値から大きく異なる．N を増やしていくと学習誤差はどこか一定の値に収束していく様子も見られる．また，テスト誤差もだいたい学習誤差と同じくらい値に収束していく様子も見られる．データが十分ある場合はテスト誤差と学習誤差は変わらないが，データが十分あるかどうかわからない場合は両方を併用するとよい．

なお，上で用いた R_N や R_N^M の大きさはどの単位で測定しているかによって決まるので（例えば

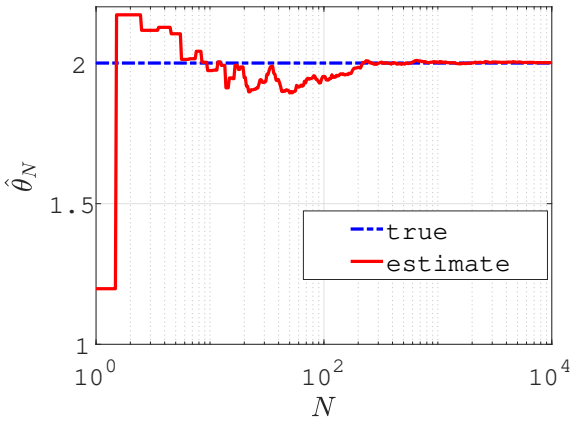

図 3.2: $\hat{\theta}_N$ の収束の様子

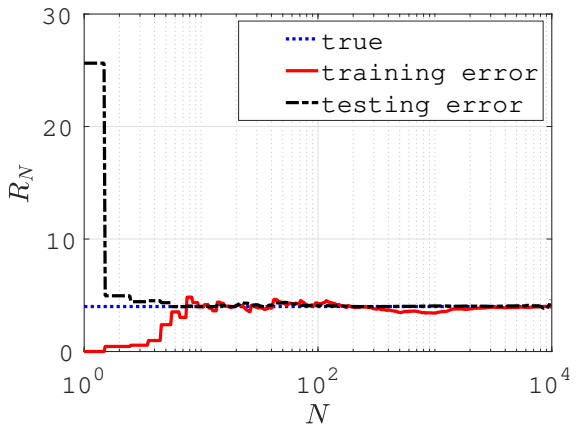

図 3.3: N に対する R_N と R_N^M

m か cm では 100 倍違うので），得られたパラメータがどれほどよくデータを説明するかの指標として用いるには向いてない．観測誤差の確率分布が不明であるならば，**決定変数（coefficient of determination, determination coefficient）**と呼ばれる量が用いられる．

$$C := \frac{\sum_{i=1}^N \|\varphi_i \hat{\theta}_N - \bar{y}\|^2}{\sum_{i=1}^N \|y_i - \bar{y}\|^2} \tag{3.13}$$

ここで $\bar{y} := \frac{1}{N}\sum_{i=1}^N y_i$ とした．最小二乗推定値 $\hat{\theta}_N$ の決定変数は $[0,1]$ の範囲に値を取り，その値が 1 に近いほどよい推定であるとみなされる．逆に 0 に近い場合は得られたパラメータ $\hat{\theta}_N$ が出力 y の推定に役に立たないことを意味する．

なお，観測誤差が正規分布に従うのであれば，$\hat{\theta}_N$ の誤差の評価も様々な統計ツールで分析することができる [2]．以下の 2 つの問題を通して，観測誤差の大きさが異なる場合の最小二乗法を比較検討されたい．

課題 1 (重回帰問題) 各入力 $x_i \in \mathbb{R}^2$, $i = 1, 2, \ldots, 10000$ に対し，次の式で観測データ $y_i \in \mathbb{R}$ が生成されているとする．

$$y_i = x_i^\top \theta + w_i \tag{3.14}$$

ここで観測誤差は $\mathcal{N}(0,1)$ に従って発生している（ただし，平均の情報のみ知っているとし，分散も分布も知らないものとする）．このとき，次の問いに答えよ．

1. θ の最小二乗誤差推定量を，与えられた全てのデータ $\{(x_i, y_i)\}_{i=1}^{10000}$ を使って求めよ．また，そのときの推定誤差共分散行列を求めよ．
2. 図 3.2 のように用いるデータ $\{(x_i, y_i)\}_{i=1}^N$ を増やしたとき，$\hat{\theta}_N$ の各要素が収束していくことを片対数グラフでプロットして確認せよ．ただし，$N = 2, 4, 8, \ldots, 2^{13} = 8192$ のときのみでよい．
3. 全てのデータを用いたときの決定変数を求めよ．

課題 2 (多項式回帰) 各入力 $x_i \in \mathbb{R}$, $i = 1, 2, \ldots, 10000$ に対し，次の式で観測データ $y_i \in \mathbb{R}$ が生成されているとする．

$$y_i = \varphi(x_i)\theta + w_i,\ \varphi(x) := \begin{bmatrix} 1 & x & x^2 & x^3 \end{bmatrix} \tag{3.15}$$

ここで観測誤差は $\mathcal{N}(0,\, 9\,)$ に従って発生している（ただし，平均の情報のみ知っているとし，分散も分布も知らないものとする）．このとき，次の問いに答えよ．

1. θ の最小二乗誤差推定量を，与えられた全てのデータ $\{(x_i, y_i)\}_{i=1}^{10000}$ を使って求めよ．また，そのときの推定誤差共分散行列を求めよ．
2. 図 3.2 のように用いるデータ $\{(x_i, y_i)\}_{i=1}^{N}$ を増やしたとき，$\hat{\theta}_N$ の各要素が収束していくことを片対数グラフでプロットして確認せよ．ただし，$N = 4, 8, \ldots, 2^{13} = 8192$ のときのみでよい．
3. 全てのデータを用いたときの決定変数を求めよ．

また，推定値が収束しない例もあることを次の問題を通して確認されたい．このような問題に最小二乗法を適用することは意味がなく，別の推定手法が必要になる．

課題 3 (分散∞の観測誤差の場合) ここで扱うデータの観測誤差は，各入力 $x_i \in \mathbb{R}^2$, $i = 1, 2, \ldots, 10000$ に対し，次の式でデータ $y_i \in \mathbb{R}$ が生成されているとする．

$$y_i = x_i^\top \theta + w_i \tag{3.16}$$

ただし，観測誤差 w_i は標準 Cauchy 分布に従うとする．このとき，次の問いに答えよ．

1. 与えられた全データ $\{(x_i, y_i)\}_{i=1}^{10000}$ を用いて θ の最小二乗誤差推定量を式 (3.5) から求めよ．
2. 図 3.2 のように用いるデータ $\{(x_i, y_i)\}_{i=1}^{N}$ を増やしたとき，$\hat{\theta}_N$ の各要素が収束していかないことを片対数グラフでプロットして確認せよ．ただし，$N = 2, 4, 8, \ldots, 2^{13} = 8192$ のときのみでよい．

データが求めたい関数を表現するのに十分であるかどうかを確認することも重要である．

課題 4 課題 2 と同じ $\varphi(x)$，θ および観測誤差の分布 $\mathcal{N}(0, 9)$ を考える．また，$x_i \in [0, 1]$, $i = 1, \ldots, 10000$ のときに，データ $\{(x_i, y_i)\}_{i=1}^{10000}$ が与えられているものとする．このとき，$\hat{\theta}_{N=10000}$ を求め，課題 2.1 の $N = 10000$ の結果と比較せよ．また，どのようにデータを取るべきか考察せよ．

注意 3.2.3 データの取り方が悪い場合やデータの数が少ない場合は，$\sum_{i=1}^{N} \varphi_i^\top \varphi_i$ が正則にならない，あるいはそれに近い状況（条件数が悪い）になることがあり，逆行列が求められないことがある．正則にならない正方行列は線形従属な行または列をもつため，統計学では**多重共線性（multicolinear）**と呼ばれる．これは $\hat{\theta}_N$ が求められないことを意味するのではなく，一意に求められないことを意味するので，そのような場合，下記のように**正則化 (regulization)** と呼ばれる項 $\lambda g(\theta)$ を追加して解くことがよく行われる．

$$\min_{\theta \in \mathbb{R}^n} \sum_{i=1}^{N} \|y_i - \varphi_i \theta\|^2 + \lambda g(\theta)$$

ここで λ は正の実数であり，$g(\theta)$ は θ の何らかの（通常非負の）関数である．この正則化は，単に解を得るための手段としてのみでなく，過学習を防ぐ役割や求めたいパラメータの特徴を反映させるために付加することもある．例えば，$g(\theta) = \|\theta\|^2$ とすれば，リッジ回帰と呼ばれる種類の問題になる．また，$g(\theta) = \|\theta\|_{\ell^1}$ と ℓ^1 ノルムを用いれば，Lasso と呼ばれるスパースな解を得るための問題になる [3, 7]．リッジ回帰は逐次最小二乗法の項で用いる．

注意 3.2.4 観測誤差が Cauchy 分布の場合を課題 3 に出したが，最小二乗法の弱点は外れ値（outlier）と呼ばれる大きな誤差に弱い（大きな誤差に合わせるようにパラメータが選ばれてしまう）点である．このような外れ値がデータに混入する場合，最小二乗法を適用する前に，事前にデータ

から外れ値を除いた方がよい．修正トンプソン法など，データから外れ値を自動的に判断して除去する方法はいくつか提案されているので，外れ値の影響が疑われる場合はデータの前処理を行うとよい．

3.2.2 重み付き最小二乗法

次に，最小二乗法における観測誤差に関する仮定を再考する．具体的には，次の 3 つを考える

- 観測誤差の同一分布性
- 観測誤差の独立性
- 観測誤差の共分散行列の情報の有無

これまでは観測誤差は独立同一分布であることや，共分散行列の値が未知であることを仮定していたが，経年劣化したセンサと新品のセンサで同じ対象を計測した場合，その観測誤差は独立ではあっても同じ統計量をもつとはいえない場合も出てくるだろう．データ数 N_1 までは同じ共分散行列 V_1 だったものが，センサを変えたので新たに取れるデータでは共分散行列が $V_2 = 0.1 \times V_1$ になっているかもしれない．仮にどちらかの分散が小さくなることが分かっているのならば（あるいは推定されているのならば），得られるデータの質は同等とはみなせないだろう．このように，データ $\{(x_i, y_i)\}_{i=1}^N$ を得たとき，それぞれのデータの信用度を考えて最小二乗法を解きたい場合もある．このとき，一般には重み行列 $Q_i \geq 0$, $i = 1, 2, \ldots, N$ を用いた次の**重み付き最小二乗法（weighted least squares）**

$$\min_{\theta \in \mathbb{R}^n} \sum_{i=1}^{N} (y_i - \varphi_i \theta)^\top Q_i (y_i - \varphi_i \theta) \tag{3.17}$$

を考えればよい．もちろん，全ての重みがゼロ行列の場合は意味がないので，そのような例は本稿では考えない．この場合，$\sum_{i=1}^N \varphi_i^\top Q_i \varphi_i$ が正則ならば，最適な推定値は

$$\hat{\theta}_N = \left(\sum_{i=1}^{N} \varphi_i^\top Q_i \varphi_i \right)^{-1} \sum_{j=1}^{N} \varphi_j^\top Q_j y_j \tag{3.18}$$

となる．重み付き最小二乗法とは通常，Q_i は対角行列とすることが多いが，本項では重み行列の場合も重み付き最小二乗法と呼ぶ．

観測誤差の共分散行列 V が既知である場合，推定誤差共分散行列 (3.7) は最小の推定誤差共分散行列ではないことが知られている（Gauss–Markov の定理，[2, 4]）．V は正定値対称行列であるので，$V = W^2$ を満たす正定値行列 W が一意に存在する（V の平方根行列）[3]．これを用いると，式 (3.2) は

$$W^{-1} y = W^{-1} \varphi(x) \theta + \underbrace{W^{-1} w}_{=:w'}, \tag{3.19}$$

と書き直すことができる．ここで新たな観測誤差 w_i' は単位行列を分散としてもつ観測誤差である．この解は，

$$\hat{\theta}_N = \left(\sum_{i=1}^{N} \varphi_i^\top W^{-2} \varphi_i \right)^{-1} \sum_{j=1}^{N} \varphi_j^\top W^{-2} y_j \tag{3.20}$$

となるので，$Q_i = aW^{-2} = aV^{-1}$ と選んだものに相当する（$a \in \mathbb{R}$ は正定数）．この観測誤差の統計量を用いて推定値 $\hat{\theta}_N$ を得る上記の手法は，**最良線形不偏推定量（best linear unbiased estimator, BLUE）**と呼ばれる推定量と一致する．「最良」の名前の通り，これは最も推定誤差および訓練誤差の分散を最小にする推定規則である．正しい観測誤差の情報を持っていれば，素朴な定式化から得られる推定規則よりもいい結果が得られることを意味する．

課題 5 $x_i \in \mathbb{R}$, $i = 1, \ldots, 1000$ に対し，観測値の次元が 2 次元となる場合を考える．

$$y_i = \varphi(x_i)\theta + w_i.$$

ここで w_i は $\mathcal{N}(0, V)$ の独立同一分布に従うとし，

$$\varphi = \begin{bmatrix} 1 & x \\ 1 & x^2 \end{bmatrix}, \; V = \begin{bmatrix} 100 & 0 \\ 0 & 1 \end{bmatrix}$$

[3] 一般的に，正則行列 U で $V = UU^\top$ となるものは無数に存在する．本稿では平方根行列を用いるが，例えば Cholesky 分解を使った下三角行列 U を使っても議論は同じくできる．

とする．このとき，式 (3.5) と式 (3.18) をそれぞれ用いて推定値を求め，それぞれの推定誤差共分散を求めよ．ただし，$Q_i = V^{-1}$ とする．また，$\hat{\theta}_N$ の各要素の収束の仕方をプロットせよ．

課題 6 先ほどの課題と同様，$x_i \in \mathbb{R},\ i = 1,\dots,1000$ に対し，観測値の次元が 2 次元となる場合を考える．

$$y_i = \varphi(x_i)\theta + w_i.$$

ここで w_i は，$i = 1,\dots,500$ は $\mathcal{N}(0, V_1)$ の独立同一分布に，$i = 501,\dots,1000$ は $\mathcal{N}(0, V_2)$ の独立同一分布に従うとし，

$$\varphi(x) = \begin{bmatrix} 1 & x \\ 1 & x^2 \end{bmatrix},\ V_1 = \begin{bmatrix} 100 & 0 \\ 0 & 1 \end{bmatrix}, V_2 = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

とする．このとき，式 (3.5) と式 (3.18) のそれぞれで推定値を求め，それぞれの収束の仕方を片対数グラフでプロットせよ．

より一般的には，観測誤差の独立性が仮定できない場合も考えられる．観測誤差が独立でない場合，次のように相関をもつ場合がある．

$$V_{i,j} = \mathbb{E}[w_i w_j^\top]$$

$\boldsymbol{V} := \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top]$ とおくと，重み付き最小二乗法の解法を拡張して，次で得られる推定値が BLUE であることが分かる．

$$\hat{\theta}_N = (\boldsymbol{\varphi}^\top \boldsymbol{V}^{-1} \boldsymbol{\varphi})^{-1} \boldsymbol{\varphi}^\top \boldsymbol{V}^{-1} \boldsymbol{y}^\top \tag{3.21}$$

ここで $\boldsymbol{V}$ はデータ数 N によってはかなり大きなサイズの行列になりえるので，仮にこの行列が分かっていたとしても，この逆行列を求めるのは時間がかかってしまう．そのため，計算時間がかかりすぎる場合や条件数が悪い場合には，何かしら工夫して逆行列を計算する必要がある．

3.3 推定値の合成と逐次最小二乗法

式 (3.5) のように得られている全てのデータをまとめて処理することは，**バッチ処理**と呼ばれる．新しくデータが追加されなければバッチ処理だけ知っておけば十分だが，データ処理の現場では入力と観測値のデータが追加されることはよくあるので，そのたびに最適化問題を解き直すことは煩わしい．とくに計測器からの情報が送られ続ける場合，データを保存するメモリの総量も膨大になるので，別の手段を用いる必要がある．幸いなことに，最小二乗推定値は出力に関して線形な解になる．このことを利用して，ここでは新たにデータが得られたときの処理の仕方について学ぶ．

3.3.1 異なるデータセットからの推定値の合成

データ $\{(x_i, y_i)\}_{i=1}^N$ を得たとき，$f(\theta, x) = \varphi(x)\theta$ の場合の式 (3.4) の問題を解くと，式 (3.5) で解が得られることがわかったが，実際のデータ処理では，別のデータセット $\{(x'_j, y'_j)\}_{j=1}^M$ を用いた推定値 $\hat{\theta}'_M$ が独立に存在する場合もある．それらは，適切な仮定の下で，最小二乗法を用いて以下のように推定値を得ることができる．

$$\hat{\theta}'_M = \underbrace{\left(\sum_{i=1}^M \varphi(x'_i)^\top \varphi(x'_i)\right)^{-1}}_{=:\Phi'_M} \sum_{j=1}^M \varphi(x'_j)^\top y'_j. \tag{3.22}$$

2 組のデータセットからそれぞれの推定値が得られているが，より多くのデータセットから推定する方が推定精度は改善することが期待される．このとき，データを合わせて改めて解き直してもよいが，2 つの推定値を用いて，推定値を更新することができる．便宜的に $\{(x'_j, y'_j)\}_{j=1}^M$ のラベルをつけ直して $\{(x_i, y_i)\}_{i=N+1}^{N+M} = \{(x'_j, y'_j)\}_{j=1}^M$ とす

ると，

$$\begin{aligned}
&\hat{\theta}_{N+M} \\
&= \left(\Phi_N^{-1} + \Phi_M'^{-1}\right)^{-1} \sum_{i=1}^{N+M} \varphi_i^\top y_i \\
&= \left(\Phi_N^{-1} + \Phi_M'^{-1}\right)^{-1} \sum_{i=1}^{N} \varphi_i^\top y_i \\
&\quad + \left(\Phi_N^{-1} + \Phi_M'^{-1}\right)^{-1} \sum_{i=N+1}^{N+M} \varphi_i^\top y_i \\
&= \left(\Phi_N^{-1} + \Phi_M'^{-1}\right)^{-1} \Phi_N^{-1}\Phi_N \sum_{i=1}^{N} \varphi_i^\top y_i \\
&\quad + \left(\Phi_N^{-1} + \Phi_M'^{-1}\right)^{-1} \Phi_M'^{-1}\Phi_M' \sum_{i=N+1}^{N+M} \varphi_i^\top y_i \\
&= \left(\Phi_N^{-1} + \Phi_M'^{-1}\right)^{-1} \left(\Phi_N^{-1}\hat{\theta}_N + \Phi_M'^{-1}\hat{\theta}'_M\right)
\end{aligned}$$

となる．したがって，新たなデータを得ることが想定されるならば，推定値 $\hat{\theta}$ のみでなく Φ_N も保存しておけば，元のデータを保存しておく必要はない[4]．

課題 7 $x_i \in \mathbb{R}, i = 1, \ldots, 10000$ に対し，次の式で観測データ $y_i \in \mathbb{R}$ が生成されているとする．

$$y_i = \varphi(x_i)\theta + w_i \tag{3.23}$$

ここで

$$\varphi(x) := \left[1 \quad \exp\left(-\tfrac{(x_i-1)^2}{2}\right) \quad \exp\left(-(x_i+1)^2\right)\right]$$

であり，w_i は平均 0，分散有限の独立同一分布から生成されているものとする．10000 組のデータの組 $\{(x_i, y_i)\}_{i=1}^{10000}$ のうち，最初の 6000 組と残りの 4000 組のデータそれぞれで $\theta \in \mathbb{R}^3$ の推定値を求め，それらから推定値を合成せよ．また，全データを使った場合の推定値と比較し，一致することを確かめよ．

観測誤差の大きさが異なる場合，重み付きの最小二乗法を使った方が推定結果がよくなることが期待される．データ数がそれぞれ N_1, N_2 のデータセットが 2 組与えられているとき，重み行列 Q_1 と Q_2 を使ったときのそれぞれのデータセットにおける推定値を $\hat{\theta}_{N_1}$，$\hat{\theta}_{N_2}$ とおくと，

$$\begin{aligned}
\hat{\theta}_{N_1+N_2} = &\left(\Phi_{Q_1,N_1}^{-1} + \Phi_{Q_2,N_2}^{-1}\right)^{-1} \\
&\times \left(\Phi_{Q_1,N_1}^{-1}\hat{\theta}_{N_1} + \Phi_{Q_2,N_2}^{-1}\hat{\theta}_{N_2}\right)
\end{aligned} \tag{3.24}$$

となる．ここで

$$\Phi_{Q,N} := \left(\sum_{i=1}^{N} \varphi_i^\top Q \varphi_i\right)^{-1}$$

とした．次の問題もそれほど難しくないので，時間があるときに考えてもらいたい．

問題 2 式 (3.24) が成り立つことを示せ．

課題 8 $x_i \in \mathbb{R}, i = 1, \ldots, 10000$ に対し，次の式でデータ $y_i \in \mathbb{R}$ が生成されているとする．

$$y_i = \varphi(x_i)\theta + w_i \tag{3.25}$$

ここで w_i は最初の 6000 組と残りの 4000 組のデータで，平均ゼロの異なる独立な分布に従うものとする（$i = 1, \ldots, 6000$ では独立同一分布，$i = 6001, \ldots, 10000$ では異なる独立同一分布）．

$$\varphi(x) := \left[1 \quad \exp\left(-\tfrac{(x_i-1)^2}{2}\right) \quad \exp\left(-(x_i+1)^2\right)\right]$$

である．10000 組のデータの組 $\{(x_i, y_i)\}_{i=1}^{10000}$ のうち，最初の 6000 組と残りの 4000 組のデータそれぞれで $\theta \in \mathbb{R}^3$ の推定値と偶然誤差 w_i の分散の推定値を求め，推定された分散を用いて推定値を合成せよ．また，全データを使い，式 (3.5) を用いて得た $\hat{\theta}_{10000}$ と，どちらが優れているか比較考察せよ（比較の仕方を考えること）．

3.3.2 逐次最小二乗法

これまでの考え方を応用して，最小二乗法を実時間処理で解くことを考えよう．具体的には，N 個のデータから推定された $\hat{\theta}_N$ と新たなデータ

[4] あくまで最小二乗法で行う場合の結果であり，他の評価基準を用いる場合は全データを用いる必要が生じることがほとんどである．

(x_{N+1}, y_{N+1}) を用いて再帰的に $\hat{\theta}_{N+1}$ を求めることを考える．最小二乗誤差推定 (3.5) より，次のように式変形できる．

$$\begin{aligned}
&\hat{\theta}_{N+1}\\
=&\Phi_{N+1}\sum_{j=1}^{N+1}\varphi_j^\top y_j\\
=&\left(\Phi_N^{-1}+\varphi_{N+1}^\top\varphi_{N+1}\right)^{-1}\\
&\times\left(\varphi_{N+1}^\top y_{N+1}+\sum_{j=1}^{N}\varphi_j^\top y_j\right)\\
=&\Big\{\Phi_N-\Phi_N\varphi_{N+1}^\top\\
&\times\left(I_m+\varphi_{N+1}\Phi_N\varphi_{N+1}^\top\right)^{-1}\varphi_{N+1}\Phi_N\Big\}\\
&\times\left(\varphi_{N+1}^\top y_{N+1}+\sum_{j=1}^{N}\varphi_j^\top y_j\right)\\
=&\hat{\theta}_N+K_{N+1}\left(y_{N+1}-\varphi_{N+1}\hat{\theta}_N\right).
\end{aligned}$$

ただし

$$K_{N+1}:=\Phi_N\varphi_{N+1}^\top\left(I_m+\varphi_{N+1}\Phi_N\varphi_{N+1}^\top\right)^{-1}.$$

ここで式変形に逆行列補題[5]を用いて，

$$\begin{aligned}
&\Phi_{N+1}\\
=&\Phi_N-\Phi_N\varphi_{N+1}^\top\\
&\times\left(I_m+\varphi_{N+1}\Phi_N\varphi_{N+1}^\top\right)^{-1}\varphi_{N+1}\Phi_N
\end{aligned}$$

を得た．したがって，新たなデータ (x_{N+1}, y_{N+1}) が得られるたびに，新たな推定値 $\hat{\theta}_{N+1}$ はそれまでのデータによる推定値 $\hat{\theta}_N$ と Φ_N を用いることで更新できるので，改めてバッチ処理をする必要がない．オンラインで行う場合，重要なのは Φ_0 の選び方である．Φ_N の定義から，Φ_0 は定義できない．そこで Φ_N そのものを用いずに，小さな正の実数 ε を用いて $\tilde{\Phi}_0=\frac{1}{\varepsilon}I_n$ とし，次のように更新すればよい．

$$\tilde{\theta}_{N+1}=\tilde{\theta}_N+\tilde{K}_{N+1}\left(y_{N+1}-\varphi_{N+1}\tilde{\theta}_N\right), \tag{3.26}$$

$$\tilde{K}_{N+1}=\tilde{\Phi}_N\varphi_{N+1}^\top\left(I_m+\varphi_{N+1}\tilde{\Phi}_N\varphi_{N+1}^\top\right)^{-1} \tag{3.27}$$

$$\tilde{\Phi}_{N+1}=\tilde{\Phi}_N-\tilde{K}_{N+1}\varphi_{N+1}\tilde{\Phi}_N \tag{3.28}$$

ただし，

$$\tilde{\theta}_0=0,\quad \tilde{\Phi}_0=\frac{1}{\varepsilon}I_n$$

とした．

ここでこの $\tilde{\Phi}_N$ と Φ_N の関係は，Φ_N が正則になる N に対して，

$$\tilde{\Phi}_N^{-1}=\Phi_N^{-1}+\varepsilon I_n=\sum_{i=1}^{N}\varphi_i^\top\varphi_i+\varepsilon I_n$$

であることに注意されたい．Φ_N^{-1} はデータを得るたびに正定値対称行列の意味で大きくなるため，$\tilde{\Phi}_N$ と Φ_N を Taylor 展開すると，

$$\tilde{\Phi}_N=\Phi_N-\varepsilon\Phi_N^2+O(\varepsilon^2)$$

であり，適切なデータ数 N が増えるにつれて Φ_N は正定値対称行列の意味で小さくなるので，ε の 1 次の項がゼロに近づくことが期待される．実際に ε の 1 次以上の項は N が増えると小さくなる項である．以下では，添字に時間の順序関係がある場合には k を用いる．

課題 9 (システム同定) 次のバネ・マス・ダンパ系を表す直線上の運動方程式を考える．

$$M\frac{d^2}{dt^2}y(t)+D\frac{d}{dt}y(t)+Ky(t)=F(t)$$

ここで M, K, D は正定数で，$F(t)$ は外力である．$y(t)$ は観測でき，$F(t)$ はこちらで設計できるものとする．このとき，得られるデータから (M, K, D) を決定したい．微小な δt では

$$\begin{aligned}
\frac{d}{dt}y(t)&\simeq\frac{y((k+1)\delta t)-y(k\delta t)}{\delta t},\\
\frac{d^2}{dt^2}y(t)&\simeq\frac{y((k+2)\delta t)-2y((k+1)\delta t)+y(k\delta t)}{\delta t^2}
\end{aligned}$$

[5] $A\in\mathbb{R}^{n\times n}$ を正則な行列とするとき，$(A+BC)^{-1}=A^{-1}-A^{-1}B(I+CA^{-1}B)^{-1}CA^{-1}$．ただし，$B$ と C は $BC\in\mathbb{R}^{n\times n}$ となる任意のサイズの行列．

と近似できることを利用し，$y_k := y(k\delta t)$，$F_k := F(k\delta t)$ とすると，

$$\begin{aligned} y_k = &\left(2 - \frac{D}{M}\delta t\right) y_{k-1} \\ &- \left(1 - \frac{D}{M}\delta t + \frac{K}{M}\delta t^2\right) y_{k-2} \\ &+ \frac{\delta t^2}{M} F_{k-2} + w_k \end{aligned}$$

を得る．ここで w_k は近似の際に生じる誤差である．したがって，問題は

$$y_k = x_k^\top \theta + w_k,\ x_k := \begin{bmatrix} y_{k-1} \\ y_{k-2} \\ F_{k-2} \end{bmatrix}$$

の θ を求める問題にすることができる．パラメータの真値を $(M, D, K) = (2, 1, 3)$ とし，$\delta t = 0.01$，w_k は $[-1, 1]$ の一様分布に従う各時刻で独立な確率変数であるとして，以下の問いに答えよ．

1. $y_{-2} = 0$, $y_{-1} = 0$ とし，$F_k = 1$, $k = -2, -1, 0, 1, 2, \ldots$ として一定の外力を加えたとき，パラメータ θ を逐次最小二乗法で求めよ．ただし，パラメータの初期値は 0 とし，$k = 10000$ までよい．また正則化項は工夫せよ．

2. $y_{-2} = 0$, $y_{-1} = 0$ とし，$F_k = \sin(\pi k/5)$, $k = -2, -1, 0, 1, 2, \ldots$ として一定の外力を加えたとき，パラメータ θ を逐次最小二乗法で求めよ．ただし，パラメータの初期値は 0 とし，$k = 10000$ までよい．また正則化項は工夫せよ．

3. $y_0 = 0$, $y_1 = 0$ とする．このとき，どのような外力を加えればパラメータの推定がうまくいくか考え実装し，逐次最小二乗法で求めよ．ただし，パラメータの初期値は 0 とし，$k = 10000$ までよい．

重み付き逐次最小二乗法の更新アルゴリズムも，逐次最小二乗法と同様の計算で求められる．とくに重み行列が単位行列の定数倍（$Q_i = \rho_i I$）のとき，次のように計算される．

$$\begin{aligned} \hat{\theta}_{N+1} =& \hat{\theta}_N + K_{N+1}(y_{N+1} - \varphi_{N+1}\hat{\theta}_N), \\ K_{N+1} =& \rho_{N+1}\Phi_N\varphi_{N+1}^\top \\ &\times \left(I_m + \rho_{N+1}\varphi_{N+1}\Phi_N\varphi_{N+1}^\top\right)^{-1}, \\ \Phi_{N+1} =& \Phi_N - K_{N+1}\varphi_{N+1}\Phi_N. \end{aligned}$$

とくに時系列データの場合には，古いデータの影響を小さく，新しいデータの影響を大きく使いたい場面が多い．そこで N 組のデータがあるときに $\rho_i = \gamma^{N-i}$，$\gamma \in (0, 1)$ とおくと，その推定値は

$$\hat{\theta}_{\gamma,N+1} = \left(\sum_{i=1}^{N} \gamma^{N-i}\varphi_i^\top\varphi_i\right)^{-1} \sum_{j=1}^{N} \gamma^{N-j}\varphi_j^\top y_j$$

となる．同様に $N+1$ 組のデータでは

$$\begin{aligned} \hat{\theta}_{\gamma,N+1} = &\left(\varphi_{N+1}^\top\varphi_{N+1} + \gamma\sum_{i=1}^{N}\gamma^{N-i}\varphi_i^\top\varphi_i\right)^{-1} \\ &\times \left(\varphi_{N+1}^\top y_{N+1} + \gamma\sum_{j=1}^{N}\gamma^{N-j}\varphi_j^\top y_j\right) \end{aligned}$$

となるので，これを整理すると次の更新アルゴリズムを得る．

$$\hat{\theta}_{\gamma,N+1} = \hat{\theta}_{\gamma,N} + K_{\gamma,N+1}(y_{N+1} - \varphi_{N+1}\hat{\theta}_{\gamma,N}) \tag{3.29}$$

$$K_{\gamma,N+1} = \Phi_{\gamma,N}\varphi_{N+1}^\top \left(\gamma I_m + \varphi_{N+1}\Phi_{\gamma,N}\varphi_{N+1}^\top\right)^{-1} \tag{3.30}$$

$$\Phi_{\gamma,N+1} = \frac{1}{\gamma}\Phi_{\gamma,N} - \frac{1}{\gamma}K_{\gamma,N+1}\varphi_{N+1}\Phi_{\gamma,N} \tag{3.31}$$

この γ を**忘却係数（forgetting factor）**と呼び，古いデータに指数的な重みをつけ，新しいデータの影響をパラメータ推定に強く反映させる効果を与える．

課題 10 (非定常時系列データの推定) 次の正弦波でゆっくりと変動する信号を考える．

$$y_k = \sin(0.0001k) + w_k,\ k = 1, \ldots, 10000.$$

ただし，w_k は -1 と $+1$ を確率 $\frac{1}{2}$ で出す Bernoulli 過程であるとし，$\theta_k := \sin(0.0001k)$ は未知とす

る．このとき，$\gamma = 0.99$ の忘却係数付きの逐次最小二乗法で θ_k を推定し，各時刻の $\hat{\theta}_{\gamma,k}$ をプロットせよ．

3.4 Kalman フィルタと Kalman スムーザ

重み付け最小二乗法の項で述べたとおり，偶然誤差の確率分布を知っている場合（統計的な性質を知っている場合），最小二乗誤差推定よりも**最小平均二乗誤差推定（minimum mean square error estimation）**と呼ばれる手法を使った方が推定精度が向上する[6]．議論を簡単にするため，N 個のデータ $\mathcal{D}_N := \{(x_i, y_i)\}_{i=1}^N$ の関数として，入力 x を入れたときの出力の推定値を $\hat{y}(\mathcal{D}_N, x)$ と表す．また，入力 x のときの出力を $y(x)$ で表す．このとき，入力 x に対する最小平均二乗誤差推定は，次の「関数 $\hat{y}$」に関する最適化問題の解になる（期待値の存在は仮定する）．

$$\min_{\hat{y}} \mathbb{E}_P \left[|y(x) - \hat{y}(\mathcal{D}_N, x)|^2 \right] \tag{3.32}$$

期待値を用いることで未知のデータに対しても二乗誤差の統計量を計算することができ，それを最小にするように推定規則を作成することができる．データとして用いている $\{y_i\}_{i=1}^N$ も確率変数であることに注意されたい．また，$\hat{y}$ は x にも依存する関数であるが，以下では $x_i = i$ と時間を表すパラメータだと思うことにする[7]．この問題 (3.32) の解は，条件付き期待値になることが知られている．条件付き期待値の厳密な定義や性質は「確率と統計」の授業や統計学の文献で勉強していただきたい．とくに $y(x) = \varphi(x)\theta + w$ と表される場合，最適化問題 (3.32) の解は θ の推定値を求める BLUE に等しい．そこで，重み行列を使った逐次最小二乗法の考え方を応用してみよう．とくにここでは線形なダイナミカルシステムの**潜在変数（hidden variable）**の推定問題を考える．

3.4.1 Kalman フィルタ

次の線形差分方程式を考える．

$$\theta_k = a_k \theta_{k-1} + v_k, \tag{3.33}$$

$$y_k = c_k \theta_k + w_k,\ k = 1, 2, \ldots \tag{3.34}$$

ここで $a_k, c_k \in \mathbb{R}$ であり，$v_k, w_k \in \mathbb{R}$ は互いに独立な平均ゼロ，分散がそれぞれ σ_v^2, σ_w^2 の Gauss 分布に従って発生し，さらに異なる時刻で独立である（白色性をもつ）とする[8]．また，$c_k \neq 0$, $k = 1, 2, \ldots$ とする．θ_0 は平均 $\bar{\theta}$, 分散 P を持つ分布に従って発生するとする（Gauss 性は不要）．ここで v_k や w_k および初期値 θ_0 は観測できないが，それらの “分布”の情報は既知であり，a_k, c_k も既知であるとしよう．観測値 y_k もセンサから各時刻で得られるものとするが，θ_k は潜在変数で直接知ることはできないとする．この節で扱う内容は，各時刻 k までの観測値 $\{y_\ell\}_{\ell=1}^k$ を用いて θ_k を推定することである．これまで扱ってきた問題は推定すべきパラメータが確定的だったが，ここでは θ_k も確率変数になっていることに注意されたい．

時刻 k までのデータによる推定値を $\hat{\theta}_k \in \mathbb{R}$ で表すとする．式 (3.26) の逐次最小二乗法では，1 ステップ前の推定値に，新たなデータ $(k, y_k) \equiv y_k$ による修正項が加えられる．式 (3.33) と (3.34) より，

$$y_k = c_k(a_k \theta_{k-1} + v_k) + w_k$$

となるため，前の時刻の推定値 $\hat{\theta}_{k-1}$ を a_k 倍したものに新たなデータから修正する項を考え，次の規則で推定値を更新することを考える．

$$\begin{aligned}\hat{\theta}_k =& a_k \hat{\theta}_{k-1} + F_k(y_k - c_k a_k \hat{\theta}_{k-1}) \\ =& a_k \hat{\theta}_{k-1} + a_k c_k F_k(\theta_{k-1} - \hat{\theta}_{k-1}) \\ &+ F_k(w_k + c_k v_k)\end{aligned}$$

[6] least mean square error (LMSE) と minimum mean square error (MMSE) は，異なる意味で用いられるので注意．LMSE は二乗誤差の「標本平均」に対する推定を意味し，MMSE は二乗誤差の「期待値」に対する推定を意味することが多いようである．

[7] そうでない場合は確率場の推定になり，Gauss 過程回帰などで用いられる．その場合，x の依存性を消すために x に関する積分や $\sup_x$ などを期待値の内側か外側につける．

[8] Gauss 性は本来は不要だが議論を簡単にするため．

ここで $F_k \in \mathbb{R}$ は自由に設計できるパラメータである．このとき，推定誤差 $\theta_k - \hat{\theta}_k$ を最小二乗平均の意味で最小化する推定器を設計したい．推定誤差分散を $V_k := \mathbb{E}[(\theta_k - \hat{\theta}_k)^2]$ とし，各時刻 $k \geq 1$ で V_k を最小にするパラメータ F_k を求めると，

$$\begin{aligned} V_k =& a_k^2(1 - c_k F_k)^2 V_{k-1} \\ &+ (1 - c_k F_k)^2 \sigma_v^2 + \sigma_w^2 F_k^2 \\ =& (a_k^2 c_k^2 V_{k-1} + c_k^2 \sigma_v^2 + \sigma_w^2) \\ &\times \left(F_k - \frac{a_k^2 c_k V_{k-1} + c_k \sigma_v^2}{a_k^2 c_k^2 V_{k-1} + c_k^2 \sigma_v^2 + \sigma_w^2} \right)^2 \\ &+ a_k^2 V_{k-1} + \sigma_v^2 - \frac{(a_k^2 c_k V_{k-1} + c_k \sigma_v^2)^2}{a_k^2 c_k^2 V_{k-1} + c_k^2 \sigma_v^2 + \sigma_w^2} \end{aligned}$$

より，

$$F_k = \frac{a_k^2 c_k V_{k-1} + c_k \sigma_v^2}{a_k^2 c_k^2 V_{k-1} + c_k^2 \sigma_v^2 + \sigma_w^2} \tag{3.35}$$

を得る．見やすくなるように変数 X_k を入れて整理すると，$\hat{\theta}_0 = \bar{\theta}, V_0 = P$ として，次の推定則を導くことができる．

$$\hat{\theta}_k = a_k \hat{\theta}_{k-1} + F_k(y_k - c_k a_k \hat{\theta}_{k-1}), \tag{3.36}$$

$$X_k = a_k^2 V_{k-1} + \sigma_v^2, \tag{3.37}$$

$$V_k = X_k - \frac{c_k^2 X_k^2}{c_k^2 X_k + \sigma_w^2} = \frac{\sigma_w^2 X_k}{c_k^2 X_k + \sigma_w^2}, \tag{3.38}$$

$$F_k = \frac{c_k X_k}{c_k^2 X_k + \sigma_w^2}. \tag{3.39}$$

この推定則は **Kalman フィルタ（Kalman filter）**と呼ばれ，システム制御や時系列解析，信号処理等で広く用いられる．θ_k, y_k が多次元であっても，行列やベクトルの平方完成を使って同様に求められる．Kalman フィルタは推定誤差分散を最小にする推定器であるが，これは定常誤差の有界性を保証するものではないことにも注意されたい．

課題 11 (Kalman フィルタ) 次の離散時間ダイナミクスおよび観測方程式が与えられているとする．

$$\theta_k = 0.9\theta_{k-1} + v_k, \tag{3.40}$$

$$y_k = 2\theta_k + w_k \tag{3.41}$$

ここで $\theta_k, y_k \in \mathbb{R}$ であり，v_k, w_k は互いに独立な $\mathcal{N}(0,1)$ に従う確率変数である．また，θ_0 は平均 3, 分散 2 をもつ確率分布に従うとする．このとき，$\{y_\ell\}_{\ell=1}^k$ までを使った θ_k, $k = 1, \ldots 100$ の推定値 $\hat{\theta}_k$ を求め，θ_k と共に時間変化をプロットせよ．

Kalman フィルタの Bayes 的な導出や応用例は，文献 [8, 9] とその参考文献を参照されたい．また，確率制御問題へどのように応用されるかは文献 [10] を参照されたい．

3.4.2 Kalman スムーザ

Kalman フィルタは，新たにデータ y_k を得るたびに，新たな時刻の潜在変数 θ_k の推定を行うアルゴリズムである．$\alpha_i := a_i(1 - c_i F_i)$ として $\hat{\theta}_k$, $k = 1, \ldots, N$ を改めて書き直すと，

$$\begin{aligned} \hat{\theta}_k =& \alpha_k \hat{\theta}_{k-1} + F_k y_k \\ =& \alpha_k \alpha_{k-1} \hat{\theta}_{k-2} + \alpha_k F_{k-1} y_{k-1} + F_k y_k \\ =& \prod_{i=1}^{k} \alpha_i \bar{\theta} + \sum_{\ell=1}^{k-1} \left(\prod_{j=\ell+1}^{k} \alpha_j \right) F_\ell y_\ell + F_k y_k \end{aligned}$$

となるので，次のように書き直すことができる．

$$\begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \vdots \\ \hat{\theta}_N \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_1 \alpha_2 \\ \vdots \\ \prod_{i=1}^N \alpha_i \end{bmatrix} \bar{\theta} + M_N \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

$$M_N := \begin{bmatrix} F_1 & 0 & \cdots & 0 \\ \alpha_2 F_1 & F_2 & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ \prod_{2=1}^N \alpha_i F_1 & \prod_{3=1}^N \alpha_i F_2 & \cdots & F_N \end{bmatrix}$$

M_N が下三角行列になることが，各時刻 k で k までのデータ $\{y_\ell\}_{\ell=1}^k$ を使っていることと等価になる．一方，差分方程式 (3.33) の解 θ_k は，過去の情報 $\theta_{k-\ell}$, $\ell = 1, \cdots$ に影響を受けているため，

$$\begin{aligned} y_k =& c_k \theta_k + w_k = c_k(a_k \theta_{k-1} + v_k) + w_k \\ =& c_k \prod_{j=i+1}^{k} a_j \theta_i + (\text{noise terms}) \end{aligned}$$

より，y_k には過去の θ に関する情報も含まれている．そのため，新たなデータ y_k から推定値 $\hat{\theta}_\ell$, $\ell < k$ を更新することもできるはずである．これを可能にするのが **Kalman スムーザ (Kalman smoother)** であり，いくつかアルゴリズムは知られているが，ここではよく用いられる **Rauch–Tung–Striebel アルゴリズム**と呼ばれるオフラインの手法を紹介する（オンラインのアルゴリズムもあるが，あまり使われない）．

時刻 1 から N までのデータ $\{y_k\}_{k=1}^N$ から Kalman フィルタによる推定値 $\{\hat{\theta}_k\}_{k=1}^N$, 推定誤差分散 $\{V_k\}_{k=1}^N$, 一段先予測誤差分散 $\{X_k\}_{k=1}^N$ が得られているものとする．時刻 k の潜在変数 θ_k を，全てのデータを用いた推定を $\hat{\theta}_k^s$ で表す．このとき，終端条件 $\hat{\theta}_N^s = \hat{\theta}_N$, $V_N^s = V_N$ とした後退方程式

$$\hat{\theta}_k^s = \hat{\theta}_k + g_k(\hat{\theta}_{k+1}^s - a_k\hat{\theta}_k) \tag{3.42}$$

$$g_k = a_k\frac{V_k}{X_{k+1}} \tag{3.43}$$

$$V_k^s = V_k + g_k^2(V_{k+1}^s - X_{k+1}) \tag{3.44}$$

の解 $\hat{\theta}_k^s$, $k = 0, \dots, N$ は，時刻 N までの全ての情報を用いた最小平均二乗誤差推定値である．導出に関しては文献 [9, 10] を参照されたい．また，V_k^s は推定誤差分散 $\mathbb{E}[(r_k - \hat{r}_k^s)^2]$ を表す．

課題 12 (Kalman スムーザ) 課題 11 の設定の元で，初期値 θ_0 の推定値とその推定誤差分散を求めよ．また，初期値の分散と比べて式 (3.42)–(3.44) からなる RTS アルゴリズムによる推定誤差分散がどれほど改善できたか，推定誤差分散と初期分散の比（分母が初期分散）で表せ．

3.5 交互最小二乗法

非線形最小二乗法は，問題によって様々な解き方があるので，一般論は存在しない．ここでは，機械学習などで多く用いられる K–**平均法（K-means clustering）**（あるいは K–平均クラスタリング）を題材に，非線形最小二乗法の一種である**交互最小二乗法（Alternating least squares）**を取り上げる．

3.5.1 交互最小二乗法

関数 $f(\theta, x)$ が $\theta = (\alpha, \beta)^\top$ に関して双線形であるとは，例えば

$$f(\theta, x) = \alpha\beta x + \beta$$

のように，推定したいパラメータ同士が掛け算の形になってしまっているというものである．$\alpha\beta$ を改めて別の変数に置き換えて推定してもよいが，それぞれのパラメータに意味のある場合もある．このような関数 $f(\theta, x)$ は，パラメータ $\theta \in \mathbb{R}^n$ を 2 つに分割すれば，分割したパラメータ $\alpha \in \mathbb{R}^{n_1}$, $\beta \in \mathbb{R}^{n_2}$, $n_1 + n_2 = n$ それぞれに対して線形になる．そこで，$f(\theta, x)$ を改めて $g(\alpha, \beta, x)$ とおいて，次のようにパラメータを求める．

Step 1 初期値 $\hat{\alpha}_0$, $\hat{\beta}_0$ を設定する．

Step 2 $j = 1, \dots$ に対し，

Step 2-1 $\beta = \hat{\beta}_{j-1}$ に固定し，α を最小化する

$$\hat{\alpha}_j = \operatorname*{arg\,min}_{\alpha\in\mathbb{R}^{n_1}} \sum_{i=1}^N \|y_i - g(\alpha, \hat{\beta}_{j-1}, x_i)\|^2$$

Step 2-2 $j = 1, \dots$ に対し，$\alpha = \hat{\alpha}_j$ を固定し，β を最小化する．

$$\hat{\beta}_j = \operatorname*{arg\,min}_{\beta\in\mathbb{R}^{n_2}} \sum_{i=1}^N \|y_i - g(\hat{\alpha}_j, \beta, x_i)\|^2$$

Step 3 収束判定条件（例えば事前に設定した $\varepsilon > 0$ に対して $(\|\alpha_{j-1} - \alpha_j\|^2 + \|\beta_{j-1} - \beta_j\|^2) < \varepsilon$）を満たすか判定し，満たさなければ j を 1 つ繰り上げて Step 2 に戻る．

同じデータを使ったまま，最小二乗問題を繰り返し解く問題になる．各最小化問題は，式 (3.5) の形で解けばよい．交互に最小化問題を解くため，交互最小二乗法と呼ばれる．アルゴリズムの性質から，$j = 1, \dots$ に対して

$$\sum_{i=1}^N \|y_i - g(\hat{\alpha}_j, \hat{\beta}_j, x_i)\|^2$$
$$\geq \sum_{i=1}^N \|y_i - g(\hat{\alpha}_{j+1}, \hat{\beta}_{j+1}, x_i)\|^2$$

なる関係も成り立つ．したがって交互最小二乗法は「悪くはならない」手法である．しかし，一般には局所的最適解への収束は必ずしも保証されるとは限らないため，初期値の工夫などが必要である．

課題 13 $x_i \in [-1,1]$, $i = 1,\dots,10000$, に対して，次の方程式で定められる入出力データが与えられているとする．

$$y_i = \alpha^\top \varphi(x_i)\beta + w_i.$$

ここで $\alpha \in \mathbb{R}^2$, $\beta \in \mathbb{R}^3$，w_i は区間 $[-1,1]$ の一様分布であるとし，

$$\varphi(x) = \begin{bmatrix} 1 & x & x^2 \\ x & x^2 & x^3 \end{bmatrix}$$

の場合を考える．10000 組の $\{(x_i, y_i)\}_{i=1}^{10000}$ から，交互最小二乗法を用いて α, β を推定せよ．このとき，10 組の異なる初期値を用意し，最もよかったパラメータとそのときに使った初期値を示せ．

異なる複数の初期値を使って最適化問題を解く方法をマルチスタート法あるいは多点スタート法という．

3.5.2 K-平均法

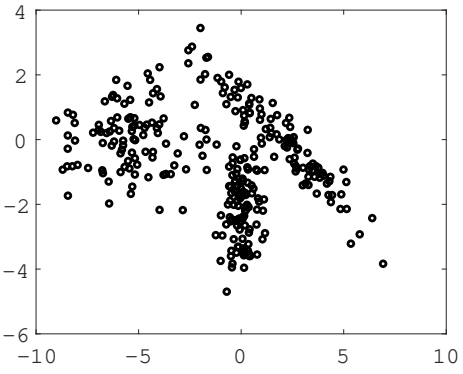

図 3.4: 観測されたデータ

K–平均法は，与えられたデータを K 個のクラスタに分類する手法であり，似たようなデータを集める代表的な方法である [4, 11]．例えば，図 3.4 は元々別々の分布から発生したデータが混在しており，元のデータは図 3.5 のように色分けされた 3 種類のデータからなる．$\{x_i\}_{i=1}^N \subset \mathbb{R}^p$ がデータとして与えられているものとする．これを K 個のクラスタに分類したい．ここで K はこちらで与える正整数とする．K をどのように決めるかの議論はここでは行わず，与えられているものとする．$\mathcal{V} := \{1,\dots,N\}$ とし，$\mathcal{V}$ の空でない部分集合 $\mathcal{V}_1,\dots,\mathcal{V}_K$ を，$\cup_{\ell=1}^K \mathcal{V}_\ell = \mathcal{V}$ かつ $\mathcal{V}_\ell \cap \mathcal{V}_k = \emptyset$, $\ell \neq k$ を満たすようにとる．この $\mathcal{V}_\ell$ がクラスタを表し，そのクラスタを特徴づける量を次で定める．

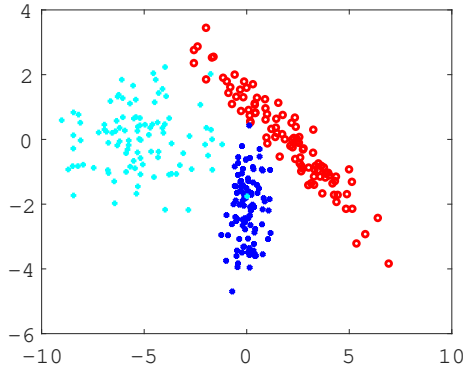

図 3.5: それぞれのデータ点を生成分布別に表した図

$$\mu(\mathcal{V}_\ell) := \arg\min_{\mu \in \mathbb{R}^p} \sum_{i_\ell \in \mathcal{V}_\ell} \|x_{i_\ell} - \mu\|^2, \ \ell = 1,\dots,K \tag{3.45}$$

これは，次のようにクラスタの中心を表す．

$$\mu(\mathcal{V}_\ell) = \frac{1}{|\mathcal{V}_\ell|} \sum_{i_\ell \in \mathcal{V}_\ell} x_{i_\ell}. \tag{3.46}$$

ただし，$|\mathcal{V}_\ell|$ は $\mathcal{V}_\ell$ の要素数を表す．問題となるのはどのようにクラスタを選ぶかであるが，これは次の最小化問題を考えればよい．

$$\min_{\{\mathcal{V}_1,\dots,\mathcal{V}_K\}} \sum_{\ell=1}^K \sum_{i_\ell \in \mathcal{V}_\ell} \|x_{i_\ell} - \mu(\mathcal{V}_\ell)\|^2. \tag{3.47}$$

ここで

$$r_{i,\ell} := \begin{cases} 1, & i \in \mathcal{V}_\ell \\ 0, & i \notin \mathcal{V}_\ell \end{cases}$$

を導入すれば，式 (3.47) は

$$\sum_{\ell=1}^{K}\sum_{i=1}^{n} r_{i,\ell}\|x_i-\mu(\mathcal{V}_\ell)\|^2$$

となる．x_i がどのクラスタに入るかを更新するには，$\mathcal{V}_\ell$ をどう変更するかの問題になるが，これは $r_{i,\ell}$ と $\mu(\mathcal{V}_\ell)$ の両方を変更しなければならないので難しい．そこで現在のクラスタ中心 $\mu(\mathcal{V}_\ell)$ を固定して μ_ℓ と表し，$r_{i,\ell}$ を更新する．$r_{i,\ell}=1$ は，x_i が $\mathcal{V}_\ell$ に所属することを意味したが，改めて x_i がどこのクラスタに所属するかは，クラスタの代表点の中で最も近いものに変更する．

$$r_{i,\ell}=\begin{cases} 1, & \ell=\underset{k=1,\dots,K}{\arg\min}\,\|x_i-\mu_k\|^2 \\ 0, & \text{otherwise}\end{cases} \tag{3.48}$$

各クラスタに所属するデータが変更されたので，$\mu(\mathcal{V}_\ell)$ も計算し直す必要があり，交互に最適化問題を解くことになる．アルゴリズムをまとめると，次の通りである．

Step 1 初期値 $\mu_\ell\in\mathbb{R}^p$, $\ell=1,\dots,K$ を設定する．

Step 2 $j=1,\dots$ に対し，

Step 2-1 式 (3.48) より $r_{i,\ell}^{(j)}$, $i=1,\dots,N$, $\ell=1,\dots,K$ を定める．

Step 2-2 次の式より $\mu_\ell^{(j)}$, $\ell=1,\dots,K$ を定める．

$$\mu_\ell^{(j)}=\frac{1}{\sum_{i=1}^N r_{i,\ell}}\sum_{i=1}^{N} r_{i,\ell}x_i.$$

Step 3 収束判定条件（例えば事前に設定した $\varepsilon>0$ に対して $\max_\ell\|\mu_\ell^{(j-1)}-\mu_\ell^{(j)}\|^2<\varepsilon$）を満たすか判定し，満たさなければ j を 1 つ繰り上げて Step 2 に戻る．

$U:=[\mu_1,\dots,\mu_K]$, $R=\{r_{i,\ell}\}$, $X=[x_1,\dots,x_N]$ と置くと，K 平均法は次のコスト関数の最小化問題となる．

$$\|X-UR\|^2 \tag{3.49}$$

R が 0 か 1 の要素しか持たないことから，連続最適化と離散最適化の両方の性質を持った難しめの最小化問題になっており，大域的最適解を求められる保証はない．また，初期値の選び方が非常に重要になる．

課題 14 与えられたデータ $x_i\in\mathbb{R}^2$, $i=1,\dots,1000$ を 3 つのクラスターに分類せよ．異なる初期値を 10 組以上選び，最も良さそうな分類をその理由とともに示せ．

K 平均法の初期値の選び方は色々工夫されているので，インターネットでも調べてみるとよい．

参考文献

[1] 松原 望, 統計学, 東京図書, 2013.

[2] 久保川 達也, 現代数理統計学の基礎, 共立出版, 2017.

[3] 小西 貞則, 多変量解析入門, 岩波書店, 2010.

[4] 森, 黒田, 足立, 最小二乗法・交互最小二乗法, 共立出版, 2017.

[5] 北川 源四郎, 時系列解析入門, 岩波書店, 2005.

[6] 持橋, 大羽，ガウス過程と機械学習, 講談社, 2019.

[7] I. Rosh, G. Y. Grabarnik, Sparse Modeling: Theory, Algorithms, and Applications, Chapman & Hall/Crc, 2006.
（Rish, Grabarnik，スパースモデリング: 理論，アルゴリズム，応用, ジャムハウス, 2019）

[8] 足立, 丸田, カルマンフィルタの基礎, 東京電機大学出版, 2012.

[9] Särkkä, Bayesian Filtering and Smoothing, Cambridge University Press, 2013.

[10] F. L. Lewis, L. Xie, D. Popa, Optimal and Robust Estimation with an Introduction to Stochastic Control Theory, CRC Press, 2008.