# PRML 輪読会
# 線形回帰モデル (3.1.1 節 – 3.1.2 節)

鈴木拓己

November 20, 2024

# Outline

# Outline

# 回帰分析の問題設定および線形回帰モデルの概要

訓練データ: $(\boldsymbol{x}_1, t_1), \ldots, (\boldsymbol{x}_N, t_N),\ \boldsymbol{x}_i \in \mathbb{R}^D,\ t_i \in \mathbb{R}.$

$$t = \underbrace{y(\boldsymbol{x})}_{\text{function}} + \underbrace{\varepsilon}_{\text{r.v.: error}} \quad \longrightarrow \quad \text{関数 } y \text{ を推定}$$

- 線形回帰モデル: 入力変数 (に対する非線形変換) の線形結合により $t$ を予測
  - 線形結合のパラメータ $\boldsymbol{w}$ に関しては線形 $\longrightarrow$ 解析が容易
  - 入力変数に関しては非線形 $\longrightarrow$ モデルの表現力が高い
- 回帰モデルの構成法
  - 関数 $y\colon \mathbb{R}^D \ni \boldsymbol{x} \longmapsto t \in \mathbb{R}$ を直接構成
    - 最も単純なアプローチ
  - 予測分布 $p(t \mid \boldsymbol{x})$ をモデル化
    - 実変数に対しては，損失関数として二乗損失関数を選ぶのが一般的であり，このとき最適解は $\mathbb{E}[t \mid \boldsymbol{x}]$ となる (1.5.5 節)

線形回帰モデルは解析的に扱いやすく，より洗練されたモデルの基礎として重要

# 最も単純な線形回帰モデル

**最も単純な線形回帰モデル**

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + w_1 x_1 + \cdots + w_D x_D, \quad \text{where} \quad \boldsymbol{x} = (x_1, \ldots, x_D)^\top \in \mathbb{R}^D.$$

- 単に線形回帰 (linear regression) と呼ばれる
- パラメータ $w_0, \ldots, w_D$ に関して線形
- 入力変数 $x_i$ に関しても線形であるため，表現力に乏しい

$\longrightarrow$ 入力変数に関して非線形な関数の線形結合を考え，モデルのクラスを拡張

線形基底関数[1]モデル

$$y(\boldsymbol{x}, \boldsymbol{w}) = \underbrace{w_0}_{\text{bias 項}} + \sum_{j=1}^{M-1} w_j \phi_j(\boldsymbol{x}), \quad \text{where} \quad \boldsymbol{x} = (x_1, \ldots, x_D)^\top \in \mathbb{R}^D. \tag{1.1}$$

- $M$: モデルの表現力を制御するハイパーパラメータ
- $\phi_j$: 基底関数 (basis function)
- 基底関数からなる特徴写像 (feature map) $\boldsymbol{\phi}$ により，入力 $\boldsymbol{x}$ を特徴空間 (feature space) へ写像:

$$\boldsymbol{\phi}\colon \mathbb{R}^D \ni \boldsymbol{x} \longmapsto \boldsymbol{\phi}(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), \ldots, \phi_{M-1}(\boldsymbol{x}))^\top \in \mathbb{R}^{M-1}$$

- $\phi_0(\boldsymbol{x}) \coloneqq 1$ とすれば，式 (1.1) は $y(\boldsymbol{x}, \boldsymbol{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x})$.

[1] https://ja.wikipedia.org/wiki/%E5%9F%BA%E5%BA%95%E9%96%A2%E6%95%B0

**Example 1 (1 変数多項式回帰)**

データ $(x, t) \in \mathbb{R} \times \mathbb{R}$ に対して，$\phi_j(x) \coloneqq x^j$ とおくと，

$$\boldsymbol{\phi} \colon \mathbb{R} \ni x \longmapsto (1, x, x^2, \ldots, x^{M-1})^\top \in \mathbb{R}^M,$$
$$y(x, \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_{M-1} x^{M-1}.$$

**Example 2 (多変数多項式回帰 (e.g. $D = 2$))**

データ $(\boldsymbol{x}, t) = ((x_1, x_2), t) \in \mathbb{R}^2 \times \mathbb{R}$ に対して，$\phi_{k,\ell} \coloneqq x_1^k x_2^\ell$ とおくと，

$$\boldsymbol{\phi} \colon \mathbb{R}^2 \ni \boldsymbol{x} \longmapsto (1, x_1, x_2, x_1 x_2, \ldots, x_1^{M_1-1} x_2^{M_2-1})^\top \in \mathbb{R}^{M_1 M_2},$$
$$y(\boldsymbol{x}, \boldsymbol{w}) = w_{0,0} + w_{1,0} x_1 + w_{0,1} x_2 + w_{1,1} x_1 x_2 + \cdots w_{M_1-1, M_2-1} x_1^{M_1-1} x_2^{M_2-1}.$$

# その他の基底関数の例 (1 変数の場合)

- スプライン関数 [2]
  - 多項式では入力空間のある領域の変化が他のすべての領域に及んでしまう
  - 入力空間を分割し，各領域で区分的に多項式をあてはめることにより解決できる
- ガウス型基底関数 (動径基底関数, RBF[3] の例)

$$\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\} \propto \mathcal{N}(\mu_j, s^2).$$

- シグモイド基底関数

$$\phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right), \quad \text{where} \quad \underbrace{\sigma(a) = \frac{1}{1+\exp(-a)}}_{\text{logistic sigmoid function}}.$$

[2] https://ja.wikipedia.org/wiki/%E3%82%B9%E3%83%97%E3%83%A9%E3%82%A4%E3%83%B3%E6%9B%B2%E7%B7%9A

[3] https://ja.wikipedia.org/wiki/%E6%94%BE%E5%B0%84%E5%9F%BA%E5%BA%95%E9%96%A2%E6%95%B0

# その他の基底関数の例 (1 変数の場合)

(左) 多項式関数 (中央) ガウス型基底関数 (右) シグモイド基底関数

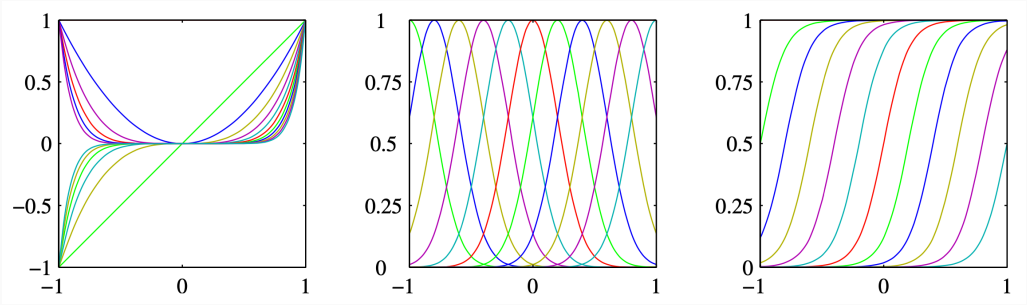


**Figure 1:** 基底関数の例

# その他の基底関数の例 (1 変数の場合)

- tanh 関数
  - 下記の式変形より，ロジスティックシグモイド関数の線形結合と等価 ($\longrightarrow$ 演習 3.1)

$$\begin{aligned}\tanh(a) &= \frac{\exp(a)-\exp(-a)}{\exp(a)+\exp(-a)} = \frac{1-\exp(-2a)}{1+\exp(-2a)} \\ &= \frac{2}{1+\exp(-2a)} - \frac{1+\exp(-2a)}{1+\exp(-2a)} = 2\underline{\sigma(2a)} - 1.\end{aligned} \tag{1.2}$$

- フーリエ基底関数
  - 様々な周波数の sin 関数，cos 関数による級数展開 (cf. フーリエ級数展開)
- ウェーブレット基底関数 [4]
  - フーリエ基底では周波数ごとに定まる基底関数は入力空間において無限に広がる
  - ウェーブレット基底は入力空間でも周波数領域でも局所的
  - 詳細は長くなりそうなので，定義と簡単な例のみ確認 (次ページ以降)

[4] https://ja.wikipedia.org/wiki/%E3%82%A6%E3%82%A7%E3%83%BC%E3%83%96%E3%83%AC%E3%83%83%E3%83%88

**Definition 1 (ウェーブレット基底)**

関数 $\psi \in L^2(\mathbb{R}) = \{f\colon \mathbb{R} \to \mathbb{C} \mid \int_{\mathbb{R}} |f(x)|^2 dx < \infty\}$ に対し，関数 $\psi_{m,n}$ を

$$\psi_{m,n}(t) = 2^{m/2}\psi(2^m t - n), \quad m \in \mathbb{Z}, \quad n \in \mathbb{Z}$$

で定義する．関数列 $\{\psi_{m,n}\}_{m\in\mathbb{Z},n\in\mathbb{Z}}$ が $L^2(\mathbb{R})$ において基底[5] をなすとき，$\psi_{m,n}$ を ウェーブレット，$\psi$ をそのマザーウェーブレット，$\{\psi_{m,n}\}_{m\in\mathbb{Z},n\in\mathbb{Z}}$ を ウェーブレット基底という．ウェーブレット基底 $\{\psi_{m,n}\}_{m\in\mathbb{Z},n\in\mathbb{Z}}$ が正規直交条件

$$\langle \psi_{m,n}, \psi_{k,\ell} \rangle = \int_{\mathbb{R}} \psi_{m,n}(t)\overline{\psi_{k,\ell}(t)}dt = \delta_{mk}\delta_{n\ell}$$

を満たすとき，$\{\psi_{m,n}\}_{m\in\mathbb{Z},n\in\mathbb{Z}}$ を正規直交ウェーブレット基底という．

[5] すなわち，$\forall f \in L^2(\mathbb{R})$ に対して，$f = \sum_{m\in\mathbb{Z}} \sum_{n\in\mathbb{Z}} \langle f, \psi_{m,n} \rangle \psi_{m,n}$.

**Example 3 (直交ウェーブレットの例: Haar ウェーブレット [6])**

マザーウェーブレット $\psi$ を

$$\psi(t) = \begin{cases} 1, & 0 \le t < 1/2, \\ -1, & 1/2 \le t < 1, \\ 0, & \textit{otherwise}, \end{cases}$$

で定めると，$\{\psi_{m,n}\}_{m\in\mathbb{Z},n\in\mathbb{Z}}$ は直交ウェーブレットとなる．

Haar ウェーブレットのグラフは次ページ．
他にもいろいろな直交ウェーブレットがある (気になる人は pdf[7] を参照してください)．

[6] https://ja.wikipedia.org/wiki/%E3%83%8F%E3%83%BC%E3%83%AB%E3%82%A6%E3%82%A7%E3%83%BC%E3%83%96%E3%83%AC%E3%83%83%E3%83%88

[7] http://wwwcs.ce.nihon-u.ac.jp/lab/moritaleb224w.pdf

# その他の基底関数の例 (1 変数の場合)

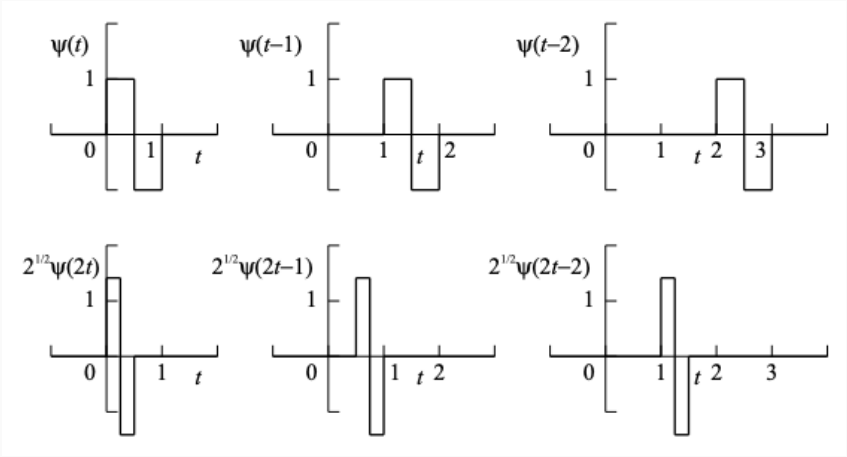


**Figure 2:** Haar ウェーブレットの場合の関数 $\psi_{0,0} = \psi, \psi_{0,1}, \psi_{0,2}, \psi_{1,0}, \psi_{1,1}, \psi_{1,2}$

# Outline

# 最尤推定と最小二乗法の関係

以後の議論は基底関数の形に依らないため，基底関数は特に限定しないことにする．
また，簡単のため，目標変数 $t$ が 1 次元の場合を扱う (多次元の場合は 3.1.5 節)．

ここでは，1 章の多項式フィッティングで扱った，最小二乗法と最尤推定との関係をより詳細に議論する．
1 章で見たように，多項式回帰において二乗和誤差関数を最小化するパラメータ $\boldsymbol{w}$ は，ガウスノイズモデルの下での最尤推定解と一致する:

$$t = \underset{\text{deterministic}}{y(\boldsymbol{x}, \boldsymbol{w})} + \underset{\text{stochastic}}{\varepsilon}, \quad \text{where} \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

すなわち，目標変数 $t$ の予測分布は，精度パラメータ $\beta$ を用いて，

$$p(t \mid \boldsymbol{x}, \boldsymbol{w}, \beta) = \mathcal{N}(t \mid y(\boldsymbol{x}, \boldsymbol{w}), \beta^{-1}), \quad \text{where} \quad \beta^{-1} = \sigma^2$$

で与えられる．

# 最尤推定と最小二乗法の関係

1.5.5 節で見たように，二乗損失関数を仮定すれば，入力 $\boldsymbol{x}$ の値に対する最適な予測値は目標変数の条件付き期待値

$$\mathbb{E}[t \mid \boldsymbol{x}] = \int_{\mathbb{R}} tp(t \mid \boldsymbol{x})dt = \int_{\mathbb{R}} t\mathcal{N}(t \mid y(\boldsymbol{x}, \boldsymbol{w}), \beta^{-1})dt = y(\boldsymbol{x}, \boldsymbol{w})$$

で与えられる．

**Remark 1**

ガウスノイズの仮定は，条件付き分布 $p(t \mid \boldsymbol{x})$ が単峰性のため，応用場面によっては不適切である可能性がある．

$\longrightarrow$ 14.5.1 節で条件付きガウス混合分布への拡張を議論 (多峰性の条件付き分布)

# 最尤推定と最小二乗法の関係

最尤推定により，予測分布 $p(t \mid \boldsymbol{x}, \boldsymbol{w}, \beta)$ のパラメータ $\boldsymbol{w}, \beta$ を推定する．
入力 $\mathbf{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ と対応する目標値 $t_1, \ldots, t_N$ からなるデータ集合を考え，
$\mathbf{t} = (t_1, \ldots, t_N)^\top$ とする．各データ点が分布 $\mathcal{N}(t \mid y(\boldsymbol{x}, \boldsymbol{w}), \beta^{-1})$ から独立に生成されたと仮定すると，尤度関数は，

$$p(\mathbf{t} \mid \mathbf{X}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid y(\boldsymbol{x}_n, \boldsymbol{w}), \beta^{-1}) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_n), \beta^{-1})$$

で与えられる．
教師あり学習問題では，入力変数の分布をモデル化しないため，$\boldsymbol{x}$ は条件としてしか現れないので，以後は尤度関数の表記から $\boldsymbol{x}$ を省略して $p(\mathbf{t} \mid \boldsymbol{w}, \beta)$ と書く．

このとき，対数尤度関数は，

$$\begin{aligned}\ln p(\mathbf{t} \mid \boldsymbol{w}, \beta) &= \ln \prod_{n=1}^{N} \mathcal{N}(t_n \mid \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_n), \beta^{-1}) = \sum_{n=1}^{N} \ln \mathcal{N}(t_n \mid \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_n), \beta^{-1}) \\ &= \sum_{n=1}^{N} \ln \left( \frac{1}{(2\pi\beta^{-1})^{1/2}} \exp\left( -\frac{(t_n - \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_n))^2}{2\beta^{-1}} \right) \right) \\ &= \sum_{n=1}^{N} \left( \frac{1}{2} \ln \beta - \frac{1}{2} \ln(2\pi) - \beta \frac{1}{2} (t_n - \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_n))^2 \right) \\ &= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta \underbrace{\frac{1}{2} \sum_{n=1}^{N} (t_n - \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_n))^2}_{E_D(\boldsymbol{w})\text{: sum-of-squares error}} \end{aligned} \tag{2.1}$$

となる．最尤推定解を得るには，これを $\boldsymbol{w}$ と $\beta$ について最大化すればよい．

式 (2.1) より，パラメータ $\boldsymbol{w}$ に関する対数尤度関数 $\ln p(\mathbf{t} \mid \boldsymbol{w}, \beta)$ の最大化は，二乗和誤差関数 $E_D(\boldsymbol{w})$ の最小化と等価である．

- ガウスノイズモデル下の最尤推定と最小二乗法との等価性

ここで，以下で与えられる計画行列 (design matrix)

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\boldsymbol{x}_1) & \phi_1(\boldsymbol{x}_1) & \cdots & \phi_{M-1}(\boldsymbol{x}_1) \\ \phi_0(\boldsymbol{x}_2) & \phi_1(\boldsymbol{x}_2) & \cdots & \phi_{M-1}(\boldsymbol{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\boldsymbol{x}_N) & \phi_1(\boldsymbol{x}_N) & \cdots & \phi_{M-1}(\boldsymbol{x}_N) \end{pmatrix} \in \mathbb{R}^{N \times M}$$

を考えると，二乗和誤差関数 $E_D(\boldsymbol{w})$ は

$$E_D(\boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} (t_n - \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_n))^2 = \frac{1}{2} (\mathbf{t} - \boldsymbol{\Phi}\boldsymbol{w})^\top (\mathbf{t} - \boldsymbol{\Phi}\boldsymbol{w}).$$

$$E_D(\boldsymbol{w}) = \frac{1}{2}(\mathbf{t} - \boldsymbol{\Phi}\boldsymbol{w})^\top(\mathbf{t} - \boldsymbol{\Phi}\boldsymbol{w}) = \frac{1}{2}\left(\mathbf{t}^\top\mathbf{t} - \mathbf{t}^\top\boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{w}^\top\boldsymbol{\Phi}^\top\mathbf{t} + \boldsymbol{w}^\top\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{w}\right)$$
$$= \frac{1}{2}\left(\boldsymbol{w}^\top\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{w} - 2\mathbf{t}^\top\boldsymbol{\Phi}\boldsymbol{w} + \mathbf{t}^\top\mathbf{t}\right)$$

であり，

(1) $\boldsymbol{\Phi}^\top\boldsymbol{\Phi}$ は対称 $\left(\because \left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^\top = \boldsymbol{\Phi}^\top\left(\boldsymbol{\Phi}^\top\right)^\top = \boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)$，

(2) $\boldsymbol{\Phi}^\top\boldsymbol{\Phi}$ はグラム行列 [8] $\Longrightarrow \boldsymbol{\Phi}^\top\boldsymbol{\Phi} \succeq 0$，

ゆえ，$E_D(\boldsymbol{w})$ の $\boldsymbol{w}$ に関する最小化は制約なし凸 2 次計画問題 [9] なので，

$$\boldsymbol{w}_{\mathsf{ML}} = \underset{\boldsymbol{w}\in\mathbb{R}^M}{\operatorname{argmin}} E_D(\boldsymbol{w}) \iff \left.\frac{\partial}{\partial\boldsymbol{w}}E_D(\boldsymbol{w})\right|_{\boldsymbol{w}=\boldsymbol{w}_{\mathsf{ML}}} = 0.$$

[8] https://mathlandscape.com/gram-matrix/

[9] http://www.me.titech.ac.jp/~mizu_lab/lib/pdf/kougisiryou/suuti/handout/11/suuti_kougi11-6.pdf

## パラメータ $\boldsymbol{w}$ に関する対数尤度関数 $\ln p(\mathbf{t} \mid \boldsymbol{w}, \beta)$ の最大化

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{w}} \ln p(\mathbf{t} \mid \boldsymbol{w}, \beta) &= \frac{\partial}{\partial \boldsymbol{w}} \left(-\beta E_D(\boldsymbol{w})\right) = -\frac{\beta}{2} \frac{\partial}{\partial \boldsymbol{w}} (\mathbf{t} - \boldsymbol{\Phi}\boldsymbol{w})^\top (\mathbf{t} - \boldsymbol{\Phi}\boldsymbol{w}) \\
&= -\frac{\beta}{2} \frac{\partial}{\partial \boldsymbol{w}} \left(\boldsymbol{w}^\top \boldsymbol{\Phi}^\top \boldsymbol{\Phi}\boldsymbol{w} - 2\mathbf{t}^\top \boldsymbol{\Phi}\boldsymbol{w} + \mathbf{t}^\top \mathbf{t}\right) \\
&= -\frac{\beta}{2} \left(2\boldsymbol{\Phi}^\top \boldsymbol{\Phi}\boldsymbol{w} - 2\boldsymbol{\Phi}^\top \mathbf{t}\right) = -\beta \left(\boldsymbol{\Phi}^\top \boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{\Phi}^\top \mathbf{t}\right) \\
&\qquad \left(\because \frac{\partial}{\partial \boldsymbol{w}} \boldsymbol{w}^\top \mathbf{A} \boldsymbol{w} = (\mathbf{A} + \mathbf{A}^\top)\boldsymbol{w}, \quad \frac{\partial}{\partial \boldsymbol{w}} \boldsymbol{a}^\top \boldsymbol{w} = \boldsymbol{a}\right)
\end{aligned}
$$

ゆえ，$\boldsymbol{\Phi}^\top \boldsymbol{\Phi}$ が正則ならば，最尤推定解 $\boldsymbol{w}_{\mathsf{ML}}$ は，

$$
\frac{\partial}{\partial \boldsymbol{w}} \ln p(\mathbf{t} \mid \boldsymbol{w}, \beta) \Big|_{\boldsymbol{w}=\boldsymbol{w}_{\mathsf{ML}}} = 0 \iff \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \boldsymbol{w}_{\mathsf{ML}} - \boldsymbol{\Phi}^\top \mathbf{t} = 0 \tag{2.2}
$$

$$
\iff \boldsymbol{w}_{\mathsf{ML}} = \left(\boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^\top \mathbf{t}. \tag{2.3}
$$

- 方程式 (2.3) を最小二乗問題の正規方程式 (normal equation) という．
- 行列 $\boldsymbol{\Phi}^{\dagger} \equiv \left(\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\top} \in \mathbb{R}^{M\times N}$ を行列 $\boldsymbol{\Phi}$ のムーア-ペンローズの擬似逆行列[10] (Moore-Penrose pseudo-inverse matrix) という．
  - 通常の逆行列の概念の非正方行列への拡張
  - 実際，$\boldsymbol{\Phi}$ が正則ならば，$\boldsymbol{\Phi}^{-1}$ が存在し，

$$\boldsymbol{\Phi}^{\dagger} \equiv \left(\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\top} = \boldsymbol{\Phi}^{-1}\left(\boldsymbol{\Phi}^{\top}\right)^{-1}\boldsymbol{\Phi}^{\top} = \boldsymbol{\Phi}^{-1}.$$

  - 任意の行列 $\mathbf{M}$ に対して，定義の 4 条件を満たす擬似逆行列 $\mathbf{M}^{\dagger}$ が一意に存在
- 式 (2.2) において $\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}$ が正則でない場合については後述．
  - 計画行列 $\boldsymbol{\Phi}$ がフルランクでない場合に発生する (多重共線性[11] (Multicollinearity))．

[10] https://ja.wikipedia.org/wiki/%E3%83%A0%E3%83%BC%E3%82%A2%E3%83%BB%E3%83%9A%E3%83%B3%E3%83%AD%E3%83%BC%E3%82%BA%E9%80%86%E8%A1%8C%E5%88%97

[11] https://ja.wikipedia.org/wiki/%E5%A4%9A%E9%87%8D%E5%85%B1%E7%B7%9A%E6%80%A7

# パラメータ $\beta$ に関する対数尤度関数 $\ln p(\mathbf{t} \mid \boldsymbol{w}_{\mathsf{ML}}, \beta)$ の最大化

次に，パラメータ $\beta$ に関して対数尤度関数 $\ln p(\mathbf{t} \mid \boldsymbol{w}_{\mathsf{ML}}, \beta)$ を最大化する．

$$\begin{aligned}\frac{\partial}{\partial \beta} \ln p(\mathbf{t} \mid \boldsymbol{w}_{\mathsf{ML}}, \beta) &= \frac{\partial}{\partial \beta} \left( \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta \frac{1}{2} \sum_{n=1}^{N} (t_n - \boldsymbol{w}_{\mathsf{ML}}^{\top} \boldsymbol{\phi}(\boldsymbol{x}_n))^2 \right) \\ &= \frac{N}{2} \frac{1}{\beta} - \frac{1}{2} \sum_{n=1}^{N} (t_n - \boldsymbol{w}_{\mathsf{ML}}^{\top} \boldsymbol{\phi}(\boldsymbol{x}_n))^2\end{aligned}$$

ゆえ，最尤推定解 $\beta_{\mathsf{ML}}$ は，

$$\left. \frac{\partial}{\partial \beta} \ln p(\mathbf{t} \mid \boldsymbol{w}_{\mathsf{ML}}, \beta) \right|_{\beta = \beta_{\mathsf{ML}}} = 0 \iff \frac{1}{\beta_{\mathsf{ML}}} = \frac{1}{N} \sum_{n=1}^{N} (\underbrace{t_n - \boldsymbol{w}_{\mathsf{ML}}^{\top} \boldsymbol{\phi}(\boldsymbol{x}_n)}_{\text{residual}})^2$$

として得られ，$\beta_{\mathsf{ML}}^{-1} = \sigma_{\mathsf{ML}}^2$ は回帰関数 $y(\boldsymbol{x}, \boldsymbol{w}_{\mathsf{ML}})$ 周りでの目標値の残差分散で与えられることがわかる．

# バイアスパラメータ $w_0$ の役割

二乗和誤差関数 $E_D(\boldsymbol{w})$ において，バイアスパラメータ $w_0$ を明示的に書くと，

$$E_D(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}(t_n - \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_n))^2 = \frac{1}{2}\sum_{n=1}^{N}\left(t_n - w_0 - \sum_{j=1}^{M-1} w_j\phi_j(\boldsymbol{x}_n)\right)^2$$

$$\Longrightarrow \frac{\partial}{\partial w_0}E_D(\boldsymbol{w}) = -\sum_{n=1}^{N}\left(t_n - w_0 - \sum_{j=1}^{M-1} w_j\phi_j(\boldsymbol{x}_n)\right)$$

ゆえ，$E_D(\boldsymbol{w})$ を最小化するパラメータ $\hat{w}_0$ は，

$$\left.\frac{\partial}{\partial w_0}E_D(\boldsymbol{w})\right|_{w_0=\hat{w}_0} = 0 \iff \hat{w}_0 = \frac{1}{N}\sum_{n=1}^{N} t_n - \sum_{j=1}^{M-1} w_j\left(\frac{1}{N}\sum_{n=1}^{N}\phi_j(\boldsymbol{x}_n)\right) \equiv \bar{t} - \sum_{j=1}^{M-1} w_j\overline{\phi_j}.$$

バイアスパラメータ $w_0$ は，訓練データ $\{(\boldsymbol{x}_n, t_n)\}_{n=1}^N$ に関する，目標値の平均 $\bar{t}$ と，基底関数の値の平均の重み付き和 $\sum_{j=1}^{M-1} w_j\overline{\phi_j}$ との差を埋め合わせる役割をしている．

# Outline

計画行列 $\boldsymbol{\Phi}$ の第 $j$ 列を $\boldsymbol{\varphi}_j = (\phi_{j-1}(\boldsymbol{x}_1), \ldots, \phi_{j-1}(\boldsymbol{x}_N))^\top \in \mathbb{R}^N$ とおくと，

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\boldsymbol{x}_1) & \phi_1(\boldsymbol{x}_1) & \cdots & \phi_{M-1}(\boldsymbol{x}_1) \\ \phi_0(\boldsymbol{x}_2) & \phi_1(\boldsymbol{x}_2) & \cdots & \phi_{M-1}(\boldsymbol{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\boldsymbol{x}_N) & \phi_1(\boldsymbol{x}_N) & \cdots & \phi_{M-1}(\boldsymbol{x}_N) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\varphi}_1 & \boldsymbol{\varphi}_2 & \cdots & \boldsymbol{\varphi}_M \end{pmatrix} \in \mathbb{R}^{N \times M}.$$

このとき，基底関数の数 $M$ がデータ点数 $N$ よりも小さければ[12]，$M$ 個のベクトル $\{\boldsymbol{\varphi}_j\}_{j=1}^M$ は $M$ 次元線形部分空間 $\mathcal{S} \subsetneq \mathbb{R}^N$ を張る:

$$\mathcal{S} = \mathrm{Im}\, \boldsymbol{\Phi} = \mathrm{span}_{\mathbb{R}}(\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_M) = \left\{ \sum_{j=1}^M w_j \boldsymbol{\varphi}_j \;\middle|\; \boldsymbol{w} \in \mathbb{R}^M \right\} \subsetneq \mathbb{R}^N.$$

[12] $M \geq N$ かつ $\boldsymbol{\Phi}$ がフルランクならば $\mathcal{S} = \mathbb{R}^N$.

ベクトル $\mathbf{y} \in \mathbb{R}^N$ を

$$\mathbf{y} \coloneqq (y(\boldsymbol{x}_1, \boldsymbol{w}), \ldots, y(\boldsymbol{x}_N, \boldsymbol{w}))^\top$$

で定義すると，$y(\boldsymbol{x}_n, \boldsymbol{w}) = \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_n)$ より，

$$\mathbf{y} = (\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_1), \ldots, \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_N))^\top = \boldsymbol{\Phi} \boldsymbol{w} \in \mathcal{S}.$$

ここで，最小二乗解 $\boldsymbol{w}_{\mathsf{ML}}$ は，$\|\cdot\|_2$ をユークリッドノルムとして，

$$\boldsymbol{w}_{\mathsf{ML}} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^M} E_D(\boldsymbol{w}) = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^M} \frac{1}{2} \|\mathbf{t} - \underbrace{\boldsymbol{\Phi} \boldsymbol{w}}_{\mathbf{y}} \|_2^2$$

で与えられるので，最小二乗解 $\boldsymbol{w}_{\mathsf{ML}} \in \mathbb{R}^M$ を求めることは，ユークリッド距離で $\mathbf{t} \in \mathbb{R}^N$ に最も近い $\boldsymbol{\Phi} \boldsymbol{w} = \mathbf{y} \in \mathcal{S}$ を選ぶことに相当する．この $\mathbf{y} \in \mathcal{S}$ は，$\mathbf{t} \in \mathbb{R}^N$ の部分空間 $\mathcal{S}$ 上への正射影に対応する (次ページ参照)．

直観的な説明

- $N = 3, M = 2$ の場合は下図

$$\begin{aligned} &\text{minimize} && \|\mathbf{t} - \mathbf{y}\|_2^2 \\ &\text{subject to} && \mathbf{y} \in \mathcal{S} \end{aligned}$$

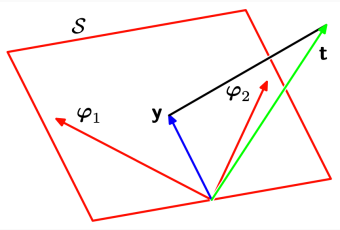


**Figure 3:** 最小二乗解の幾何学的解釈

理論的な説明

行列 $\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\top$ は任意のベクトル $\boldsymbol{v} \in \mathbb{R}^N$ を $\mathcal{S} = \mathrm{span}_{\mathbb{R}}(\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_M)$ に正射影する ($\because$ 演習 3.2). したがって,

$$\boldsymbol{w}_{\mathrm{ML}} = \left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\top\mathbf{t}$$

より,

$$\hat{\mathbf{y}} = \boldsymbol{\Phi}\boldsymbol{w}_{\mathrm{ML}} = \boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\top\mathbf{t} \quad (3.1)$$

は $\mathbf{t} \in \mathbb{R}^N$ の部分空間 $\mathcal{S}$ 上への正射影となっている.

# $\boldsymbol{\Phi}^\top\boldsymbol{\Phi}$ が正則でない (非正則に近い) 場合の対処法

式 (2.2) において $\boldsymbol{\Phi}^\top\boldsymbol{\Phi}$ が正則でない場合には，最尤推定解 $\boldsymbol{w}_{\mathrm{ML}}$ を求められない．
特に，2 つ以上の基底ベクトル $\boldsymbol{\varphi}_j$ が線形従属の場合，$\operatorname{rank}\boldsymbol{\Phi} < \min(M, N)$ となり，
このような問題が発生する．また，2 つ以上の基底ベクトル $\boldsymbol{\varphi}_j$ が線形従属に近い場合
にも同様の問題が発生し，求まるパラメータ $\boldsymbol{w}_{\mathrm{ML}}$ が数値的に不安定になる．
この問題の対処法として以下のような手法がある．

- 特異値分解 [13] (singular value decomposition; SVD)
  - 陽に逆行列 $\left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^{-1}$ を計算せずとも，擬似逆行列 $\boldsymbol{\Phi}^\dagger$ を計算することができる．
- 正則化最小二乗法 ($\longrightarrow$ 3.1.4 節)
  - 正則化項を加えることにより，$\boldsymbol{\Phi}^\top\boldsymbol{\Phi}$ が正則でない場合でも最適解 $\boldsymbol{w}_{\mathrm{ML}}$ が得られる．
  - 特に，$q = 2$ (Ridge 回帰) の場合は解析的に最適解が求められる．

$$\text{minimize} \quad \|\mathbf{t} - \boldsymbol{\Phi}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_q^q \quad \text{subject to} \quad \boldsymbol{w} \in \mathbb{R}^M, \quad \text{where } \|\cdot\|_q \text{ is } L^q\text{-norm}^{14}.$$

[13] https://ja.wikipedia.org/wiki/%E7%89%B9%E7%95%B0%E5%80%A4%E5%88%86%E8%A7%A3
[14] https://ja.wikipedia.org/wiki/Lp%E7%A9%BA%E9%96%93

**Theorem 1 (特異値分解定理 [15])**

$\mathbf{M} \in \mathbb{C}^{m\times n}$ を $\operatorname{rank}\mathbf{M} = r$ の行列とする．このとき，$\mathbf{\Sigma} \in \mathbb{C}^{m\times n}$ が一意に存在し，

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{*} \tag{3.2}$$

の分解が成り立つ．ただし，$\mathbf{U}, \mathbf{V}$ はそれぞれ $m \times m$, $n \times n$ のユニタリ行列であり，$\mathbf{\Sigma}$ は半正定値行列 $\mathbf{M}\mathbf{M}^{*}$ あるいは $\mathbf{M}^{*}\mathbf{M}$ の正の固有値 [16] の平方根 $\sigma_1 \geq \cdots \geq \sigma_r > 0$ および $q = \min(m, n)$, $\sigma_{r+1} = \cdots = \sigma_q = 0$ を用いて

$$\mathbf{\Sigma} = \begin{cases} \begin{pmatrix} \mathbf{\Delta} & \mathbf{O} \end{pmatrix}, & (m < n), \\ \mathbf{\Delta}, & (m = n), \\ \begin{pmatrix} \mathbf{\Delta} \\ \mathbf{O} \end{pmatrix}, & (m > n), \end{cases} \quad \textit{where} \quad \mathbf{\Delta} = \operatorname{diag}(\sigma_1, \ldots, \sigma_q) = \begin{pmatrix} \sigma_1 & & \mathbf{O} \\ & \ddots & \\ \mathbf{O} & & \sigma_q \end{pmatrix}$$

で与えられる．分解 (3.2) を $\mathbf{M}$ の特異値分解，$\sigma_1, \ldots, \sigma_q$ を $\mathbf{M}$ の特異値と呼ぶ．

[15] 証明: https://qiita.com/gyu-don/items/2e4fd7e945d8c349afb5

[16] $\mathbf{M}\mathbf{M}^{*}$ と $\mathbf{M}^{*}\mathbf{M}$ の非零固有値は一致する: http://teagis.ip.is.saga-u.ac.jp/svd.pdf

ユニタリ行列による変換は等長変換なので，特異値分解 $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ におけるそれぞれの線形変換は，

- $\mathbf{V}^*$：回転 ($\because$ ユニタリ行列)
- $\mathbf{\Sigma}$：拡大・縮小
- $\mathbf{U}$：回転 ($\because$ ユニタリ行列)

を表す．また，$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ より

$$\mathbf{M}\mathbf{M}^* = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^\top\mathbf{U}^*, \quad \mathbf{M}^*\mathbf{M} = \mathbf{V}\mathbf{\Sigma}^\top\mathbf{\Sigma}\mathbf{V}^*,$$

ゆえ，特異値分解における $\mathbf{U}, \mathbf{V}$ はそれぞれ $\mathbf{M}\mathbf{M}^*, \mathbf{M}^*\mathbf{M}$ の固有値分解によって得られる．

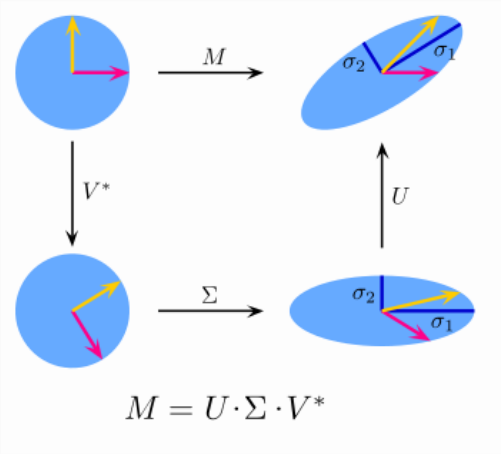


**Figure 4:** 特異値分解の図示

# 特異値分解による擬似逆行列 $\boldsymbol{\Phi}^\dagger$ の計算

特異値分解定理により，計画行列 $\boldsymbol{\Phi} \in \mathbb{R}^{N\times M}$ に対して，直交行列 $\mathbf{U} \in \mathbb{R}^{N\times N}$, $\mathbf{V} \in \mathbb{R}^{M\times M}$ および行列 $\boldsymbol{\Sigma} \in \mathbb{R}^{N\times M}$ が存在し，$\boldsymbol{\Phi} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ の分解が成り立つ．
いま，$\mathbf{U}, \mathbf{V}$ は直交行列ゆえ，

$$\mathbf{U}\mathbf{U}^\top = \mathbf{U}^\top\mathbf{U} = \mathbf{I}, \quad \mathbf{V}\mathbf{V}^\top = \mathbf{V}^\top\mathbf{V} = \mathbf{I}, \quad \mathbf{U}^{-1} = \mathbf{U}^\top, \quad \mathbf{V}^{-1} = \mathbf{V}^\top$$

が成り立つことに注意すると，

$$\begin{aligned}
\boldsymbol{\Phi}^\dagger &= \left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\top = \left(\mathbf{V}\boldsymbol{\Sigma}^\top\mathbf{U}^\top\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top\right)^{-1}\mathbf{V}\boldsymbol{\Sigma}^\top\mathbf{U}^\top \\
&= \mathbf{V}\left(\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}\right)^{-1}\mathbf{V}^\top\mathbf{V}\boldsymbol{\Sigma}^\top\mathbf{U}^\top = \mathbf{V}\left(\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma}^\top\mathbf{U}^\top \\
&\equiv \mathbf{V}\boldsymbol{\Sigma}^\dagger\mathbf{U}^\top. \qquad (\boldsymbol{\Sigma}^\dagger \in \mathbb{R}^{M\times N})
\end{aligned}$$

ただし，$\boldsymbol{\Sigma}^\dagger \equiv \left(\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma}^\top$ は，$\boldsymbol{\Sigma}$ の非零成分 [17] の逆数を成分とする行列の転置．

[17] 数値計算では，許容誤差よりも大きい要素のみが非零と見なされ，他の要素は 0 に置き換えられる．

# Outline

# まとめ

- 線形基底関数モデル
  - 基底関数 $\phi_j$ からなる特徴写像 $\boldsymbol{\phi}$ によって線形回帰モデルを拡張
  - 様々な基底関数が使われる
- 最尤推定と最小二乗法
  - 最小二乗解とガウスノイズモデル下の最尤推定解は一致
  - 計画行列 $\boldsymbol{\Phi}$ がフルランクならば，最小二乗解 $\boldsymbol{w}_{\mathrm{ML}}$ は解析的に求められる
- 最小二乗法の幾何学
  - 最小二乗解 $\boldsymbol{w}_{\mathrm{ML}}$ の求解は，訓練データの目標値からなるベクトル $\mathbf{t} \in \mathbb{R}^N$ のモデル空間 $\mathcal{S} = \mathrm{Im}\, \boldsymbol{\Phi} = \mathrm{span}_{\mathbb{R}}(\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_M)$ 上への正射影を求めることに対応する
  - $\boldsymbol{\Phi}^\top \boldsymbol{\Phi}$ が非正則に近い場合の対処法
    * 特異値分解
    * 正則化最小二乗法
    * 正則化の一種として，主成分回帰 [18] (principal component regression; PCR) などもある

[18] https://en.wikipedia.org/wiki/Principal_component_regression

# Outline

## 演習 3.1 (基本) www

tanh 関数とロジスティックシグモイド関数は次のように関係付けられることを示せ．

$$\tanh(a) = 2\sigma(2a) - 1. \tag{5.1}$$

さらに，次の形のロジスティックシグモイド関数の線形結合

$$y(x, \boldsymbol{w}) = w_0 + \sum_{j=1}^{M} w_j \sigma\left(\frac{x - \mu_j}{s}\right)$$

は次の形の tanh 関数の線形結合

$$y(x, \boldsymbol{u}) = u_0 + \sum_{j=1}^{M} u_j \tanh\left(\frac{x - \mu_j}{2s}\right)$$

と等価であることを示し，新しいパラメータ $\{u_0, \ldots, u_M\}$ ともとのパラメータ $\{w_0, \ldots, w_M\}$ を関係付ける式を求めよ．

(解答) 式 (1.2) より，式 (5.1) が成り立つ．これを用いると，

$$y(x, \boldsymbol{u}) = u_0 + \sum_{j=1}^{M} u_j \tanh\left(\frac{x - \mu_j}{2s}\right) = u_0 + \sum_{j=1}^{M} u_j \left\{2\sigma\left(\frac{x - \mu_j}{s}\right) - 1\right\} = u_0 - \sum_{j=1}^{M} u_j + \sum_{j=1}^{M} 2u_j \sigma\left(\frac{x - \mu_j}{s}\right)$$

ゆえ，パラメータの関係式は $w_0 = u_0 - \sum_{j=1}^{M} u_j$, $w_j = 2u_j$ $(j = 1, \ldots, M)$ で与えられる．

## 演習 3.2 (標準)

行列

$$\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\top$$

は任意のベクトル $\boldsymbol{v}$ を $\boldsymbol{\Phi}$ の列ベクトルで張られる空間の上に正射影することを示せ．そしてこの結果を使って，最小二乗解 (2.3) は図 3 で示した多様体 $\mathcal{S}$ の上にベクトル $\mathbf{t}$ を正射影することに対応していることを示せ．

(解答) 行列 $\boldsymbol{\Phi} = (\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_M) \in \mathbb{R}^{N\times M}$ に対して，部分空間 $\mathcal{S} \in \mathbb{R}^N$ を $\mathcal{S} = \mathrm{span}_{\mathbb{R}}(\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_M)$ とおく．任意のベクトル $\boldsymbol{v} \in \mathbb{R}^N$ の $\mathcal{S}$ 上への正射影 $\boldsymbol{p}$ は，$\boldsymbol{w} = (w_1, \ldots, w_M) \in \mathbb{R}^M$ を用いて，$\boldsymbol{p} = \sum_{j=1}^M w_j\boldsymbol{\varphi}_j = \boldsymbol{\Phi}\boldsymbol{w}$ と表せる．このとき，ベクトル $\boldsymbol{v}-\boldsymbol{p}$ と $\mathcal{S}$ の基底 $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_M$ は直交するので，

$$\begin{aligned}
\left\langle \boldsymbol{v}-\boldsymbol{p}, \boldsymbol{\varphi}_j \right\rangle = 0, \quad \forall j \in \{1, \ldots, M\} &\iff \boldsymbol{\Phi}^\top(\boldsymbol{v}-\boldsymbol{p}) = 0 \\
&\iff \boldsymbol{\Phi}^\top(\boldsymbol{v}-\boldsymbol{\Phi}\boldsymbol{w}) = 0 \\
&\iff \boldsymbol{w} = \left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\top\boldsymbol{v}
\end{aligned}$$

が成り立つ．したがって，$\boldsymbol{v}$ の正射影 $\boldsymbol{p}$ は，

$$\boldsymbol{p} = \boldsymbol{\Phi}\boldsymbol{w} = \boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\top\boldsymbol{v}$$

で表されるので，行列 $\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\top$ は任意のベクトル $\boldsymbol{v}$ を $\boldsymbol{\Phi}$ の列ベクトルで張られる空間の上に正射影する．最小二乗解 (2.3) が多様体 $\mathcal{S}$ の上にベクトル $\mathbf{t}$ を正射影することに対応していることは，式 (3.1) より成り立つ． □

## 演習 3.3 (基本)

それぞれのデータ点 $t_n$ に重み要素 $r_n > 0$ が割り当てられており，二乗和誤差関数が

$$E_D(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N} r_n \left\{ t_n - \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_n) \right\}^2$$

となるデータ集合を考える．このとき，この誤差関数を最小にする解 $\boldsymbol{w}^*$ についての式を求めよ．また，(i) ノイズの分散がデータに依存する場合，(ii) データ点に重複がある場合に照らして，それぞれ重み付き二乗和誤差関数の解釈を与えよ．

(解答) 行列 $\mathbf{R}$ を $\mathbf{R} = \mathrm{diag}(r_1, \ldots, r_N) \in \mathbb{R}^{N\times N}$ とすると，

$$\begin{aligned} E_D(\boldsymbol{w}) &= \frac{1}{2}\sum_{n=1}^{N} r_n \left\{ t_n - \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_n) \right\}^2 = \frac{1}{2}(\mathbf{t} - \boldsymbol{\Phi}\boldsymbol{w})^\top \mathbf{R}(\mathbf{t} - \boldsymbol{\Phi}\boldsymbol{w}) \\ &= \frac{1}{2}(\mathbf{t}^\top - \boldsymbol{w}^\top \boldsymbol{\Phi}^\top)\mathbf{R}(\mathbf{t} - \boldsymbol{\Phi}\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{w}^\top \boldsymbol{\Phi}^\top \mathbf{R}\boldsymbol{\Phi}\boldsymbol{w} - 2\mathbf{t}^\top \mathbf{R}\boldsymbol{\Phi}\boldsymbol{w} + \mathbf{t}^\top \mathbf{R}\mathbf{t}) \end{aligned}$$

ゆえ，

$$\frac{\partial}{\partial \boldsymbol{w}} E_D(\boldsymbol{w}) = \boldsymbol{\Phi}^\top \mathbf{R}\boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{\Phi}^\top \mathbf{R}\mathbf{t} \quad \Longrightarrow \quad \frac{\partial}{\partial \boldsymbol{w}} E_D(\boldsymbol{w})\Big|_{\boldsymbol{w}=\boldsymbol{w}^*} = 0 \iff \boldsymbol{w}^* = \left(\boldsymbol{\Phi}^\top \mathbf{R}\boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^\top \mathbf{R}\mathbf{t}.$$

# 演習問題 (Appendix)

## 演習 3.3 (基本)

それぞれのデータ点 $t_n$ に重み要素 $r_n > 0$ が割り当てられており，二乗和誤差関数が

$$E_D(\boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} r_n \left\{ t_n - \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_n) \right\}^2$$

となるデータ集合を考える．このとき，この誤差関数を最小にする解 $\boldsymbol{w}^*$ についての式を求めよ．また，(i) ノイズの分散がデータに依存する場合，(ii) データ点に重複がある場合に照らして，それぞれ重み付き二乗和誤差関数の解釈を与えよ．

(解答続き) (i) ノイズの分散がデータに依存する場合，式 (2.1) より，重み付き二乗和誤差関数はデータ点 $t_n$ のノイズの分散が $r_n^{-1}$ で与えられている場合の二乗和誤差関数と解釈できる．(ii) データ点に重複がある場合，重み付き二乗和誤差関数はデータ点 $(\boldsymbol{x}_n, t_n)$ が $r_n$ 個重複している場合の二乗和誤差関数と解釈できる．