

## 1 Bayes 理論の目的

Bayes 理論は、パラメータの集合  $W \subset \mathbb{R}^d$ 、真の分布  $q(x)$ 、確率モデル  $p(x|w)$ 、事前分布  $\varphi(w)$  が事前に与えられて構築される理論である。理論の目的は以下のようにまとめられる：

### Bayes 理論の目的

真の分布  $q(x)$ 、確率モデル  $p(x|w)$ 、事前分布  $\varphi(w)$  が与えられたとする。自然数  $n = 1, 2, 3, \dots$  に対して、我々は何らかの処方箋に従って予測分布と呼ばれる確率分布の系列  $\{p_n^*(x)\}_{n=1}^\infty$  を構成する。予測分布の構成手法は Bayes 推測、事後確率最大化法、最尤推定法、平均プラグイン法などが存在する。たとえば Bayes 推測を用いた場合では、予測分布を

$$p_n^*(x) := \int p(x|w)p(w|X^n) dw = \langle p(x|w) \rangle_{w|X^n} \quad (1)$$

と定める。ここで、 $p(w|X^n)$  は事後分布と呼ばれる確率分布であり、真の分布  $q(x)$  から得られる  $n$  個の独立なサンプル  $X^n$ （データセット）を用いて、

$$p(w|X^n) \propto \varphi(w) \prod_{i=1}^n p(X_i|w)^\beta \quad (2)$$

と定義される。Bayes 理論の主な目的は、予測分布の構成手法として Bayes 推測を用いた場合、 $n$  の増大に対して

$$D_{\text{KL}}[q(x) \| p_n^*(x)] = \int q(x) \ln \frac{q(x)}{p_n^*(x)} dx = \mathbb{E}_X \left[ \ln \frac{q(x)}{p_n^*(x)} \right] \quad (3)$$

がどのような振る舞いをするか、すなわちどのように予測分布  $p_n^*(x)$  が真の分布  $q(x)$  に近づいていくかを調べるのが目的である。また、

$$D_{\text{KL}}[q(x) \| p_n^*(x)] = \mathbb{E}_X [\ln q(x)] - \mathbb{E}_X [\ln p_n^*(x)] \quad (4)$$

であり、右辺第一項は真の分布のみで定まる定数である。したがって右辺第二項の振る舞いを調べられれば良い。予測分布を Bayes 推測で構成した場合、

$$-\mathbb{E}_X [\ln p_n^*(x)] = -\mathbb{E}_X \left[ \ln \langle p(X|w) \rangle_{w|X^n} \right] \quad (5)$$

となり、この値は汎化損失  $G_n$  と呼ばれるものになっている：

$$G_n := -\mathbb{E}_X \left[ \ln \langle p(X|w) \rangle_{w|X^n} \right]。 \quad (6)$$

しかし我々は真の分布にアクセスできないため、原理的に  $G_n$  の値を求めることはできない。Bayes 理論の目的を言い換えると、この  $G_n$  の値をいかにして推定するかということになる<sup>a</sup>。特に

- 事後分布による期待値計算を如何にして実行するか（原理的には計算可能だが…？）
- 真の分布による期待値計算をどのような推定量として評価するか

ということが重要である。また、 $G_n$  は出現サンプルに依存した値であることに注意。

<sup>a</sup> 原理的に計算可能な  $T_n$  を推定量として見るのではダメ？  $G_n$  を推定すること自体をしたいわけではない？ たぶん情報量規準はこの疑問と関係している気がする。

事後分布はモデルと事前分布から定まるため、原理的には計算可能である。実用上では、事後分布やモデルが複雑であれば積分計算が実行できないため、積分の収束先からのズレを見るオーダー評価しか行えない。オーダー評価の処方箋を与えているのが正則理論や一般理論と呼ばれる理論である。

(情報量規準についてまだ理解していないので、この段落は自分の予想) 真の分布による期待値計算は、原理的に何らかの推定量として扱う必要がある。推定量として扱う方法を与えるために、情報量規準と呼ばれる量を導入する。そのため、異なる情報量規準を用いると異なる推定結果となり、評価の意味合いが変わってくる。

## 1.1 いくつかの概念の定義と本書の仮定について

### 真の分布に対して最適なパラメータの集合

パラメータの集合  $W$  に対し、真の分布  $q(x)$  とモデル  $p(x|w)$  の間の KL divergence を最小にするパラメータの集合を  $W_0$  と定義する：

$$W_0 := \{w \in W \mid w = \arg \min_w D_{\text{KL}}[q(x) \parallel p(x|w)]\} \quad (7)$$

この集合のことを、真の分布に対して最適なパラメータの集合と呼ぶ。

### 実質的にユニーク

任意の  $w_0 \in W_0$  について、 $p(x|w_0)$  がユニークな確率分布  $p_0(x)$  を表すとき、真の分布に対して最適な確率分布は実質的にユニークであるという。

以下では実質的にユニークが常に実現されていることを仮定する。渡辺ペイズ本では、より強い仮定として、「相対的に有限な分散を持つ」という状況が達成されていることを仮定する。<sup>\*1</sup>

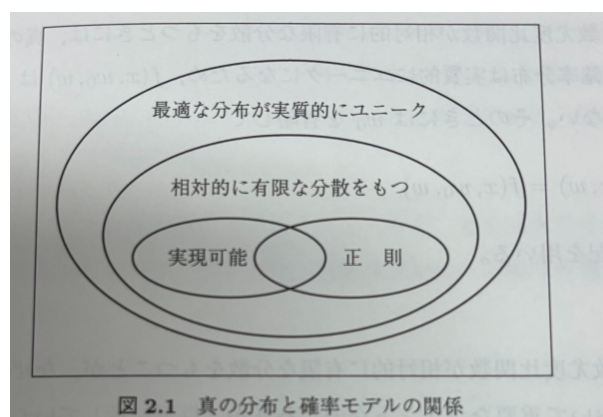


図 1 真の分布と確率モデルに対する関係（渡辺ペイズ p35）

以下の説明は、「相対的に有限な分散を持つ」ことが達成されている、すなわち「最適な分布が実質的にユニーク」な状況を前提にしている。

<sup>\*1</sup> 「相対的に有限な分散を持つ」ことが達成されていない場合、後で定義する重要な量である  $K_n(w)$  の分散がその平均でバウンドできなくなるため、事後分布がサンプルに応じて大きく変動してしまうことが問題となるらしいです。これ以上はよくわかりません。（渡辺ペイズ p36, 注意 12 を参照。）

## 1.2 事後分布の期待値計算を行いたい

本当は、任意のモデルと事前分布の設計に対して

$$\ln \langle p(X|w) \rangle_{w|X^n} \quad (8)$$

の計算を行いたい。しかしそれすら難しいので、 $n$  に対するオーダー評価で我慢する<sup>\*2</sup>。 $p(X|w)$  は  $n$  に依存しない関数なので、事後分布が  $n$  に対してどのような振る舞いをしているかを見ればオーダー評価の意味では十分。すなわち示すべきことは、 $O(x, w)$  を  $n$  に依存しない関数として、

$$\langle O(x, w) \rangle_{w|X^n} := \int O(x, w) p(w|X^n) dw \quad (9)$$

のオーダー評価を行うことである。それを実行するため、 $p(w|X^n)$  を天下りのだが（ラプラス近似を用いたため）次のように変形する：

$$p(w|X^n) \propto \varphi(w) \prod_{i=1}^n p(X_i|w)^\beta \propto \varphi(w) \exp(-n\beta K_n(w)) \quad (10)$$

ここで  $K_n(w)$  は経験誤差関数であり、

$$K_n(w) := \frac{1}{n} \sum_{i=1}^n f(X_i, w) := \frac{1}{n} \sum_{i=1}^n \ln \frac{p(X_i|w_0)}{p(X_i|w)} \quad (11)$$

と定義される。なお、 $f(x, w)$  は対数尤度比関数と呼ばれ、 $K_n(w)$  は  $f(x, w)$  のサンプル平均である。この変形を用いると、

$$\langle O(x, w) \rangle_{w|X^n} = \frac{\int_W O(x, w) \exp(-n\beta K_n(w)) \varphi(w) dw}{\int_W \exp(-n\beta K_n(w)) \varphi(w) dw} \quad (12)$$

となる。このように変形することで、事後分布の確率的な振る舞いを  $K_n(w)$  に押し付けることができた。さらに指数関数の肩が  $-n$  に比例しているため、 $n$  が十分に大きな場合、式 (12) の積分計算に寄与する領域は  $K_n(w)$  の最小値付近のみと考えることができる。したがって、

1.  $K_n(w)$  の最小値を取る点が、パラメータ空間  $W$  全体でどのような分布をしているか
2.  $K_n(w)$  の最小値周りにおける関数の局所的な振る舞い

を調べることが重要である。 $K_n(w)$  の  $w$  に対する振る舞い方は、真の分布とモデル設計に依存して決まるため、第三章では正則理論という枠組みの元で  $K_n(w)$  の振る舞いを考える。なお、 $n \rightarrow \infty$  極限における  $K_n(w)$  の収束先を平均誤差関数  $K(w)$  と呼ぶ：

$$K_n(w) \rightarrow K(w) := \mathbb{E}_X[f(X, w)] = \int f(x, w) q(x) dx \quad (13)$$

$K(w)$  は真の分布  $q(x)$  とモデル設計方法  $p(x|w)$  に依存して決まる関数である。

---

<sup>\*2</sup> オーダー評価をしようと思うと、原理的に計算できない量  $K(w)$  を出現させる必要がある。これは収束先からのズレを見たいためしょうがないのか？

### 1.3 期待値計算の評価において一般的に成り立つ事実

#### 期待値計算における積分の分割

事後分布による期待値計算は以下の形で表される：

$$\langle O(x, w) \rangle_{w|X^n} = \frac{\int_W O(x, w) \exp(-n\beta K_n(w)) \varphi(w) dw}{\int_W \exp(-n\beta K_n(w)) \varphi(w) dw}。 \quad (14)$$

分母は分子の積分において  $O(x, w) = 1$  を代入した形であるため、分子の  $n$  に対する振る舞いを見れば全体の振る舞いがわかる。分子の積分領域を2つに分割し、 $n$  に対する主要項（第一項目）、非主要項（第二項目）へと分割する：<sup>a b</sup>

$$\int_{K(w) < \epsilon_n} O(x, w) \exp(-n\beta K_n(w)) \varphi(w) dw + \int_{K(w) \geq \epsilon_n} O(x, w) \exp(-n\beta K_n(w)) \varphi(w) dw。 \quad (15)$$

ただし、値  $\epsilon_n > 0$  は  $n$  の単調減少関数であり、

$$\lim_{n \rightarrow \infty} \epsilon_n = 0, \quad \lim_{n \rightarrow \infty} \sqrt{n} \epsilon_n = \infty \quad (16)$$

を満たすように選ぶ。このとき、非主要項（第二項目）は正則理論の仮定なしで、

$$\int_{K(w) \geq \epsilon_n} O(x, w) \exp(-n\beta K_n(w)) \varphi(w) dw = o_p(\exp(-\sqrt{n})) \quad (17)$$

になることが示される。この事実を示すために必要なのは  $K_n(w)$  の  $n$  に対する振る舞いの調査のみであり、正則理論の仮定は必要なく常に成り立つ。したがって、

$$\langle O(x, w) \rangle_{w|X^n} = \frac{\int_{K(w) < \epsilon_n} O(x, w) \exp(-n\beta K_n(w)) \varphi(w) dw}{\int_{K(w) < \epsilon_n} \exp(-n\beta K_n(w)) \varphi(w) dw} + o_p(\exp(-\sqrt{n}))。 \quad (18)$$

が常に成立する。なお、 $K(w)$  ができてきているので、計算できない量になっているのは注意。

<sup>a</sup> 主要項、非主要項という言葉は、各項のオーダー評価が意味をなすくらい  $n$  が大きな領域ということを前提に使っている。また、二項目が非主要項であることは、第一項目の評価を行って初めて分かる事実ではあるが、それは認めてください。

<sup>b</sup>  $O(x, w)$  は  $n$  に依存していない関数であることを前提としており、これが  $n$  に依存していればオーダー評価も変わってくる。

以降は主要項

$$\int_{K(w) < \epsilon_n} O(x, w) \exp(-n\beta K_n(w)) \varphi(w) dw \quad (19)$$

のオーダー評価を行う。 $K(w)$  に性質の良い仮定を課してオーダー評価を行っているのが正則理論（第三章）であり、そうでない場合は一般理論（第四章）である。<sup>\*3</sup>

<sup>\*3</sup>  $K_n(w)$  ではなく  $K(w)$  に仮定を課していることに注意。正則理論の枠組みでは、 $K(w) < \epsilon_n$  の領域内部において

$$\exp(-n\beta K_n(w)) \propto \mathcal{N}(w_0 + \hat{\xi}_n / \sqrt{n}, (n\beta J(w_0))^{-1}) (1 + o_p(1)) \quad (20)$$

と表される。なお、 $J(w_0)$  は正定値行列で、 $J(w) = \nabla^2 K(w)$  と定義されている。また、 $\hat{\xi}_n$  は  $K_n(w_0)$  の確率的な振る舞いを担う確率変数である。積分領域はこの分布の広がりよりも十分に大きく、 $\mathbb{R}^d$  全体での積分に置き換えて Gauss 積分が適用できる。

## 2 正則理論

### 2.1 正則理論での仮定

正則理論とは、真の分布とモデル設計（と、サンプルの数）が、以下の3つの条件を満たすことを仮定する理論である<sup>\*4</sup>：

1.  $K(w)$  を最小にするパラメータが唯一  $w_0$  である。これは、最適なパラメータの集合が  $W_0 = \{w_0\}$  と表されることと同じ意味である。
2.  $K(w)$  の Hessian： $J(w) = \nabla^2 K(w)$  に対し、 $J := J(w_0)$  が正定値である。
3. サンプルの数  $n$  が十分に大きいこと。

これらの仮定によって、

- 主要項の積分範囲を考える際には  $w = w_0$  の近傍のみでよくなる。
- $w = w_0$  近傍で主要項の被積分関数は Gaussian として扱える。

という嬉しい性質が成り立つ。結論だけ述べると、この仮定の元で主要項のオーダー評価は次のようになる。

#### 期待値のオーダー評価（正則理論の場合）

正則理論という仮定を課している場合、以下のオーダー評価が成立する：

$$\langle O(x, w) \rangle_{w|X^n} = \frac{\int_{\mathbb{R}_d} O(x, w) \exp\left(-\frac{n\beta}{2} \left\| J^{1/2} \left( w - w_0 - \frac{\xi_n}{\sqrt{n}} \right) \right\|^2\right) dw}{\int_{\mathbb{R}_d} \exp\left(-\frac{n\beta}{2} \left\| J^{1/2} \left( w - w_0 - \frac{\xi_n}{\sqrt{n}} \right) \right\|^2\right) dw} (1 + o_p(1)). \quad (23)$$

2章の残りでは、この計算途中で大事だと思われる式変形について説明する。

<sup>\*4</sup> 渡辺ペイズ本では  $K(w)$  ではなく  $L(w)$  についての仮定をしていたが、これらは基準点を変えただけであり、

$$L(w) = L(w_0) + K(w) \quad (21)$$

によって結びつけられるため、 $K(w)$  に関する仮定として考えても良い。なお、 $L(w)$  は平均対数損失であり、

$$L(w) = -\mathbb{E}_X [\ln p(X|w)] \quad (22)$$

と定義される。

## 2.2 正則理論で用いる文字の定義

正則理論の仮定のもとでの主要項の振る舞いの結果を述べる前に、少しだけ文字の定義を行う。まずは  $K_n(w)$  を次のように変形する：

$$K_n(w) = K(w) - (K(w) - K_n(w)) = K(w) - \frac{1}{\sqrt{n}}\eta_n(w)。 \quad (24)$$

ここで  $\eta_n(w)$  は  $K_n(w)$  の確率的な振る舞いを担う確率過程であり、次のように定義される：

$$\eta_n(w) := \frac{1}{\sqrt{n}} \sum_{i=1}^n (K(w) - f(X_i, w))。 \quad (25)$$

この確率過程は経験過程と呼ばれるものであり、 $n \rightarrow \infty$  でとある確率過程に法則収束するという良い性質を持っている。この性質があることで、収束先の確率分布からのズレとして  $K_n(w)$  をオーダー評価することが可能になる。必要となる確率変数  $\xi_n, \hat{\xi}_n$  を

$$\xi_n = J^{-1/2}(w_0) \nabla \eta_n(w_0)、 \quad (26)$$

$$\hat{\xi}_n = J^{-1/2}(w_0) \xi_n = J^{-1} \nabla \eta_n(w_0) \quad (27)$$

と定義する。 $\nabla \eta_n(w)$  がすべての  $w$  で  $\mathcal{N}(0, I(w))$  に法則収束するという性質を持っているため、

- 確率変数  $\xi_n$  は正規分布  $\mathcal{N}(0, J^{-1/2} I J^{-1/2})$  に法則収束する。
- 確率変数  $\hat{\xi}_n$  は正規分布  $\mathcal{N}(0, J^{-1} I J^{-1})$  に法則収束する。

ここで  $I = I(w_0)$  であり、 $I(w)$  は以下で定義される：

$$I(w) := \mathbb{E}_X [\nabla f(X, w) (\nabla f(X, w))^\top] - \nabla K(w) (\nabla K(w))^\top。 \quad (28)$$

## 2.3 正則理論における主要項のオーダー評価

正則理論の仮定のもとでは、 $w = w_0$  近傍の  $K_n(w)$  の振る舞いを見れば良い。 $K_n(w)$  を平方完成して Gauss 積分として扱いたいため、平均値の定理を利用して  $w = w_0$  周りの展開を考える。そのため

$$K_n(w) = K(w) - \frac{1}{\sqrt{n}} \eta_n(w) \quad (29)$$

の右辺の項にそれぞれ平均値の定理を用いると、 $w$  により定まる 2 つの値  $w^*, w^{**}$  が存在し、

$$K(w) = \frac{1}{2} (w - w_0) \cdot J(w^*) (w - w_0), \quad (30)$$

$$\eta_n(w) = (w - w_0) \cdot \nabla \eta_n(w^{**})。 \quad (31)$$

が成立する。 $K(w)$  の一次項が出てこないのは、 $\nabla K(w_0) = 0$  が成立しているためである。これらの結果から、 $K_n(w)$  は  $w$  の二次式として表される：

$$K_n(w) = \frac{1}{2} (w - w_0) \cdot J(w^*) (w - w_0) - \frac{1}{\sqrt{n}} (w - w_0) \cdot \nabla \eta_n(w^{**})。 \quad (32)$$

なお、平方完成を行うためには  $J(w^*)$  が正則（対角化可能）であることが必要である。渡辺ペイズ本では、 $w^*$  が  $w_0$  近傍では  $J(w^*)$  が常に正則ということを仮定しているが、その妥当性がどこから来ているのかは不明。正則であることを認めたうえで、 $K_n(w)$  は

$$nK_n(w) = \frac{n}{2} \left\| J(w^*)^{1/2} \left( w - w_0 - \frac{1}{\sqrt{n}} J(w^*)^{-1} \nabla \eta_n(w^{**}) \right) \right\|^2 - \frac{1}{2} \left\| J(w^*)^{-1/2} \nabla \eta_n(w^{**}) \right\|^2 \quad (33)$$

と平方完成できる。

また、 $n \rightarrow \infty$  のとき  $\epsilon_n \rightarrow 0$  であり、正則理論の仮定 1 のおかげで、 $w^*, w^{**} \rightarrow w_0$  が成立する。したがって一般の  $n$  ではそこからのズレとして

$$J(w^*) = J + o_p(1), \quad \nabla \eta_n(w^{**}) = \nabla \eta_n(w_0) + o_p(1) \quad (34)$$

が成立する。このオーダー評価の結果を用いると、主要項は次のように表せる：

$$\begin{aligned} & \int_{K(w) < \epsilon_n} O(x, w) \exp(-n\beta K_n(w)) \varphi(w) \, dw \\ &= \exp\left(\frac{\beta}{2} \|\xi_n\|^2\right) \varphi(w_0) \int_{K(w) < \epsilon_n} O(x, w) \exp\left(-\frac{n\beta}{2} \left\| J^{1/2} \left( w - w_0 - \frac{\hat{\xi}_n}{\sqrt{n}} \right) \right\|^2\right) \, dw (1 + o_p(1)) \\ &= \exp\left(\frac{\beta}{2} \|\xi_n\|^2\right) \varphi(w_0) \int_{\mathbb{R}^d} O(x, w) \exp\left(-\frac{n\beta}{2} \left\| J^{1/2} \left( w - w_0 - \frac{\hat{\xi}_n}{\sqrt{n}} \right) \right\|^2\right) \, dw (1 + o_p(1))。 \end{aligned} \quad (35)$$

最後の等式は、積分範囲を  $\mathbb{R}^d$  全体に広げてよいことを利用している。これにより、先ほど提示したオーダー評価の形が得られる：

#### 期待値のオーダー評価（正則理論の場合）

正則理論という仮定を課している場合、以下のオーダー評価が成立する：

$$\langle O(x, w) \rangle_{w|X^n} = \frac{\int_{\mathbb{R}^d} O(x, w) \exp\left(-\frac{n\beta}{2} \left\| J^{1/2} \left( w - w_0 - \frac{\hat{\xi}_n}{\sqrt{n}} \right) \right\|^2\right) \, dw}{\int_{\mathbb{R}^d} \exp\left(-\frac{n\beta}{2} \left\| J^{1/2} \left( w - w_0 - \frac{\hat{\xi}_n}{\sqrt{n}} \right) \right\|^2\right) \, dw} (1 + o_p(1))。 \quad (36)$$

### 3 正則理論における汎化損失の振る舞い

正則理論においては、事後分布による期待値計算のオーダー評価が可能になった。その結果を用いると  $G_n$  は

$$G_n = -\mathbb{E}_X \left[ \ln \langle p(X|w) \rangle_{w|X^n} \right] = L(w_0) + \frac{1}{n} \left( \frac{d}{2\beta} + \frac{1}{2} \|\xi_n\|^2 - \frac{1}{2\beta} \text{tr}\{IJ^{-1}\} \right) + o_p\left(\frac{1}{n}\right) \quad (37)$$

となる。これには原理的に計算が不可能な量が入っており、これは

- 真の分布による期待値を外から取ったから
- 事後分布の期待値計算をオーダー評価するために収束先の値を見たから

という 2 つの理由がある。真の分布をしらないと計算不可能な量は

$$L(w_0) = -\mathbb{E}_X [\ln p(X|w_0)], \quad (38)$$

$$\xi_n = J^{-1/2} \nabla \eta_n(w_0), \quad (39)$$

$$I = I(w_0) = \mathbb{E}_X [\nabla f(X, w_0) (\nabla f(X, w_0))^\top] - \nabla K(w_0) (\nabla K(w_0))^\top, \quad (40)$$

$$J = J(w_0) = \nabla^2 K(w_0) \quad (41)$$

である。また、 $w_0$  の値も真の分布をしらないと求められない。いったいこれからどうするのか。ちなみに経験損失  $T_n$  は原理的に計算できる：

$$T_n = -\frac{1}{n} \sum_{i=1}^n \ln \langle p(X_i|w) \rangle_{w|X^n}. \quad (42)$$

経験損失のオーダー評価は

$$T_n = L_n(w_0) + \frac{1}{n} \left( \frac{d}{2\beta} - \frac{1}{2} \|\xi_n\|^2 - \frac{1}{2\beta} \text{tr}\{IJ^{-1}\} \right) + o_p\left(\frac{1}{n}\right) \quad (43)$$

となる。

$$L_n(w_0) = -\frac{1}{n} \sum_{i=1}^n \ln p(X_i|w_0), \quad (44)$$

$$(45)$$

は  $w_0$  を別の量に置き換えれば計算可能ではある。また、サンプルによる期待値を取ると、

$$\mathbb{E}[G_n] = L(w_0) + \frac{1}{n} \left( \frac{\lambda - \nu}{\beta} + \nu \right) + o\left(\frac{1}{n}\right), \quad (46)$$

$$\mathbb{E}[T_n] = L(w_0) + \frac{1}{n} \left( \frac{\lambda - \nu}{\beta} - \nu \right) + o\left(\frac{1}{n}\right) \quad (47)$$

と表される。ここで

$$\lambda = \frac{d}{2}, \quad \nu = \frac{1}{2} \text{tr}\{IJ^{-1}\} \quad (48)$$

である。



## 4 Bayes 推測以外の方法で予測分布を構成する

予測分布の構成方法として Bayes 推測以外の手法を用いた場合にどのようなになるかを説明する。特に、最尤推定法と事後確率最大化法（MAP 推定法）が統一的に説明できることを述べる。これらはパラメータを点推定するような方法である。まず、ハイパーパラメータ  $\beta > 0$  に依存する関数

$$\mathcal{L}_n^\beta(w) := -\frac{1}{n} \sum_{i=1}^n \ln p(X_i|w) - \frac{1}{n\beta} \ln \varphi(w) \quad (49)$$

を最小にするパラメータを  $\hat{w}_n^\beta$  とおく。このようなパラメータが唯一とは限らない。このとき、予測分布を

$$p_n^*(x) = p(X|\hat{w}_n^\beta) \quad (50)$$

と定義する。 $\beta \rightarrow \infty$  とした場合が最尤推定<sup>\*5</sup>であり、 $\beta = 1$  とした場合が事後確率最大化法（MAP 推定）に対応する。すなわち  $\beta$  は事前分布と学習データをどれくらい重視するかを調整する意味を持っている。

### 推定量の一致性

式 (7) で定義した集合  $W_0$  の元が唯一  $w_0$  であるとき、任意の  $\beta$  について、 $n \rightarrow \infty$  のとき  $K(\hat{w}_n^\beta)$  の値は 0 に確率収束する。また、正則理論を仮定していれば、 $\hat{w}_n^\beta \rightarrow w_0$  に確率収束する。

これらの方法で得られた  $\hat{w}_n^\beta$  を用いて、汎化損失と経験損失は

$$L(\hat{w}_n^\beta) = -\mathbb{E}_X[\ln p(X|\hat{w}_n^\beta)] \quad (51)$$

$$L_n(\hat{w}_n^\beta) = -\frac{1}{n} \sum_{i=1}^n [\ln p(X_i|\hat{w}_n^\beta)] \quad (52)$$

と定義される。

<sup>\*5</sup> 最尤推定はデータを完全に信じてモデルの対数尤度を最大化するパラメータを選択するような推定方法であるため過学習を起こしやすい？

## 5 情報量規準

この章は自分の中でその意味を捉えられていないまま書いています。

- $G_n$  そのものを情報量規準で推定したいわけではない？
- $G_n$  のオーダー評価の結果を用いて計算できない量は微量として扱えるようにし、主要項は計算できる量で推定したい？
- $E[G_n]$  の意味で、主要項は計算できる量で推定したい？

情報量規準として BIC, RIC, TIC, AIC がある。これらの定義を以下で述べる。前の輪読では AIC は大事な量として扱ってなかったが、大事な量として扱う必要があると思う。TIC は大事ではない。

$G_n$  を「データセットの平均として」推定したいのが RIC で、AIC、BIC のような直感的な理解はできていない。

先に述べておくと、AIC と BIC の使い分け方について、ふわっとした自分の理解としては、

- AIC：モデルがどれほど新しいデータセットに適合しているかを評価しているかを見ている。AIC では単純なモデルを選びたいという気持ちはあまり大きくない。
- BIC：同じくらいの説明能力があるモデルであれば、よりパラメータが少ないモデルを選択する。特に、モデルが真の分布を再現できるもの同士を比較するときは、よりパラメータが少ない（説明変数が少ない）モデルを選択する。これはパラメータによるペナルティ項の影響が大きいという数式から理解できそう。

以上はネットで調べた情報をもとに自分が理解したことで、この本の数式から正しくこの解釈が出てくるかはあまり見ていない。（なんとなくはこの解釈で良さそうな雰囲気。）以下は軽く調べて参考になりそうなもの

- [AIC（赤池情報量基準）に関する解説 - Statistics School（実用的な理解をするためならこれで十分そう）](#)
- [AIC と MDL と BIC, 赤池さんの記事](#)

- データ解析のための統計モデリング入門(久保拓弥) 以下、緑本  
独特な語り口調、独特な構成の本ですが、すばらしい本です。特にAICに関する説明(第4章)は圧巻です。初心者はどうしても(モデルの)サンプルへの当てはまりに目が行きがちですが、それは意味のないことであって、あくまでも母集団の推定をすべきだということが分かりやすく丁寧に説明されています。ぜひ読むべきです。また、Rを用いて分析されていて、そのコードも隅々まで公開されています。Rの演習という意味でも利用価値のある本です。  
さらに、著者本人による解説動画もありますので、併せてご覧ください。

図2 Qiitaの記事の一節。統計検定準1級 合格体験記

## 5.1 BIC

BIC の定義は次である：

$$\text{BIC} := - \sum_{i=1}^n \ln p(X_i | \hat{w}_n^\beta) + \frac{d}{2} \ln n \quad (53)$$

複数のモデルの比較において BIC が小さいほど適切であると考えるとき、この値のことをベイズ情報量規準という。(対数尤度が大きく、パラメータが少ないモデルを選びたい気持ち。) BIC が他の指標と大きく違うのは、パラメータによる正則化項が  $n$  に比例していることである。

BIC は、真の分布は単純なモデルで表現できるということがわかっているときに使うような指標だと思った。あとで述べる AIC は「単純なモデルを選ぶ」ということにはあまり関心がなさそうだった。(この理解がただしそうかを後で考えたい)

### BIC について定性的に理解する

解釈をしやすくするために  $d$  が同じモデルを 2 つ持ってきて、 $\beta$  も固定する。データセットも同じものを用意する。事前分布も同じにする。もっともらしいパラメータ  $\hat{w}_n^\beta$  を決めたとき、BIC が小さいほうが適切ということは、予測分布  $p_n^*(x)$  に関する対数尤度

$$\sum_{i=1}^n \ln p_n^*(X_i) \quad (54)$$

が大きいモデルのほうが適切であるということを意味する。これは最尤推定法を擬似的に利用したいということだと思った。また、表現能力が必要以上に高すぎると過学習のおそれがあるので、同じくらいの表現能力を持っているモデルでは、よりパラメータが少ないモデルを選択している。これ以上の意味はわからない。自由エネルギーの主要項が BIC そのものらしいので、自由エネルギーの意味を理解しないといけない気がする。

## 5.2 RIC

RIC の定義は以下である：

$$\text{RIC} := T_n + \frac{1}{n} \text{tr}\{IJ^{-1}\}。 \quad (55)$$

これは計算できない量である。しかし、RIC の  $1/n$  よりも大きなオーダー項はすべて計算できる量としてまとめられる：

$$\text{RIC} = L_n(\hat{w}_n^\beta) + \frac{1}{n} \left( \frac{d}{2\beta} + \left( 1 - \frac{1}{2\beta} \text{tr}\{I_n(\hat{w}_n^\beta)J_n(\hat{w}_n^\beta)^{-1}\} \right) \right) + o_p\left(\frac{1}{n}\right)。 \quad (56)$$

ここで計算できる新たな量として

$$I_n(w) := \frac{1}{n} \sum_{i=1}^n \nabla \ln p(X_i|w) (\nabla \ln p(X_i|w))^\top \quad (57)$$

$$J_n(w) := -\frac{1}{n} \sum_{i=1}^n \nabla^2 \ln p(X_i|w) \quad (58)$$

$$(59)$$

を定義した。この量は、大数の法則と  $\hat{w}_n^\beta \rightarrow w_0$  を用いて

$$I(w_0) = I_n(\hat{w}_n^\beta) + o_p(1), \quad J(w_0) = J_n(\hat{w}_n^\beta) + o_p(1) \quad (60)$$

となることが示せるらしい。

RIC では常に

$$\mathbb{E}[G_n] = \mathbb{E}[\text{RIC}] + o\left(\frac{1}{n}\right) \quad (61)$$

が成立する。よって  $1/n$  より大きなオーダーにおいて、汎化損失  $G_n$  のサンプル期待値を推定したいのが目的な気がする。本当はチューニングしきったモデルの中で  $G_n$  が小さくなる方がどちらであるかを見たいんだけど、それが難しくてサンプル期待値の意味でしか見れないということ。

## 5.3 TIC

TIC の定義は以下である：

$$\text{TIC} := L_n(\hat{w}_n^\beta) + \frac{1}{n} \text{tr}\{I_n(\hat{w}_n^\beta)J_n^{-1}(\hat{w}_n^\beta)\}。 \quad (62)$$

これは Bayes 以外の推測方法の場合に適用できる RIC というような立ち位置だと思う。実際に

$$\mathbb{E}[L(\hat{w}_n^\beta)] = \mathbb{E}[\text{TIC}] + o\left(\frac{1}{n}\right) \quad (63)$$

が成立する。

## 5.4 AIC

AIC の定義は以下である：

$$\text{AIC} := L_n(\hat{w}_n^\beta) + \frac{d}{n} = -\frac{1}{n} \sum_{i=1}^n [\ln p(X|\hat{w}_n^\beta)] + \frac{d}{n}。 \quad (64)$$

この量は、モデルが真の分布が実現可能な場合、

$$\text{RIC} = \text{AIC} + o_p\left(\frac{1}{n}\right) \quad (65)$$

が成立する。これは  $\text{tr}\{IJ^{-1}\} = d$ 、すなわち  $\text{tr}\{I_n(\hat{w}_n^\beta)J_n^{-1}(\hat{w}_n^\beta)\} = d + o_p(1/n)$  であるため、しかし、実現可能でない場合はこの関係式が成り立たない。実現可能でない場合は RIC と AIC は異なる量であるため、AIC も大事な量だと思う。

### AIC と BIC の意味合い（注意 31 の意味を考えてみる。）

AIC を規準と使う場合、サンプルの現れ方の揺れに応じて選択されるモデルが変動しやすいらしい。すなわち、AIC は感度が高い（データセットの与え方に依存して判定が変わってしまう）ような指標である。

たくさんのモデルを用意しておいて、AIC を規準として採用した場合を考える。AIC を使って生き残ったモデルは、どのようなデータセットが与えられても常に AIC の意味で良いモデルという意味合いを持つ。これは、データセットをあまり信用しておらず、堅牢なモデルを選びたいときに使う気がする。

その一方で、BIC は AIC と比べて感度が低い。つまりデータセットを変化させても選択されるモデルは変わりにくい。仮にデータセットが信頼できない場合（データセットに外部からのノイズが乗っているようなイメージのとき）に BIC を使ってしまうと、ノイズが乗ったデータセットを正しいと思い込んだ上でモデルが選択されてしまう。

以上の解釈が正しいかどうかを chatGPT に聞いてみた。

#### ✅ あなたの主張に対する整理

##### 1. AICは「感度が高い」指標で、データの揺れによって選ばれるモデルが変わりやすい

- ✅ 正しい理解です。
- AICは汎化性能を重視しているため、サンプルが少し違えば「より予測がうまくいきそうなモデル」に乗り換えることが多い。
- よって、データに忠実な判断をする反面、サンプル依存性（ばらつき）に敏感になります。

##### 2. AICで生き残ったモデルは「どんなデータでも良いモデル」という意味になる

- ❌ やや誤解があります。
- AICで選ばれたモデルは、\*\*「与えられたデータに基づいて、将来のデータに対する予測性能が高いと見込まれるモデル」\*\*です。
- したがって、異なるデータセットが与えられれば、選ばれるモデルも変わることがあります。
- 「常にAICの意味で良いモデル」というのは誤りで、AICの判断は与えられたデータに強く依存します。

##### 3. BICは感度が低く、モデル選択がデータ変更に対して安定する

- ✅ 基本的に正しい理解です。
- BICはモデルの複雑さに対してより厳しいペナルティを課すため、複雑なモデルを選びにくく、モデル選択の一貫性が高い。

##### 4. BICは「ノイズのあるデータ」に対して過信してしまい、正しくないモデルを選ぶ可能性がある

- ✅ この指摘も良い感覚です。
- BICは真のモデルに近いものを見つけようとする一方で、現在のデータを“真の分布”と仮定する性質があります。
- したがって、もしデータに外れ値やノイズが多ければ、その「間違っただけのデータを正しい」と仮定した上で判断するため、頑健性に欠ける可能性があります。

図 3 ChatGPT に AIC と BIC の使い分けについて聞いてみた結果。

## 6 以下議論用