統計科学
状態空間モデルを用いた時系列データの解析

松田孟留
◎理化学研究所脳神経科学研究センター

近年，さまざまな分野で大規模なデータが得られるようになってきています．統計科学はデータをもとに現象の理解や予測を行うための方法を研究する学問で，さまざまな数学が登場します．統計科学では，データは確率変数の実現値であるとみなし，その確率分布を「統計モデル」によって表現します．本稿では，時間とともに変化する現象の解析によく用いられる統計モデルである「状態空間モデル」を例として，統計科学の考え方を紹介します．

1……条件つき確率とベイズの公式

まず確率の用語を確認しておきます．事象 A が起きる確率を $P(A)$ とします．たとえば，サイコロの目を表す確率変数を X とすると，

$$P(X=k)=\frac{1}{6} \qquad (k=1,\cdots,6)$$

です．事象 A が起きたという条件のもとでの事象 B の確率 (**条件つき確率**) は

$$P(B \mid A)=\frac{P(A\cap B)}{P(A)}$$

と定義されます．たとえば,「サイコロの目が偶数である」という条件のもとで「サイコロの目が 4 以上である」確率は

$$\begin{aligned} &P(X\geqq 4 \mid X\in\{2,4,6\}) \\ &\quad=\frac{P(X\in\{4,6\})}{P(X\in\{2,4,6\})}=\frac{1/3}{1/2}=\frac{2}{3} \end{aligned}$$

となります．2 つの確率変数 X と Y について，$p(x \mid y)=P(X=x \mid Y=y)$ のように書くことにすると，条件つき確率の定義より

$$p(x \mid y)=\frac{p(x)p(y \mid x)}{p(y)}=\frac{p(x)p(y \mid x)}{\sum_{x'}p(x')p(y \mid x')}$$

が成り立ち，**ベイズの公式**とよばれます．ここで，$\sum_{x'}$ は X のとりうる値すべてにわたる和です．この式は $Y = y$ という情報をもとに X の確率分布を更新したと解釈できるので，$p(x)$ は**事前分布**，$p(x \mid y)$ は**事後分布**とよばれます．ベイズの公式をもとに未知の量に関する推論を行う枠組みを**ベイズ統計学**といい，さまざまな分野で応用されています．

2……壺の逐次ベイズ推定

ベイズの公式の応用例として次の問題を考えます．

> 黒玉と白玉がたくさん入った壺が 2 つある．壺 A には黒玉と白玉が 2 : 1 の割合で入っていて，壺 B には黒玉と白玉が 1 : 2 の割合で入っている．いま P さんが無作為に壺を選び，そこから 3 つの玉を順に取り出したところ，その色は
>
> 黒，黒，白
>
> であった．このとき，P さんが選んだ壺が壺 A である確率を求めよ．

壺の種類を X，t 回目に取り出した玉の色を Y_t とします．問題の設定を条件つき確率で表現すると，

$$P(Y_t = \text{黒} \mid X = A) = \frac{2}{3}$$

などとなります．求める確率は条件つき確率

$$P(X = A \mid Y_1 = \text{黒}, Y_2 = \text{黒}, Y_3 = \text{白})$$

です．これを直接計算することも可能なのですが，ここでは後の拡張に備えて**逐次ベイズ推定**という考え方で問題を解いてみます．$X = A$ である確率を玉を取り出す (情報が得られる) ごとに更新していくイ

メージです.

まず，玉を取り出す前の時点では，壺を無作為に選んだことから

$$P(X = A) = \frac{1}{2}$$

です．そして，1 つ目の玉 (黒) を取り出した時点での確率 $P(X = A \mid Y_1 = \text{黒})$ は，ベイズの公式より

$$\frac{1/2 \cdot 2/3}{1/2 \cdot 2/3 + 1/2 \cdot 1/3} = \frac{2}{3}$$

と得られます．これは，1 つ目の玉の色の情報をもとに事前分布 $p(x)$ を事後分布 $p(x \mid y_1)$ に更新したと解釈できます．このように捉えると，今度は $p(x \mid y_1)$ を事前分布とみなすことで事後分布 $p(x \mid y_1, y_2)$ を

$$p(x \mid y_1, y_2) = \frac{p(x \mid y_1)p(y_2 \mid x, y_1)}{\sum_{x'} p(x' \mid y_1)p(y_2 \mid x', y_1)}$$

と計算できます．いま $p(y_2 \mid x, y_1) = p(y_2 \mid x)$ なので，2 つ目の玉 (黒) を取り出した時点での確率 $P(X = A \mid Y_1 = \text{黒}, Y_2 = \text{黒})$ は

$$\frac{2/3 \cdot 2/3}{2/3 \cdot 2/3 + 1/3 \cdot 1/3} = \frac{4}{5}$$

となります．2 回連続で黒が出たことから壺 A である確率が高そうですが，実際

$$\begin{aligned} P(X = A) &< P(X = A \mid Y_1 = \text{黒}) \\ &< P(X = A \mid Y_1 = \text{黒}, Y_2 = \text{黒}) \end{aligned}$$

となっています．同様に計算することで，3 つ目の玉 (白) を取り出した時点での確率 $P(X = A \mid Y_1 = \text{黒}, Y_2 = \text{黒}, Y_3 = \text{白})$ は

$$\frac{4/5 \cdot 1/3}{4/5 \cdot 1/3 + 1/5 \cdot 2/3} = \frac{2}{3}$$

となります．これで，求める確率は 2/3 であることがわかりました.

3……入れ替わる壺の逐次ベイズ推定

次に，推定対象である壺の種類が時間とともに変化しうる場合を考えます．これは状態空間モデルの簡単な例になっています.

今度は Q さんが無作為に壺を選び，3 つの玉を取り出した．ただし Q さんは玉を 1 つ取り出す度に 1/3 の確率で気が変わって壺を変更する．3 つの玉の色は順に

黒，黒，白

であった．このとき，Q さんが 3 つ目の玉を取り出した壺が壺 A である確率を求めよ．

Q さんが t 回目に玉を取り出した壺の種類を X_t，t 回目に取り出した玉の色を Y_t とします．問題の設定から X_t の時間変化は図 1 のようなグラフで表現できます．矢印についた値は遷移確率で，たとえば

$$P(X_{t+1} = A \mid X_t = A) = \frac{2}{3}$$

です．このような確率変数の列 X_t を**マルコフ連鎖**といいます．求める確率は

$$P(X_3 = A \mid Y_1 = \text{黒}, Y_2 = \text{黒}, Y_3 = \text{白})$$

です．

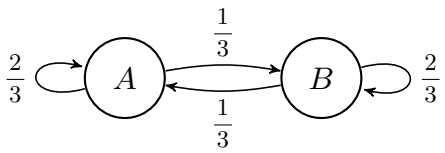

図 1　マルコフ連鎖の状態遷移図

この問題も玉を取り出すごとに確率を更新していくことで解けます．ただし，前節では

$$p(x) \to p(x \mid y_1) \to p(x \mid y_1, y_2) \to p(x \mid y_1, y_2, y_3)$$

という順番で計算したのに対して，今回は

$$\begin{aligned} p(x_1) &\to p(x_1 \mid y_1) \\ &\to p(x_2 \mid y_1) \to p(x_2 \mid y_1, y_2) \\ &\to p(x_3 \mid y_1, y_2) \to p(x_3 \mid y_1, y_2, y_3) \end{aligned}$$

という順番で計算します．まず，前節と同じ議論により $P(X_1 = A) = 1/2$，$P(X_1 = A \mid Y_1 = \text{黒}) = 2/3$ です．すると，X_t がマルコフ連鎖であることから，

$p(x_2 \mid y_1)$ が

$$
\begin{aligned}
&P(X_2 = A \mid Y_1 = \text{黒}) \\
&\quad = \sum_{x_1} P(X_1 = x_1 \mid Y_1 = \text{黒}) P(X_2 = A \mid X_1 = x_1) \\
&\quad = \frac{2}{3} \cdot \frac{2}{3} + \frac{1}{3} \cdot \frac{1}{3} = \frac{5}{9}
\end{aligned}
$$

と計算できます．これを事前分布としてベイズの公式を用いることで，$p(x_2 \mid y_1, y_2)$ が

$$
\begin{aligned}
&P(X_2 = A \mid Y_1 = \text{黒}, Y_2 = \text{黒}) \\
&\quad = \frac{5/9 \cdot 2/3}{5/9 \cdot 2/3 + 4/9 \cdot 1/3} = \frac{5}{7}
\end{aligned}
$$

と得られます．同様に計算していくと，

$$
\begin{aligned}
&P(X_3 = A \mid Y_1 = \text{黒}, Y_2 = \text{黒}) \\
&\quad = \frac{5}{7} \cdot \frac{2}{3} + \frac{2}{7} \cdot \frac{1}{3} = \frac{4}{7},
\end{aligned}
$$

$$
\begin{aligned}
&P(X_3 = A \mid Y_1 = \text{黒}, Y_2 = \text{黒}, Y_3 = \text{白}) \\
&\quad = \frac{4/7 \cdot 1/3}{4/7 \cdot 1/3 + 3/7 \cdot 2/3} = \frac{2}{5}
\end{aligned}
$$

となります．これで，求める確率は $2/5$ であることがわかりました．

4……状態空間モデル

前節の問題では，玉の色 Y_1, Y_2, Y_3 をもとに壺の種類 X_1, X_2, X_3 を逐次的に推定しました．このように時系列データ Y_t の背後に潜む未知の状態系列 X_t を推測するために用いられるのが**状態空間モデル**です [3]．状態空間モデルは，状態が従うマルコフ過程 $p(x_{t+1} \mid x_t)$(システムモデル) と，各状態で観測データが従う条件つき分布 $p(y_t \mid x_t)$(観測モデル) からなります．前節では状態が有限個 (A か B) の値をとる状態空間モデルを考えていたことになります．このような状態空間モデルは**隠れマルコフモデル**と呼ばれ，音声認識などで広く用いられています．

状態空間モデルでは，条件つき分布 $p(x_s \mid y_1, \cdots, y_t)$ を計算することで時系列データの背後に潜むダイナミクスを読み取ることができます．時間の前後関係によって予測 $(s > t)$，フィルタリング $(s = t)$，スムージング $(s < t)$ と呼ばれます．前節の計算は一

期先予測 ($s = t+1$) とフィルタリングを交互にくり返したことに対応します．ほかにも状態空間モデルの種類に応じていろいろな計算方法 (アルゴリズム) が開発されています [2].

状態空間モデルはさまざまな分野でデータ解析に用いられています [1]．たとえば，天気予報では時々刻々と得られる気象データをもとに大気の状態を推定しています．また，我々の脳は五感を通して得られる情報をもとに外界を逐次的に予測していると考えられます．応用上は，時系列データの特性に合わせて適切な状態空間モデルを構築することが重要です．モデルに含まれる未知のパラメータ (**例**：壺に入った黒玉と白玉の割合) をデータから推定したり，複数のモデルの候補 (**例**：P さんか Q さんか) からデータに当てはまるものを選択することも多いです.

5……YouTuber とサザエさんに挑戦

状態空間モデルを用いたデータ解析の例として，じゃんけんの手の逐次予測を行った結果を紹介します．用いたデータは，YouTuber のヒカキンさんが動画中のじゃんけんコーナーで出した手の時系列データ[1)]とテレビアニメ『サザエさん』のじゃんけんコーナーでサザエさんが出した手の時系列データ[2)]です.

t 回目の手を $Y_t \in \{$ グー, チョキ, パー $\}$ とします．ここでは，t 回目にグー・チョキ・パーを出す確率 (g_t, c_t, p_t) を変換した

$$X_t = \left(\log \frac{g_t}{p_t}, \log \frac{c_t}{p_t} \right)$$

を状態とした状態空間モデルを考えました．この変換は，$g_t > 0, c_t > 0, p_t > 0, g_t + c_t + p_t = 1$ をみたす (g_t, c_t, p_t) と $X_t \in \mathbb{R}^2$ の間の全単射になっています．状態 X_t の従うマルコフ過程はランダムウォーク

$$p(x_{t+1} \mid x_t) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{\|x_{t+1} - x_t\|^2}{2\sigma^2} \right)$$

としました．つまり，X_{t+1} は平均 X_t，分散 σ^2 の正

1) 2014 年 11 月 3 日?2022 年 2 月 4 日の 1659 回分，[4] より.

2) 1991 年 11 月 10 日?2021 年 12 月 26 日の 1518 回分，[5] より.

規分布に従うとモデル化しました．この状態空間モデルに粒子フィルタと自己組織化 (σ^2 の逐次推定) という手法 [3] を用いることで，一期先予測分布 $p(x_{t+1} \mid y_1, \cdots, y_t)$ を逐次的に計算しました．そして，次に出す手の予測分布

$$p(y_{t+1} \mid y_1, \cdots, y_t) = \sum_{x_{t+1}} p(y_{t+1} \mid x_{t+1}) p(x_{t+1} \mid y_1, \cdots, y_t)$$

を計算し,「最も出す確率が高いと予測される手に勝つ手を出す」という戦略でヒカキンさん・サザエさんと対戦してみました．

ヒカキンさんのデータに対する予測確率 $p(y_{t+1} \mid y_1, \cdots, y_t)$ は図 2 のようになりました．全体的にグーの確率が高めである一方で，最近はチョキを出しやすくパーを出しにくいという予測になっています．この予測確率に基づいて対戦した結果，勝ち・あいこ・負けの回数はそれぞれ 629，499，531 でした．なお，毎回 1/3 の確率でランダムに手を出す戦略で 629 回以上勝つ確率は $p = 3.8 \times 10^{-5}$ と小さいです (二項分布の検定)．よって，状態空間モデルでヒカキンさんの手の出し方のくせを読み取ることによって，高い勝率を達成できたと考えられます．

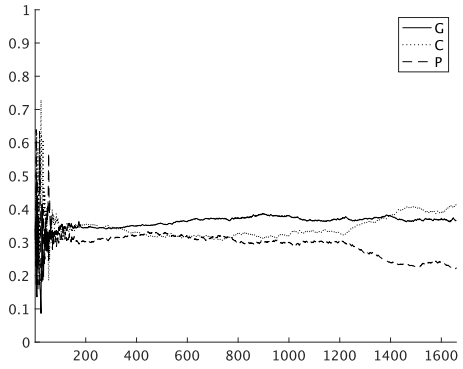

図 2　ヒカキンさんの手の予測確率

一方，サザエさんのデータに対する予測確率 $p(y_{t+1} \mid y_1, \cdots, y_t)$ は図 3 のようになりました．3 つの手の予測確率は均等に近くなっており，サザエさんが強敵であることが窺えます．この予測確率に基づいて

対戦した結果，勝ち・あいこ・負けの回数はそれぞれ 391，572，555 でした．今度は負け越しになっています！　なお，ランダムな戦略で勝ちが 391 回以下になる確率は $p = 1.9 \times 10^{-10}$ ととても小さいです．つまり，今回用いたモデルはサザエさんに対しては特に相性が悪いようです．実は，サザエさんの出す手には「同じ手を続けて出しにくい」などの傾向があることが指摘されています [5]．現状の状態空間モデルではこうした依存構造は捉えられないため，適切にモデルを変更することでサザエさんにも勝てるようになるかもしれません．

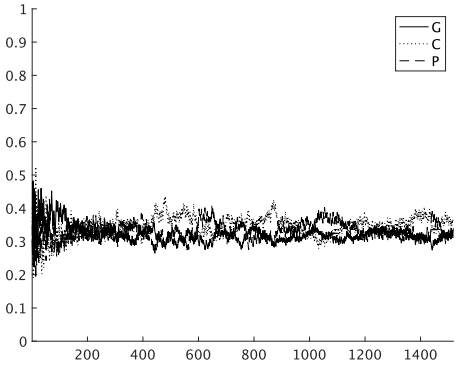

図 3　サザエさんの手の予測確率

参考文献

［1］ 赤池弘次，北川源四郎編 (2020).『時系列解析の実際 (新装版)』，朝倉書店.

［2］ 樋口知之，上野玄太，中野慎也，中村和幸，吉田亮 (2011).『データ同化入門——次世代のシミュレーション技術』，朝倉書店.

［3］ 北川源四郎 (2020).『R による時系列モデリング入門』，岩波書店.

［4］ ヒカキンぶんぶんじゃんけん記録室
`https://hikakinjunken.tk`

［5］ サザエさんじゃんけん研究所
`http://park11.wakwak.com/~hkn/`

[まつだ たける]