

統計科学のフロンティア 4

# 階層ベイズモデルと その周辺

統計科学のフロンティア 4

甘利俊一 竹内啓 竹村彰通 伊庭幸人 編

# 階層ベイズモデルと その周辺

時系列・画像・認知への応用

石黒真木夫 松本隆  
乾敏郎 田邊國士

岩波書店

## 編集にあたって

# 階層ベイズ法——ベイズ統計の新しい展開

この巻の目的は、柔軟なモデリングのための道具として、さまざまな分野で注目を浴びている「階層ベイズ法・経験ベイズ法」と「罰金付き推定」の世界を解説することである。応用としては、時系列の解析と予測、確率密度の推定、画像再構成、視覚の認知科学、非適切逆問題と数値解析など、幅広い話題がとりあげられている。

階層ベイズ法とはどういう考え方かを簡単に説明する。従来の統計的モデリングでは、データ  $\mathbf{y}$  について、確率モデル  $P(\mathbf{y}|\mathbf{x})$  を考えるとき、パラメータ  $\mathbf{x} = \{x_i\}$  の独立な要素となるべく少なくするのが基本であった。これは頻度主義でもベイズ統計でも同じである。パラメータ数が多いほうが、空間的・時間的な変動や対象の個性を表現するには有利であるが、単にパラメータ数を増やしたのでは、有限個のデータの背後にある法則を表現する能力がかえって下がり、未知のことが予測できなくなってしまう。

これに対して、階層ベイズ法では以下のように考える。まず、パラメータは十分たくさん用意する。たとえば、 $\mathbf{x}$  として、各時刻でのシステムの状態、真の画像の各画素の値、ニューラルネットの各結合の値、各個体の特性をあらわす値、などをそのまま使う。その一方で、 $\mathbf{x}$  についての事前の知識を事前分布  $P(\mathbf{x}|\alpha)$  で表現し、ベイズの公式(ベイズの定理)によって求めた事後分布

$$P(\mathbf{x}|\mathbf{y}) \propto P(\mathbf{y}|\mathbf{x})P(\mathbf{x}|\alpha)$$

によって推論を行う。 $\alpha$  は、事前知識の内容やその信頼度をあらわすパラメータで、ハイパープラメータ(超パラメータ)と呼ばれる<sup>\*1</sup>。情報が不十分で多義的な解釈が可能なときに  $\mathbf{x}$  をどうするかを、事前分布  $P(\mathbf{x}|\alpha)$  で

---

\*1 事前知識を(ソフトな)拘束条件とみなせば、拘束条件に含まれる定数や拘束の強さが  $\alpha$  で表現されるわけである。実際には、 $P(\mathbf{y}|\mathbf{x})$  に含まれているパラメータのうち、 $\mathbf{x}$  全体に関係するもの(たとえば画像や時系列に加わった雑音の強さ)についても  $\alpha$  と並行してハイパープラメータとして扱うが、以下では省略する。

表現することで、多数のパラメータについての安定した推論を可能にするわけである。これは、多数用意したパラメータにソフトな制約をつけることで柔軟さを保ちつつ、個数を減らすのと同等の効果を狙っているともみられる。

ここまでが第1段階である。第2段階は、ハイパーパラメータ  $\alpha$  の決定である。階層ベイズ法では、 $\alpha$  についても事前分布  $P(\alpha)$  を仮定して、 $(\mathbf{x}, \alpha)$  という拡張された空間での事後分布

$$P(\mathbf{x}, \alpha | \mathbf{y}) \propto P(\mathbf{y} | \mathbf{x})P(\mathbf{x} | \alpha)P(\alpha)$$

を考えることで、ハイパーパラメータの自由度を扱う。モデルが階層的な形  $P(\mathbf{y} | \mathbf{x})P(\mathbf{x} | \alpha)P(\alpha)$  に表現されるのが「階層ベイズモデル」と呼ばれる理由である。実際の応用では、 $\alpha$  の周辺分布(周辺密度)

$$P(\alpha | \mathbf{y}) \propto \int P(\mathbf{y} | \mathbf{x})P(\mathbf{x} | \alpha)P(\alpha) d\mathbf{x}$$

を最大にする  $\alpha$  を求め、これを用いて  $\mathbf{x}$  についての推論を行ってもほぼ同じ結果になることが多い。 $P(\alpha)$  に一様分布を仮定すると、これは、周辺尤度

$$\int P(\mathbf{y} | \mathbf{x})P(\mathbf{x} | \alpha) d\mathbf{x}$$

を最大化する  $\alpha$  を選ぶのに等しい。このような、 $\alpha$  を点推定で置き換える方式を、特に「経験ベイズ法」ということがある。ただし、この名称は、別の種類の「経験的」方法、たとえば、別に用意したデータベースから事前の頻度を推定して事前分布を定める方法にも用いられることがあるので、注意が必要である。

以上のうち、第1段階の部分は、後述するように、ベイズ統計的な解釈に限定せずに、罰金付き推定という視点から眺めることもできる。また、第2段階の「 $\mathbf{x}$  について和をとった量で  $\alpha$  を推定し、それを用いて  $\mathbf{x}$  についての推論を行う」という考え方、陽に「階層ベイズモデル」と呼ばれているものだけではなく、時系列の状態空間モデル、隠れマルコフモデル、有限混合分布モデル、欠測を含むモデル、潜在変数モデル、等々に基づく手

法にも共通である<sup>\*2</sup>.

こうした目でみると「階層ベイズ法」「経験ベイズ法」に関連して論じられる範囲は広範なものになるが、本巻は、そのうち、核となる分野、この手法の精神が最も明確にあらわれるような話題を選んで構成されている。ここに含まれない話題のいくつか、たとえば、離散状態の隠れマルコフモデルや有限混合分布モデルは「統計科学のフロンティア」シリーズのほかの巻で扱われている。また、シリーズ12巻『計算統計Ⅱ』では、マルコフ連鎖モンテカルロ法との関連でベイジアン・モデリングが論じられる予定である。必要に応じて、これらとあわせて読まれるとよいと思う。

以下、各部の内容を簡単に紹介する。まず、石黒による「事前情報を利用した複雑な系の解析」であるが、離散データの解析、密度推定、季節調整などを題材に、赤池によって提唱され、統計数理研究所のグループによって展開されたベイズ型情報処理の考え方方が実践的に述べられている。石黒の研究の出発点は時系列解析にあるが、第Ⅰ部では、系列事象の解析と汎用的な側面の両方に目配りした解説がなされている<sup>\*3</sup>。また、カルマンフィルタやガウス近似を中心に、数値的方法の基礎も扱われている。

このグループの研究は、経験ベイズ法そのものに関する先駆的なものであるが、季節調整のような一意性のない分解問題への応用は特に独自性が高く、この分野の知識が既にある読者にも興味深いであろう。

次が、松本による「非線形ダイナミカルシステムの再構成と予測」である。ここで扱われるのは時系列予測の問題であるが、石黒のアプローチとは異なり、変数の時間変化を記述する非線形の方程式を推定する手法がとられている。松本の方法では、一般的の非線形関係を記述するためにニューラルネット(多層パーセプトロン)を用いるが、その際、推定すべき結合定

\*2 分野によっては、ここでいう  $x$  に相当するものを「状態(state)」「画素(pixel)」「欠測値(missing data)」「潜在変数(latent variable)」等と呼び、 $\alpha$  を「パラメータ」、周辺尤度を単に「尤度」と呼ぶので、対応関係を考える際には注意を要する。

\*3 時系列の状態空間モデルという立場で統一したテキストとしては、北川源四郎『時系列解析入門』(岩波書店、2005)がある。ただし、同書の「状態」「パラメータ」「AIC」は、石黒の解説の「パラメータ」「超パラメータ」「ABIC」に、それぞれ対応するので注意が必要である。

数の数がデータ数に比して多すぎるという問題が生じる。これを階層ベイズ的な枠組で処理するというのが、方法の骨子である。同一入力を持つ結合をグループ化して共通の事前分布を設定し、重要度の低い結合の強さが弱くなるように制約することで、予測能力の向上を狙っている。線形モデルの場合のリッジ回帰の一般化ともいえる。最後の章では、ハミルトニアン・モンテカルロ(ハイブリッド・モンテカルロ)の応用にも触れている。

松本の定式化は、マッカイによるニューラルネットの階層ベイズ的取り扱いを時系列に適用したもので、赤池グループのそれとは起源が異なっている。細部の考え方には違いがあるが、はじめての読者はむしろ一致する点のほうを強く感じられるかもしれない。

本論の3番目が、乾による「視覚計算とマルコフ確率場」である。ここでは、一見まったく違う分野である視覚の認知科学が扱われる。視覚計算の基礎からはじまり、標準正則化による定式化が解説され、後半では、マルコフ確率場を事前分布とする画像再構成から、乾と川人による双方向性結合に基づく脳理論に及ぶ、まとめた解説が与えられている。

この定式化では、視覚情報処理は網膜への投影像などの限定された情報から3次元構造を復元する不良設定問題として扱われる。われわれは、運動する物体や映写された映画が連続的に動いて見えるのも、物体の輪郭がはっきり識別できるのも、当たり前であると考えがちである。しかし、視覚情報処理の研究は、それらの背後に、脳の中に組み込まれている事前知識あるいは外界のモデルと、外界から感覚器により得たデータとの複雑な相互作用があることを示唆している。この立場では、「錯覚」——たとえば「床屋の棒」の赤青の縞が上下に動いて見えること——は「推論」につきもの必然的な結果として説明されるのである。こうした見方を知ることは、統計的モデリングを学ぶ上で大変重要だと考える。

掉尾を飾るのが、田邊による補論「帰納推論と経験ベイズ法——逆問題の処理をめぐって」である。非適切逆問題(不良設定逆問題)という視点から説き起こして、この分野の全体を貫く思想をコンパクトに述べた内容である。概念的・思想的な内容を先ず把握したい読者は、この序文に続いて田邊の解説を読まれることをお勧めする。

「階層的モデリング」の最も素朴な形は、たとえば「顧客－顧客のグループ」「個人－学級－学校」のように、外界に明示的に存在する階層性をベイズの枠組みを利用してモデル化するものかもしれない。本巻では、こうしたタイプのモデルはあまり取り上げなかった。この点に違和感を持つ読者もいるかもしれないが、「階層」の概念はより広い意味を持つことを強調したい。



本巻のあらましを知るには、ここまで解説で十分であるが、興味のある読者のために、この分野における「ベイズ統計」の意味について、また、本シリーズの3巻『モデル選択』との関連について、もう少し論じることにする。

まず、ベイズ統計の立場から見ると、周辺尤度最大化によりハイパーパラメータ  $\alpha$  を定めることは「あるデータからそのデータを解析するための事前分布を適応的に定める」ともみられることを注意しておく。改めてこのように表現すると、ベイズ統計の基本に反するものとして眉をひそめる向きもあるかもしれない。この点は解釈が微妙に分かれるところで、この序文のはじめの解説や本文の松本の解説では「周辺尤度の最大化＝拡張された空間  $(\mathbf{x}, \alpha)$  でのベイズ法の近似」と割り切っているのに対し、石黒や田邊の解説では、ハイパーパラメータの最尤推定とパラメータについてのベイズ法を組み合わせたものとして説明している<sup>\*4</sup>。

こうした理念的なことも興味深いが、実際に応用するには、どのような状況が「データからそれ自身を解析するための知識や拘束条件を定める」ことを許しているのかを、個々の場合に反省することが、むしろ重要かもしれない。まず、 $\mathbf{x}$  の要素数に比べて、ハイパーパラメータ  $\alpha$  の数が相対

---

\*4 なお、これに関連して、さらに上の階層のモデル選択(尤度や事前分布の分布形の選択、アーキテクチャの選択)についても立場の相違がある。赤池や石黒が周辺尤度に AIC 的な補正を行うことを想定しているのに対し、マッカイや松本は各階層について純粹のベイズの枠組を用いる立場であり、田邊はデジタルなモデル選択自体を避けて可能な限りひとつのモデルに埋め込むべきだとしている。

的に少ないことが基本であろう<sup>\*5</sup>. また,  $\mathbf{x}$  の一部についての知識が残りの要素に関してもなんらかの情報を与えることが必要である<sup>\*6</sup>. たとえば, 「滑らかな曲線でつなぐ」場合を考えると, 通常の階層ベイズ・経験ベイズの扱いでは, 曲線全体で「滑らかさの度合い」がある程度一様であることが前提とされている. ニューラルネットの結合定数の推定の場合は, 結合の強さが同じオーダーにあることが期待できるような組み分けを行ってから, 事前分布を設定している. こうした前提のもとでは, 多数の  $\mathbf{x}$  に共通する性質をデータから学習して, 個々の  $\mathbf{x}$  の推定にフィードバックするのは自然である. それを数理的に表現したのが, 階層ベイズ・経験ベイズの手法であると考えられる.

次に, ベイズの立場を離れて, より広い立場から問題を眺めてみよう. 石黒や田邊が論じているように「第1段階」の  $\mathbf{x}$  の推定の部分だけなら, 必ずしもベイズ的な枠組みを想定しなくとも,  $-\log P(\mathbf{x}|\alpha)$  の部分を「罰金」とみなすことで, 「罰金付き推定」としての定式化が可能である. 逆にいえば, ベイズ的な構造を仮定せずに, ハイパーパラメータ  $\alpha$  を決める手法があれば, 全体を非ベイズ的(頻度主義的)な定式化として完結させることも可能ということになる.

このような方法として最も簡単なものは「データ  $\mathbf{y}$  の一部を“テスト用データ”として分けておき, 残りを“学習用データ”として, 後者のみで  $\mathbf{x}$  の推論を行い, 前者を最もよく予測する  $\alpha$  を選ぶ」ことであり, 交差確認法(交差検証法, cross-validation)と呼ばれる. それ以外の手法としては, 石黒の解説にある EIC, 3巻『モデル選択』で触れられる予定の GIC などがある<sup>\*7</sup>. また, ある種のモデルと損失関数, 仮定のもとで「必ず結果が良

\*5 階層数が3つ以上ある場合などでは, 各階層にいろいろな役割分担がありうるので, 一概には言えないかもしれないが.

\*6 スタイン推定を学ばれた読者はこれには異論があるかもしれない. しかし, スタイン推定の示しているのは「 $\mathbf{x}$  の要素が互いに無関係でも損失がないようにできる」ということで, 積極的に良い結果になるのは, やはり要素間になんらかの関連がある場合ではないだろうか. また, スタイン推定の理論でカバーされる範囲は, 経験ベイズ法一般よりはるかに狭いことも注意.

\*7 本書の直前に刊行された下記の書物にも GIC, EIC や正則化法との関連を含む解説がある. 小西貞則, 北川源四郎『情報量規準』(シリーズ予測と発見の科学2, 朝倉書店, 2004).

くなる(悪くならない)ような罰金の付け方」が存在するケースがあり、スタイル推定量として知られているが、これについても 3 卷で論じられる。

ここで、なぜ、3 卷『モデル選択』が出てくるのかと疑問に思う読者もいると思う。これは、「モデル選択」というと、離散的なモデルの中からデジタルに 1 つを選ぶことであると一般に考えられているからであるが、本巻で論じたような事前分布や罰金によってパラメータ  $x$  にソフトな制約をつける手法<sup>\*8</sup>も広い意味のモデル選択である。たとえば、ニューラルネットへの応用では、結合強度に罰金を与えることで予測能力を向上させようとしているが、これは、重要でない結合を切る(ゼロとおく)のと同じ目的の操作である<sup>\*9</sup>。「モデル選択」をこのように一般的に捉えれば、3 卷と 4 卷の内容に関連があるのは当然である。両者の分担は、大まかに言えば、4 卷がベイズ的・実践的であるのに対し、3 卷は非ベイズ的・理論的ということになるが、これはあくまでおよその傾向であって、たとえば、3 卷で論じられる MDL は広い意味でのベイズ的手法に属するし、4 卷の EIC はどちらかといえば非ベイズ的な概念である。また、3 卷にも豊富な例が含まれる予定である。

統計科学は異質な理念や価値観が絡み合う世界である。本シリーズでは、テキストとしての効率を最適化するより、むしろ差異をそのまま提示して、読者自身に考えてもらうことを重視した。読者が多様性と共通性を楽しめつつ、豊富な内容を有効に活用されることを願っている。

(伊庭幸人)

---

\*8 枠組みによっては、これとは別に、分布族の形やアーキテクチャについてのデジタルなモデル選択があるが、これらはより上の階層にあるとみなされている。

\*9 これに関連して、カーネル法における正則化について、6 卷『パターン認識と学習の統計学』の II 部に簡潔な解説がある。本巻では触れなかった  $\ell_1$  ノルムによるスパース化についても論じられている。



## 目 次

編集にあたって

## 第Ⅰ部 事前情報を利用した複雑な系の解析

石黒真木夫 1

## 第Ⅱ部 非線形ダイナミカルシステムの再構成と予測

松本隆 89

## 第Ⅲ部 視覚計算とマルコフ確率場 乾敏郎 171

補 論 帰納推論と経験ベイズ法  
—逆問題の処理をめぐって— 田邊國士 235

索 引 253

裝丁 蟻名優子

# I

---

## 事前情報を利用した複雑な系の解析

石黒真木夫

## 目 次

1 はじめに	3
1.1 知りたいことの量とデータの量	3
1.2 問題の出どころ	4
1.3 簡単な例題	5
1.4 ベイズ型情報処理	7
1.5 この稿の構成	13
1.6 記号	13
2 ベイズ型情報処理の適用例	16
2.1 ベイズ型2値回帰	16
2.2 密度関数推定	18
2.3 季節調整法	19
3 ベイズ型情報処理の技術要素	23
3.1 <i>AIC</i>	23
3.2 ガウス分布の場合のベイズ公式	26
3.3 時系列データの場合	38
3.4 粒子ベイズ	46
3.5 2次近似	50
3.6 「滑らかな変化」を扱う技術	51
4 ベイズを越えて	53
4.1 MAP推定	53
4.2 情報量規準 <i>EIC</i>	57
4.3 数値例	60
4.4 仮想的観測	70
5 おちばひろい	73
5.1 縦と横	73
5.2 局所的モデリング	75
5.3 絵解きベイズ定理	76
6 最後に	81
6.1 「滑らかさ」以外の「事前情報」	81
6.2 能動的解析/実験計画との接点	82
付録	83
A.1 <i>AIC</i> 最小化法の論理	83
A.2 Householder法	83
参考文献	86

# 1 はじめに

情報処理にもいろいろある。本稿は統計科学的情報処理について論ずるものである。統計科学的情報処理をさらにデータの量と知りたいことの量で分類する。われわれが目標とするのは、数万個のデータに基づいて数万個のパラメータを推定することを要する問題の扱いである。

この種類の問題を取り扱う典型的な方法としてベイズ的な方法がある。この章でベイズ的処理を紹介し、そのベイズ的処理の構成要素に即して全体の組立てを説明する。本章を含めて本稿のいたるところで利用するベイズ公式の説明もこの章に置いた。全体を通じて利用する数式の記法に関する規約もここに置いた。

## 1.1 知りたいことの量とデータの量

「データの量」はデータがコンピュータのメモリで専有するアドレスの数で数えることとしよう。知りたいことの量は、何らかの数式の形と、そのパラメータの値まで含めてわかれれば「知りたいことがわかった」ということにして、そのパラメータの個数をもって知りたいことの量とする。たとえばある集団から 100 人の身長を計ったデータから、その集団の平均身長を推定する問題は 100 個のデータから 1 個のパラメータを推定する問題ということになる。

「身長推定問題」がどういう数式と関係しているのかと考える読者がいるかもしれない。あるデータの平均値を求めるということをそのデータの分布にガウス分布モデル

$$\frac{1}{\sqrt{2\pi}\sigma^2} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}$$

を仮定して、そのパラメータ  $\mu$  を推定することと解釈することができる。

そう考えれば、この式がこの場合の「何らかの式」ということになる。あなたは平均を求める時にデータの分布がガウス分布であるかどうか気になどしていないかもしれない。そんなあなたに問題を1つ。データの分布がガウス分布と似ても似つかぬものだったら、その分布の平均値にはどういう使い道があるだろう？

われわれの目標にこの身長平均推定問題は入れない。数十のデータに基づいて数個のパラメータを推定する問題も枠の外に置く。数個のデータに基づいて数百のパラメータを推定する問題は扱わなくてはならないだろう。

当面100個程度のデータに基づいて100個程度のパラメータを推定する問題を境界として、その「あたり」から上を考えていくことにしよう。

## 1.2 問題の出どころ

現実の世界で起こっていることを正確に把握し、合理的に対処するためには、わかりやすいモデルに基づいた検討が重要である。現実的な問題であればあるほど、モデルが多くのパラメータを含み、そのパラメータの値を定めるために必要となるデータの量も多くなる。「わかりやすい＝パラメータ数が少ない」ではない。わかりやすくするためにパラメータ数を増やすことが必要な場合もある。問題意識が明確な場合には知りたいこと自体のモデルはそれほど多くないパラメータで記述されることも多いが、現実の「きたない」データはそれほどストレートに知りたいことを明かしてくれない。データが完璧な観測計画に基づいて、よく整備された装置によって整然と集められたものだったらいいのだが、そうでないことが多い。観測装置が予期した通りに整備されていないなど予想外の出来事で計画通りの観測ができないことがあるし、もともと別の目的のために作られた観測計画に基づいて集めたデータを他の目的に流用したりもする。このような場合、観測系が抱える「きたない」面を抱えこんだパラメータ数の多いモデルが必要になってくる。

例えば、地上における大気の底からの天体観測においては、大気も観測系の一部として考慮しなければならず、この部分まで完璧に整備した装置

による観測は不可能である。人工衛星の利用などによる大気圏外に出ての観測がこの問題を解決する。このような不定要素を観測技術の改良によってできるだけ除くのが王道である。しかし、コストその他の点で王道をとれない場合もあるし、ある時点での技術の粋を尽くして王道をつきつめてもパラメータの数をおさえられないことがある。このような場面がわれわれが扱おうとする問題の出どころである。

### 1.3 簡単な例題

方眼紙にフリーハンドで曲線を引いた(図1)。横軸を  $x$ 、縦軸を  $y$  として、 $x = 0.5 \text{ cm}, 1.0 \text{ cm}, 1.5 \text{ cm}, \dots$  の位置での曲線の  $y$  値に乱数を加えて点を打ち(図2上)，それから元の曲線を消した(図2下)。こうして図3のようなデータが得られた。

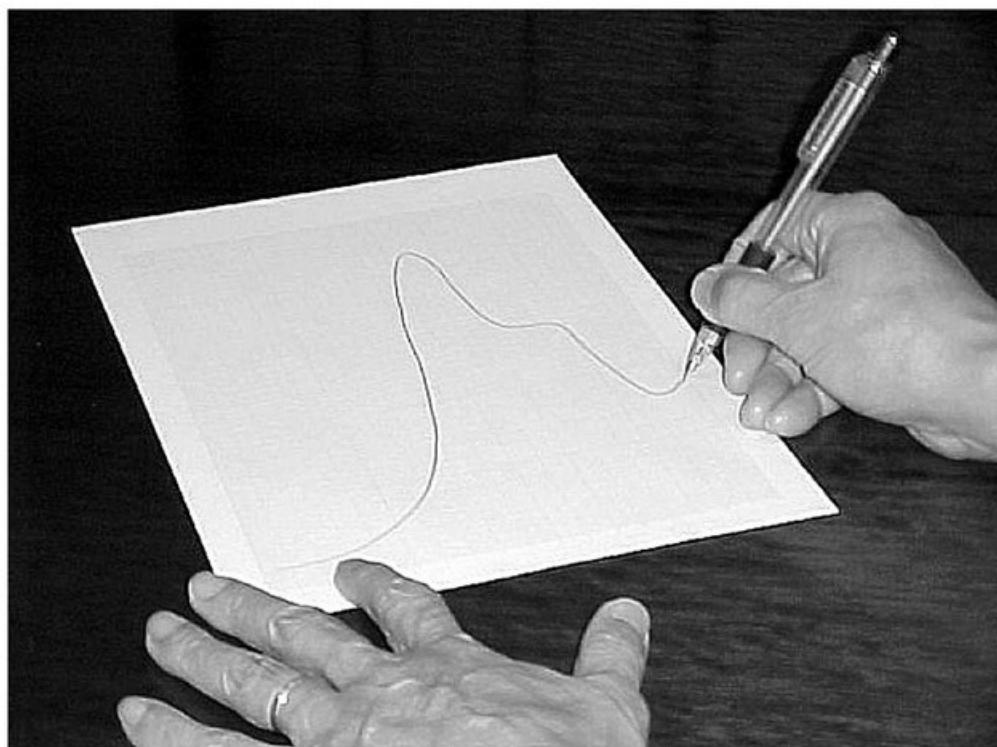


図 1 フリーハンドで曲線を引く

このデータから元の曲線を推定するのを「簡単な例題」とする。以後「手書き曲線の推定」問題、と呼ぶ。

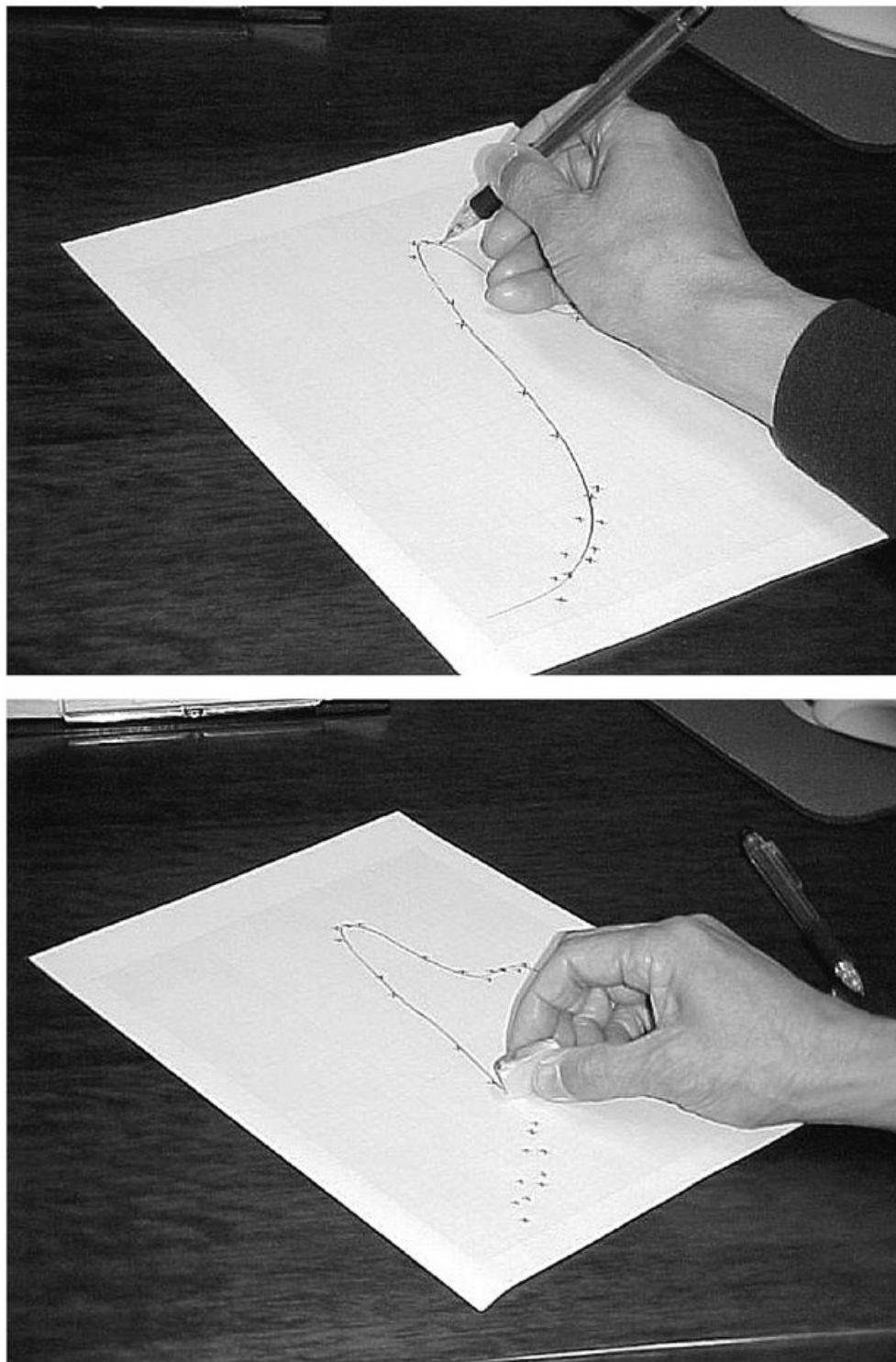


図 2 データを作る

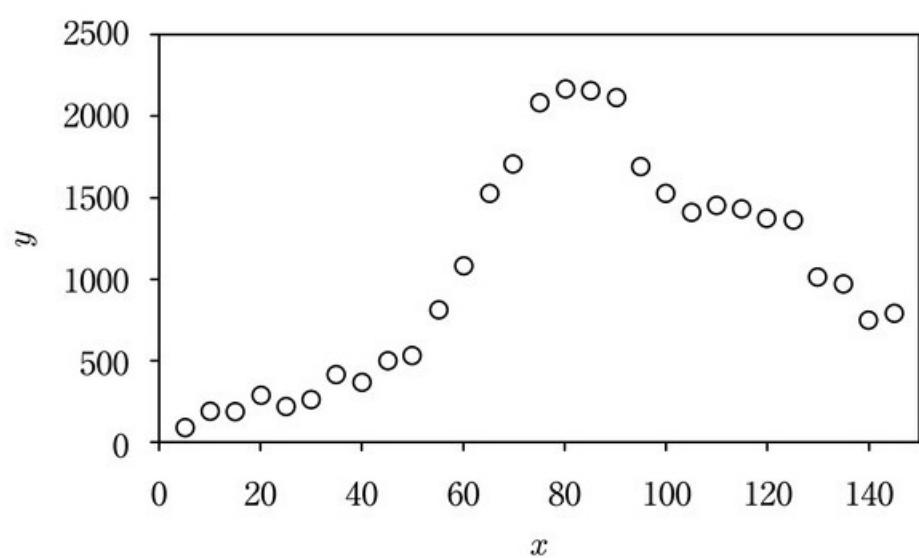


図 3 データ

## 1.4 ベイズ型情報処理

知りたいことを数式の形で表現しよう。元の曲線を  $y = f_*(x)$  と書くことにすれば、問題はデータから  $f_*$  の形を推定する問題となる。 $x = 0.1 \text{ cm}, 0.2 \text{ cm}, 0.3 \text{ cm}, \dots$  の位置における  $f_*(x)$  の値がわかれればいいことにしよう。 $f_i = f_*(0.1i)$  と書くことになると、 $\{f_1, f_2, \dots, f_{5n}\}$  が「知りたいこと」となる。曲線の形についてどの程度の「解像度」で知りたいのかを「暗に」表現してしまっていることに注意。観測値の方が  $0.5 \text{ cm}$  おきであるから

$$y_i \equiv f_{5i} + r_i \quad (1)$$

である。 $n$  個のデータ  $\{y_1, y_2, \dots, y_n\}$  から  $5n$  個の値  $\{f_1, f_2, \dots, f_{5n}\}$  を推定する問題になる。

$$r_i \sim N(0, \sigma^2) \quad (2)$$

と仮定すると。(1)式と(2)式のセットで曲線の形とデータを結びつける統計的モデルが構成され、 $\{f_1, f_2, \dots, f_{5n}\}$  に加えて  $\sigma^2$  も推定することにすれば統計学的な問題となる。

$\{r_i\}$  を互いに独立な確率変数とすると、ガウス分布の確率密度関数を  $\phi$  で表して、データ全体の確率密度関数が

$$P(\mathbf{y} | \mathbf{f}; \sigma^2) \equiv \prod_{i=1}^n \phi(y_i | f_{5i}; \sigma^2) \quad (3)$$

と書ける。 $\mathbf{y} = (y_1, \dots, y_n)^T$ ,  $\mathbf{f} = (f_1, \dots, f_{5n})^T$  である。この式は  $\mathbf{y}$ ,  $\mathbf{f}$  および  $\sigma^2$  を変数とする関数であるが、 $\mathbf{y}$  に観測値を入れて固定すれば  $\mathbf{f}$  と  $\sigma^2$ だけを変数とする関数、尤度関数となる。尤度関数が書ければ、まず最尤法を使うことを考えるのが普通だが、このモデルのパラメータの数は  $5n + 1$  であり。データの数が  $n$  であるから最尤法は働かない。無理やり使えば  $\hat{f}_{5i} = y_i$  ( $i = 1, 2, \dots, n$ ) となるであろうが、 $f_6, f_7, \dots, f_9$  などの中間の値は定まらず期待した答にはならない。たとえば、多項式回帰モデル

$$f_i \equiv \sum_{m=0}^M a_m i^m \quad (4)$$

は  $M$  を十分小さくとれば、最尤法で扱えるモデルになる。にもかかわらず

このような“reparametrization”(パラメータを別のパラメータで書き直すこと。ここでは  $f$  というパラメータを  $a$  で書き直す例である)はここでは採用しない。フリーハンドで描いた曲線が、多項式で表せるとは思えないし、 $a_0, a_1, a_2, \dots$  という係数に実用的な意味があるとも思えないからである。“reparametrization”ではなく、ベイズの公式を利用してみよう。

**ベイズ公式** 簡単な式である。2つの確率変数の同時分布、周辺分布、条件付分布の間の関係である。パラメータ  $\theta$  に依存する確率変数  $X$  の分布  $P(x|\theta)$  が与えられているものとする。 $\theta$  の分布  $P(\theta)$  も与えれば、 $X$  と  $\theta$  の同時分布  $P(x,\theta)$  は

$$P(x,\theta) \equiv P(x|\theta)P(\theta)$$

である。同時分布を細工して次の一群の式を得る。

$$\begin{aligned} P(x) &\equiv \int P(x,\theta)d\theta \\ P(\theta|x) &\equiv \frac{P(x,\theta)}{P(x)} \end{aligned}$$

$$P(\theta|x)P(x) = P(x,\theta)$$

$$P(x|\theta)P(\theta) = P(\theta|x)P(x)$$

最初の2つは定義式である。数式とはさみは使いようで切れる。われわれはこれらの式において  $x$  にデータの役割を与える。 $\theta$  はデータ分布のパラメータということになり、 $P(\theta)$  は  $\theta$  の出方についてデータを見るまでもなくわれわれが知っていることを  $\theta$  の分布の形で表現するものとなる。「 $\theta$  の事前分布」と名付ける。このような役割の非対称性をきわだたせるために  $\theta$  の分布を  $P(\theta)$  でなく  $\pi(\theta)$  と書くことになると最後の式は

$$P(x|\theta)\pi(\theta) = \pi(\theta|x)P(x)$$

となる。

今後この公式をあちこちで使う。この公式を使ったことがわかるようにこの公式による式の変形を

$$\underbrace{P(x|\theta)\pi(\theta)}_{\text{B}} = \underbrace{\pi(\theta|x)P(x)}_{\text{B}}$$