経営工学におけるモデル選択

関　庸一

1. モデル選択への問題の定式化

経営工学の対象とする問題には「データでものをいう」必要がある場合が多い。たとえば、品質管理、経営戦略決定の問題などでは、工程の現状や、経営管理データなど、"何らかの誤差"を含むデータから意思決定を迫られる場合が多い。このような場合、観測データからその背後の統計的モデルを予測して、それに基づき意思決定することになる。この場合の統計的モデルとは、データ発生源の確率分布である。

たとえば、連続的に製品が加工されている工程を考えてみよう。その製品には重要な特性があり、この特性値に異常が発見されれば、工程を一旦止めて工程を正常な状態に調整し直す必要があり、加工と同時にその特性値を観測しているとする。

このような状況において、もし、特性値が測定誤差や工程の偶然変動を含まずに観測できる場合には意思決定は簡単となる。標準値と比較して問題となるズレが生じていたら工程を止めて調整をすれば良い。ところが、無視できない誤差が付随した観測値しか得られない場合には、問題は難しくなる。観測された特性値は常に、大なり小なり標準値と比較してズレが生じているわけで、いま意思決定のために必要な、工程特性の真の値は直接には得られないこととなる。

たとえば、特性値$y_i\ (i=1,\cdots,n)$が、既知の分散σ^2の正規誤差を伴って観測されると考えてよい状況であるとしよう。このとき考えられる工程の状態は、「工程特性の真の値が所与の標準的な値μである」(管理状態)と「工程特性の真の値は所与の標準的な値μからずれた未知のμ'である」の二つであって、対応して二つの確率モデル$y_i \sim N(\mu, \sigma^2)$と$y_i \sim N(\mu', \sigma^2)$のいずれが正しいか、観測される特性値から選択し、もし、ズレが存在するなら、μ'の推定値$\hat{\mu}'$から意思決定を行なうこととなる。

せき　よういち　群馬大学情報工学科
〒376 桐生市天神町1-5-1

このように、問題解決上関心対象とする量が直接観測できず、確率変数の実現値としてのみ得られる場合を考えると、対象問題に関して想定される確率モデルをいくつか用意し、データから、どのモデルが正しいか？モデルのパラメータがどんな値か？を推定して必要な行動を選択判断するという形で問題解決へのアプローチを定式化できる。このようなモデル選択の考え方で定式化できる問題は多く、モデル選択に対して有効な方法論を与えることは、大変重要と考えられる。このモデル選択の方法論として近年注目を集めている基準として、この特集対象のMDL基準がある。

問題をモデル選択として定式化するときには、

1. どんなモデルのクラスを想定するか(どのようなパラメータを持たせるか)という現象のモデル化
2. データから、どのモデルを選ぶべきかというモデル選択の基準
3. 推定されたモデルから、現実のコストなどの評価基準に基づき行動を決定する方法

のそれぞれに、選択肢が考えられる。これらは本来不可分で、一体にして考えなければ、問題解決として一貫した手続きとならない。特にモデルと行動を対にしてモデル選択を考える必要がある。

この稿では2.,3.の関係を危険率でとらえられる場合を中心として単純な正規分布モデルを例として解説することとする。

2. MDL原理

MDL原理[6, 7, 5]は想定したモデルのクラスの中からモデルを選ぶ基準として、「与えられたデータを最も短く記述できるモデルが良いモデル」と判断するものである。ここで、データの記述の長さとは、データを(例えば2進)語頭符号化([4]参照)したときの符号長のことで、モデルクラスに関する知識だけを共有する伝達相手が復号化可能な形式で符号化するときに必要となる仮想的な符号長だと考えれば良い。

摸式的に表現すると、[モデルの記述長] + [モデルを知った下でのデータの記述長] と表現できる。ここで、[モデルの記述長] はモデルクラスのどのモデルであるかを指定するための記述長である。個々のモデルがパラメータ化されている場合には、[パラメータタイプ m の指定] $:l_m$ と [パラメータ値$\boldsymbol{\theta}$の指定] $:l(\boldsymbol{\theta}|m)$ の二つに分けて記述できることとなる。ただし、$\boldsymbol{\theta}$は、推定を必要とするパラメータのベクトルで、m ごとに定まる次元数 p を持つものとする。また、各要素間には拘束条件がなく、自由度 p であるようにパラメータ化されているとする。[モデルを知った下でのデータ$\boldsymbol{y}$の記述長] は、モデルの確率密度関数 $P(\boldsymbol{y}|\boldsymbol{\theta},m)$ と観測データ$\boldsymbol{y}$から計算される対数尤度の符号を替えたもの$-\log P(\boldsymbol{y}|\boldsymbol{\theta},m)$と考えて良い。以上から、パラメータタイプ m のモデルを用いて符号化を行なった際の符号長 $L(\boldsymbol{y},\boldsymbol{\theta},m)$ は次式となる。

$$L(\boldsymbol{y},\boldsymbol{\theta},m) = -\log P(\boldsymbol{y}|\boldsymbol{\theta},\ m) + l(\boldsymbol{\theta}|m) + l_m \quad (1)$$

MDL 原理では (1) 式が最小となるようなモデルを選ぶこととなる。ここで、第 1 項を最小化する$\boldsymbol{\theta}$は最尤推定量$\hat{\boldsymbol{\theta}}$ であり[1]、$\boldsymbol{\theta}$の値を$\hat{\boldsymbol{\theta}}$に高い精度で近づけるほど小さくなるが、第 2 項は$\boldsymbol{\theta}$の記述精度を上げると大きくなるので、このトレードオフ関係を考慮し、まるめ精度$\boldsymbol{\delta}$について最小化する。パラメータが独立に推定される場合には、

$$L(\boldsymbol{y},m) = -\log P(\boldsymbol{y}|\hat{\boldsymbol{\theta}}+\boldsymbol{\delta},\ m) + \sum_{i=1}^{p} L^*\left(\frac{\hat{\theta}_i}{\delta_i}\right) + l_m \quad (2)$$

を$\boldsymbol{\delta}$について最小化することとなる。ここで、$L^*(x)$ は実数 x を整数化した上で符号化した際の、x の符号長を与える適当な関数である。$L^*(x)$ は $O(\log(x))$ のオーダーにすることが可能なので、結局、普通は、漸近近似した上で、モデル選択に無関係な定数項を省いて、次のような基準となる。

$$L(\boldsymbol{y},m) = -\log P(\boldsymbol{y}|\hat{\boldsymbol{\theta}},\ m) + \frac{p}{2}\log(dn) + l_m \quad (3)$$

ここで、d は確率分布から定まるある定数である。単に $d=1$ として用いることもある。

前節の工程特性値の例だと表1のようになる。確率密度関数は

$$\begin{aligned} P(\boldsymbol{y}|\theta,m) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i-\theta)^2}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{RSS(\theta)}{2\sigma^2}\right) \end{aligned} \quad (4)$$

表 1: 工程特性変化モデルの仮想的符号

モデルの記述		
パラメータタイプの指定	パラメータ値の指定	モデルの下でのデータの記述長
m_0(変化なし)	不用	$-\log P(y\|\mu,m_0)$
m_1(変化あり)	μ'	$-\log P(y\|\mu',m_1)$

となる。ただし、$m=m_0$のとき$\theta=\mu$、$m=m_1$のとき$\theta=\hat{\mu}'$として、$RSS(\theta)$ は残差平方和$\sum_i(y_i-\theta)^2$である。結局、各記述長は以下となる[2]。

$$\begin{aligned} L(\boldsymbol{y},m_0) &= \frac{\log e}{2\sigma^2}RSS(\mu) + \frac{n}{2}\log(2\pi\sigma^2) + l_{m_0} \\ L(\boldsymbol{y},m_1) &= \frac{\log e}{2\sigma^2}RSS(\hat{\mu}') + \frac{n}{2}\log(2\pi\sigma^2) \\ &\quad + \frac{1}{2}\log\frac{en}{\sigma^2} + l_{m_1} \end{aligned}$$

両モデルのいずれが適当かは、次の値が正であれば m_0、負であればm_1と判定されることとなる。

$$\begin{aligned} \Delta L &= L(\boldsymbol{y},m_1) - L(\boldsymbol{y},m_0) \\ &= \frac{\log e}{2\sigma^2}n(\hat{\mu}'-\mu)^2 + \frac{1}{2}\log\frac{en}{\sigma^2} + l_{m_1} - l_{m_0} \end{aligned}$$

現実には、経時的に観測されるデータ系列のどこで特性値が変化したかが不明であるし、分散の変化もありうることが普通である。そこで、モデル符号長l_mに関し、変化点からの経過観測数、変化パタン (平均のみの変化、分散のみの変化、両者の変化) などの考慮が必要である [8]。

3. モデル選択アプローチの利点

MDL 原理に従えば、前節のように与えられたMDL基準が最も小さな (記述長最小の) モデルを選び、かつ、パラメータ推定量としてはそのモデルの下での、最尤推定量を用いることとなる。

これによれば、AIC 原理など他のモデル選択基準と同様、従来の統計学における確率基準に従う検定と、最小二乗基準や最尤原理などに基づく推定を統合した使い方が可能となる。つまり、普通の推測統計学における検定と推定を一貫した基準で同時に行なってくれる。

一般的な検定では、確率基準に基づいた判断が行なわれる。つまり、ある方向へ判断を誤る確率 (危険率) を 5%や 1%に押えて、その他の誤り確率の最小化を目指すのが普通の定式化である。たとえば、二群のデータの母平均値の差の有無の検定なら帰無仮説 H_0 : $\mu_1=\mu_2$、対立仮説 $H_1:\mu_1\neq\mu_2$ なる二つのモデルを

[1]最尤推定量については [3] など数理統計学の教科書参照

[2]本稿では対数の底は 2 とする。

考えて、帰無仮説が正しいのに対立仮説を採用する確率$\Pr(H_1|H_0)$を危険率として、その逆の誤りを犯す確率$\Pr(H_0|H_1)$を最小化する。

しかし、この考え方はモデルが3つ以上多数ある場合の多重比較を行なおうとすると困難に直面する。誤りのパタンが組合せ的に増加するため、どんな誤り方の確率を管理するかを決めるのが難しくなる。

また、たとえば、母平均値の差に関する判断をしようとすると、標本平均の差や群間分散などに基づいて誤りパタンごとに検定統計量を構成することになるが、これらは独立にならないことが多く、確率基準で判断手続きを定める上で必要な確率計算が複雑になってしまう。このような状況として次の例を上げよう。

4. 標本のプーリング問題

複数の条件(k個)の下で実験を行なって連続特性値が得られた場合、それらの間で母平均に違いがあるかを知りたい場合を考えてみよう。このようなデータは、従来、分散分析を行なって、F検定をすることが最も一般的である。つまり、帰無仮説$H_0:\mu_1=\mu_2=\cdots=\mu_k$を仮定して残差を計算し、この仮定の下では非常に少ない確率でしか生じ得ないデータが出ていたら、帰無仮説を棄却して母平均値の間に違いがあると判断することになる[1]。

ところが、このやり方では、具体的にどのような母平均値間の違いがあるかについては判断してくれない。H_0が棄却された場合でも、一部の母平均には差がない場合も考えうる。考えうる母平均のプーリングパタンは、k母平均が

- すべて異なる :1通り
- ある一対のみが等しい :$\binom{n}{2}$通り
- ある二対のみが等しい :$\binom{n}{2}\binom{n-2}{2}/2$通り

⋮

- すべてが等しい (H_0) :1通り

のパタンだけある。このどのパタンになるかという情報を推測しようとすると、前節で述べたような多重比較の難しい問題になってしまう。一方、MDL原理に従えば、k個の母平均の違いに関する考えうるすべてのモデルを考えて、それらのモデルに関するMDL基準を算出し、単純に、それらの最小の値を取る母平均モデルを採用すれば良いことになる。

この方針は、比較すべきモデルが、あまりに多過ぎて非現実的に思われるかも知れないが、実際に具体的データが与えられた場合を考えると、適当な前提の下では、パラメータの最尤推定を行なったときに明らかにMDL基準が悪くなるモデルを、始めから考慮対象から外すことができる。

特に、母平均の間に順序制約がなり立つ場合には考慮対象となる母平均の大小関係パタンが制限されて、有効な方法論となる[2]。たとえば、投薬しても悪影響を及ぼす可能性があり得ないことが知られている医薬品の有効性の実験で、投薬量を増やしていくk水準において薬効を測定した場合などである。このとき考えられる母平均の大小関係パタンは、投薬量順に並べたk水準の隣合ういずれの間で違いがあるかないかだけのパタンとなり、合計2^{k-1}通りとなる。

このように、MDL原理を用いると多重比較の問題もすっきりと定式化できることになる。ただし、このままだと、具体的な問題に適用した場合、モデル選択の誤りの確率が考慮されていないという問題がある。この点について次節で考えてみよう。

5. モデル選択の誤り確率

前節で取り上げた順序制約のあるk群のプーリングの問題で2^{k-1}個すべてのモデルについてl_mを等しいとおいて、モデル記述長の項の効果がなくなるようにした場合を考えてみよう。これは、各モデルの価値について何の知識ももたない状況であるとすれば、最も基礎的な方法であると考えられる。

このとき、どのような誤り確率が生じるかを、$k=2$の最も単純な場合について調べてみよう。この場合、普通の初等統計学で習う平均値の差の検定が適用できる状況でもある。モデルは2つの水準をプールしてしまう1群モデルと、両母平均が異なるとして分離して扱う2群モデルの二つしかない。

今、母平均の差がΔだけあるデータ数nの2群データから判定手続きが2群モデルと判定する検出率を$P(\Delta,n,\mu)$と表すとする。ここでμは真の全平均の大きさとする。誤りの確率を捉えるには、差があるのに検出に失敗する確率、つまり$\Delta>0$の場合の$1-P(\Delta,n,\mu)$と、本来差がないのに誤って2群モデルを採用してしまう確率つまり、危険率$P(0,n,\mu)$の2つを見れば良い。漸近近似する前の(2)式を直接最適化した場合 :ΔL_0と、漸近近似した形の(3)式の場合 :ΔL_1、および、普通の平均値の差の検定の場合 :Z test について得られる確率を図1に示す。ΔL_0はμの

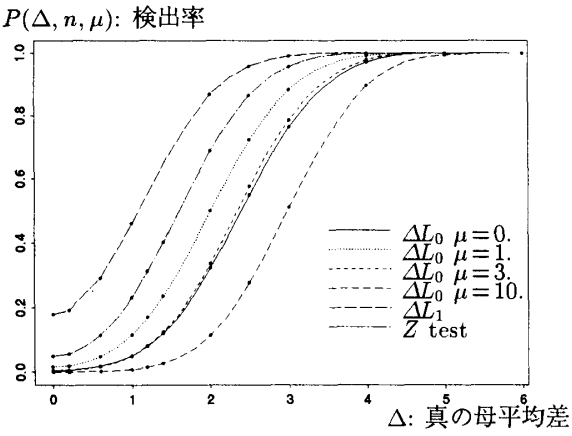

図 1: 2群モデルの検出率 ($n = 6$)

Z test は危険率5%の母分散既知の平均値の差の検定の結果

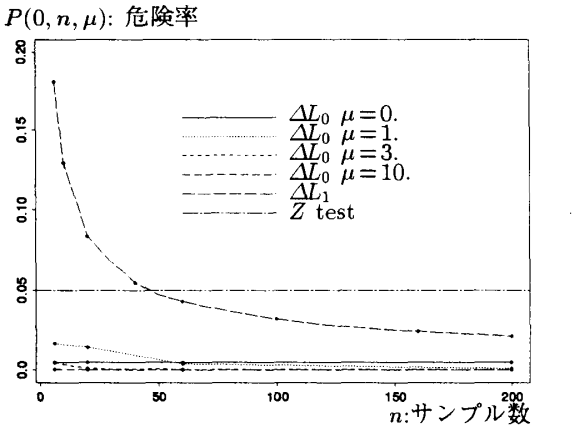

図 2: 危険率 $\alpha_0(n, \mu)$, $\alpha_2(n)$

推定量を符号化するのでμ の大きさに依存する。

どの場合もΔが大きくなるにつれて、2群モデルの検出率 $P(\Delta, n, \mu)$ は同じように増加するが、方式によって位置が左右にずれていることが分かる。これに応じて検出率と危険率のバランスが変わることになる。特にΔL_1では、危険率がかなり高めになっており、普通の平均値の差の検定の結果とかなりずれてきていることがわかる。検出率と危険率のアンバランスは、漸近近似の際に定数項を省いているために生じたもので、定数項を含めたΔL_1は、この場合はほぼ妥当な結果になっている。

しかし、サンプル数が増えると、$\Delta > 0$ の場合の検出率はΔの大きさに限らず 1 に近付いてゆく。また、危険率についても、図2に示すように普通は 0.0 に近付いていく[3]。これは、(3) 式の MDL には、漸近一致性がなり立つためである。

[3] $\mu = 0$ のときのみ、漸近一致性がなく一定の確率となる。

6. モデルの事前分布の定め方

MDL 基準 (1) 式の最後のモデル記述長の項 l_m についてはいままで具体的には議論してこなかったが、これの設定には自由度が残されている。つまり、判定基準としては恣意性が残されている。逆に考えれば、モデル記述長を自然に考えることができるような問題設定ならば、AIC 基準などと違い、パラメータ数のみで表現できない構造的なモデルの相違も自然とモデル選択の対象とできる可能性がある。

前節の危険率と検出力のアンバランスに関しては、具体的な目標があれば、l_m をすべてのモデルについて等しくせず適当に調整してやることで、バランスを変えることができる。たとえば、従来の検定と同等の危険率でモデル選択を行ないたいという目的があれば、H_0の場合のモデルの記述長のみを特別扱いし、調節する方法も与えうる [2]。

理論的には、このような調整を与えることは次のように解釈できる。MDL 基準（符号長）は劣確率分布としての意味を持つ [9]。つまり、データ$\boldsymbol{y}$、モデル m に対して、次のような事前同時分布を考えて、これに対応した最短符号を考えていることとなる。

$$\begin{aligned} Q(\boldsymbol{y}) &= 2^{-L(\boldsymbol{y},m)} \\ &= P(\boldsymbol{y}|\hat{\boldsymbol{\theta}}, m) \cdot 2^{-l(\hat{\boldsymbol{\theta}}|m)} \cdot 2^{-l_m} \end{aligned} \tag{5}$$

この最後の $g(m) = 2^{-l_m}$ の項は、モデルタイプに関する事前分布であり、これにより従来の統計学のベイズ流の解釈ができる。

結局、MDL 基準のモデル記述長に関しては、次の2つの見方ができる。第1には、問題設定に対応して、各モデルに対する妥当な事前知識（事前確率）を想定できれば、それに沿ったモデル選択が可能となる。特に、すべてのモデルを平等な候補として考えて良い場合には、それらに等しいモデル記述長 l_m を与えることになり、この項はMDL 基準の大小比較でキャンセルされるので、モデル記述長は無視できることとなる。ただし、この場合には、MDL 基準を導出する際、近似するため切ってしまったパラメータ記述長の定数項の詳細な検討が必要となる。

第2は、問題解決上必要なモデル選択の誤り確率 $\Pr(m'|m)$ の設定が与えられる場合である。この場合はその誤り確率が実現するように、モデルクラス上の事前確率分布 $g(m)$ を定める。これを用い、MDL 基準のモデル記述長を $-\log g(m)$ と設定することで、与

えられたモデル選択の誤り確率に従った判定手続きを設定してやることが可能となる。この場合は、導出の際に漸近近似などで無視していた定数項は事前確率分布に吸収されてしまうので、厳密に求める必要はない。

第1節で述べたように、現実問題では現実のコストを考慮したモデル選択が必要となる。このとき、コストがモデルのパラメータタイプの誤り選択確率のみの関数であれば、上の第2のアプローチによりモデルクラス上に事前確率分布をもとめてやることにより、意思決定問題としてのモデル選択が可能となる。しかし、現実の決定の損失は、モデルの選択誤り確率だけでは計り切れず、その時の母数の値など、沢山の状況に依存して決まるものであろう。このような場合に、どのような定式化を行なえば良いかについては今後の課題とされるところが多い。

7. データ量とモデルクラス

このようにMDL基準に基づく、データを短く圧縮するのに最適なモデルは、データ数(データの持つモデルに関する情報)に見あった程度に詳しいモデルになる(overfitが適度に抑制される)ことが期待される。つまり、データ中のモデルに関する情報量が多い時には真のモデルを当ててほしいが、情報量が少ない時は、適度に要約して、簡潔なモデルを提案してもらいたいと考える場合に適当な基準であることとなる。データ分量に見あう以上の複雑さを拒否するという考え方となる。

現実の決定問題の場合に、真のモデルというものが想定できるかどうかは議論の分かれることかも知れないが、一応、真の確率モデルがあるとしてみよう。モデル選択の結果、真のモデルとズレの大きなモデルが選択される場合の原因には、どのような可能性があるか考えてみよう。

まず、第1には選択の考慮対象としたモデルクラスが不適当である場合である。つまり、考慮対象としたモデルクラスが真のモデルを含まず、真のモデルに最も近いものでもズレが大きい場合である。これを避けるためには、考慮対象とするモデルクラスを十分広くすれば良いようにも思えるが、極端な場合、データ数より多い自由パラメータを持つモデルの推定は不可能だし、次項のような問題も発生する。問題の現象に対応したよいモデル化が必要になる。

第2にはデータが足りない場合である。MDL基準はデータが少ない場合には、前述のようにデータ圧縮に不必要なほどの複雑なモデルは、それが真のモデルであるかどうかに関わらず拒否してしまう。よって、真のモデルの複雑さとノイズレベルに比べて、データ数が少ない場合には、真のモデルからはずれた簡潔なモデルが選ばれることになる。

8. さいごに

データを短く記述できるモデルが良いモデルであるという発想は概念的には大変理解しやすいが、序論で述べたように、具体的な問題では、最小記述長基準でモデル選択をすると現実の利得(評価関数)とどう関連してくるか？が問題となる。今後、現実の利得を記述長に上手に還元する考え方の開発が必要とされている。

また、問題の定式化が複雑になると確率モデルクラスが大きくなり、MDLによるモデル選択は離散的最適化問題となる。これに適した最適化の手法の開発が必要されるであろう。

参考文献

[1] 広津千尋, 分散分析, 教育出版, (1976)

[2] Hoshino, N. and Seki, Y., A Test based on MDL Criterion for Comparing Increasing Dose Levels with a Zero Dose Control, *Communications in Statistics: Theory and Methods*, *25*, 8, (1996)

[3] 稲垣 宣生, 数理統計学, 裳華房, (1990)

[4] 韓太舜, 情報圧縮とはなにか, 数理科学, 290, 5-15, (1987)

[5] 韓太舜, 小林欣吾, 情報と符号化の数理, 岩波書店, (1994)

[6] Rissanen, J., Modeling by shortest data description, *Automatica*, *14*, 465-471, (1978)

[7] Rissanen, J., A Universal Prior for Integers and Estimation by Minimum Description Length, *The Annals of Statistics*, *11*, 2, 416-431, (1983)

[8] 関庸一, 橋本巧, MDL基準に基づく正規母集団変化時点検出に関する研究, 日本経営工学会誌, *47*, 3, (1996)

[9] 山西建司, 韓 太舜, MDL入門:情報理論の立場から, 人工知能学会誌, 7, 3, 427-434, (1992).