

階層構造をもつ時系列モデルによる情報処理：ベイズ流自己調節と並列計算

Information processing based on hierarchical time series models: Bayesian self-tuning and parallel computing

樋口 知之[1]

Tomoyuki Higuchi[1]

[1] 統数研

[1] Inst. Stat. Math.

<http://www.ism.ac.jp/higuchi/>

時系列モデルの場合広い範囲のモデルが、その数理的構造から一般状態空間モデルと呼ぶ漸化式によって表現できる。最近では、最も自由なモデリングを保証する、自己組織型状態空間モデルの開発と、それを実現する計算集約的な数値計算技法の考案が精力的に行われている。利用するモデルを高層化することで自由なモデル構成が可能になり、その結果、不適切なモデルの利用のために妨げられていた情報の抽出が可能になるようすを具体的例を通じて解説する。また、利用するモデルのクラスを上げるごとに、人間の判断が必要だった手続きが徐々に自動化され、結果として時系列データの自動処理が可能になっていくプロセスを説明する。

データには誤差が含まれるので、その適切な処理によってはいじめて、本質的・一般的な部分とたまたまそのデータだけに生じた偶然の部分に分離できる。科学研究における観測誤差の処理は既に前世紀より行われてきた。しかしながら、非定常性・非線形性などを伴う複雑・大規模な多変量現象の解析においては、比較的簡単な理論モデルによって現象を完全に記述し、現実のデータとの差異は独立な観測誤差とみなせることは、ほとんどないといつてよい。その場合、現在の知識では表現できない部分とそれ以外の部分を適切なモデリングに基づいて合理的に分解することによって、新しい発見の可能性が生じる。先端科学領域においては、未知の現象はしばしば誤差限界ぎりぎりのところに発現する。このような中で、未知の部分だけを単に観測誤差とみなす処理では発見の緒を見出すことはできない。先端領域における発見のためには、その未知の部分に対する我々の期待を積極的にモデルの形で表現し、能動的に情報抽出を行うことが必要である。いうまでもなく、このような統計的モデリングにおいては「適切な」モデルの利用が本質的に重要である。データが誤差を含まない場合には、あるいはデータ解析の目的が単にそのデータを精密に表現することにあるならば、「記述能力の高い」モデルを用いればよい。しかし、実際に多くの場合には、そのデータ解析を通して、より「普遍的な」知識の抽出・発見を目指している。統計的モデリングにおいては、その本質部分を予測能力という観点から捉えてきた。よいモデルを求めるための要素としては、適切なモデル族とモデル評価基準が必要であるが、さらにそれを現実の問題に適用可能とするためには、効率的な計算法も必要となる。

本発表では、統計モデルのなかで、大量データの自動処理の領域で有用な時系列モデルを取り扱う。時系列の自由なモデリングのためには、一般にはデータ数 N に比例した未知数を持つモデルが必要となる。このようなモデルでは、単純な計算法では $O(N^3)$ の計算法が必要であり、実用化は困難である。一般的にいえば、柔軟性・表現能力の高いモデルの採用は、未知数の爆発的増加を誘因し、その推定にかなりの計算資源を必要とする。モデルの数理的構造に踏み込んだ巧妙な計算法なくしては、その安定した推定値を得ることは不可能の場合もある。高度なモデルにもとづいて大量データを処理する状況では、さらに効率的な計算法の開発が必須である。時系列モデルの場合、広い範囲のモデルが、その数理的構造(具体的にはマルコフ性)から一般状態空間モデルと呼ぶ漸化式によって表現できる。時系列モデルの場合は、普通データが連続値をとる状況を想定しているが、離散値しかとりえない対象---例えば、DNA の A,T,G,C---にもこのモデルは適用できる。この一般状態空間モデルの離散値版は、隠れマルコフモデルと呼ばれ、DNA sequence の解析や、音声認識などの野で、表現能力の高いモデルとして活発に利用されている。この漸化式による表現方法は、モデルの表現能力を保持しながら、未知数の推定に $O(N)$ の計算法---逐次フィルタおよび平滑化のアルゴリズム---を実現する。このように一般状態空間モデルは、自由なモデリングのプラットフォームであると同時に、効率的な計算法を生成することができる。上述したように、 $O(N)$ に減らされた計算量であるが歴史的には、さらなる自由なモデリングの追求が、より複雑・高度なモデルの未知数推定を実現する計算技法の開発を要請してきた。最近では、最も自由なモデリングを保証する、自己組織型状態空間モデルの開発と、それを実現する計算集約的な数値計算技法の考案が精力的に行われている。本発表では、利用するモデルを高層化することで自由なモデル構成が可能になり、その結果、不適切なモデルの利用のために妨げられていた情報の抽出が可能になるようすを具体的例を通じて解説する。また同時に、利用するモデルのクラスを上げるごとに、人間の判断が必要だった手続きが徐々に自動化され、結果として時系列データの自動処理が可能になっていくプロセスを説明する。すなわち、統計モデルの高層化により、大量時系列の自動処理と新しい情報抽出の二つが同時に

可能になることを明らかにする .