数理解析研究所講究録
第2043巻 2017年 1-5

繁殖行動の同調モデルとパラメータの AKB 推定

統計数理研究所　島谷健一郎

Synchronicity model for reproduction and the AKB estimation of parameters

The Institute of Statistical Mathematics　Ichiro Ken Shimatani

1. 序

動物の同調行動は、確率論的不確実性を入れたモデルを立て、パラメータを適当に与えてシミュレーションを遂行すること自体は決して難しいことではない。しかし、実データがあるとき、データに基づいた未知パラメータ推定や最適化、複数のモデルを考案したときのデータに基づく比較 (数式や不確実性の与え方の異なるモデルのパラメータ値は何らかの意味で最適化されていないと不公平) は、一般には極めて困難である。

このような、シミュレーションは容易だが実データとの照合に確立された手法を有さないモデルにおける、最も手っ取り早く簡便な解決法に、approximate Bayesian computation（略して ABC 法、または近似ベイズ法）がある。荒っぽくは、様々なパラメータ値（一般にはパラメータは複数あるのでベクトル）で大量にシミュレーションを行い、実データと近い人工データを生成したパラメータを選ぶだけである。こんな「モデルやデータ解析の素人のやりそうな幼稚な手法」を、ベイズの枠組みで数学的に(近似的に)正当化したものが ABC 法である。

初めてこの荒っぽい手法が本格的に研究に用いられてから、手法は繰り返し改良を加えられ、数学的にも整備され、一口に ABC 法といっても数多くのアルゴリズムが乱立している。そんな中、Fukumizu (2013) が提唱した kernel Bayes 理論を ABC 法に応用する approximate kernel Bayesian algorithm (略して AKB 法) は、いくつかの点で他の計算法(と数学的根拠)を凌駕する。

本稿では、AKB 法の計算アルゴリズムを紹介し、そのひとつの応用例を述べる。

2. ABC 法と AKB 法

ABC 法では、与えられた実データと、それを説明することが期待されるモデル（シミュレーションの実行(人工データ生成)が容易で、（未知）パラメータを含む）に対し、以下の手順でパラメータ推定を行う。

1. パラメータに事前分布を設定する。
2. 事前分布からランダムに選んだパラメータを用いて人工データを生成する。
3. 実データと人工データを比較し、実データに「近い」人工データを生成したパラメータを抽出する。
4. 抽出されたパラメータたちは、事後分布の近似となる。
5. パラメータの推定値を 1 個の数値で代表させたいときは、抽出されたパラメータ（事後分布）の標本平均、中央値、最頻値、などを目的に応じて用いればよい。
6. モデルによる予測は、事後分布からランダムに抽出したパラメータを用いたシミュレーションを行うことで、パラメータ推定の分散を考慮した推定ができる。

実際にこの方法を実行する場合、実データとすべての数値が近い人工データは滅多に作成されない。そこで、妥協案として採用されるのが、実データに対しいくつかの集約統計量(summary statistics)を計算し、それらが近い人工データを生成したパラメータを抽出するというものである。

集約統計量の中には、そのモデルを適用する範囲では元データと同じ情報量を持つ十分統計量(sufficient statistics)というものがある。例えば、連続的な数値データに 1 次元の正規分布モデルを適用する場合、全データの数値を用いなくても、平均と分散という集約統計量だけから未知パラメータ（平均と分散）を最尤推定できる。このとき、平均と分散の対は、正規分布モデルの十分統計量と呼ばれる。

たいていのデータとモデルでは、十分統計量は知られていない。そこで、データの特性を表現しそうな統計量を工夫することになる。

事前分布は恣意的に定める。過去の研究から尤もらしい部分の一様分布にしたり、尤もらしいあたりに集中する平均と分散の正規分布にする。あるいは、全実数や全正の数などの無情報事前分布にする。

ABC 法には、集約統計量による情報消失と事前分布の恣意性以外に、以下のような問題がある。

1. 集約統計量だけで比べるにしても、実データに近い人工データを作るには膨大なシミュレーション回数を要する。
2. どのくらい「近い」人工データを作ったパラメータを事後分布の近似に使う

か、規準があいまいである。

3. 集約統計量は多いほうがよいが、強い相関を持つ統計量があると、それが表すデータの特性に偏ったパラメータ抽出を招く。

これに対し、AKB 法は以下のような手順を踏む。

・事前分布から選んだパラメータで生成した人工データと実データからカーネル函数を用いて決めるウェイトでパラメータのウェイト和をとると、事後分布の平均の推定値となる

この手順には、以下のような利点がある。

1. すべてのデータを使うので、少ないシミュレーション回数で済む。
2. 「データの近さ」に関する恣意性は消えた（代わりにカーネル函数の幅を決める恣意性が加わったが、概してこの恣意性の結果への影響は小さい）。
3. 計算の中で Ridge 回帰補正という手法をはさむため、集約統計量間の相関の影響は軽減されると期待できる（ここでも新たに恣意的に決めたパラメータを用いるが、概して結果への影響は小さい）。

さらに、ABC 法により抽出されたパラメータたちが事後分布の近似であるためにはサンプルを増やすと事後分布に収束することを証明するのが望ましいが、AKB 法では、シミュレーション回数を増やすと、ウェイト和が事後分布の平均に収束することを証明している。さらに、平均だけでなく信頼区間も計算可能で、その計算手順も公表されている。

次節で、事後分布の平均の推定法を述べる。

3. AKB 法の手順

j 番目($j = 1, \cdots, n$, n はシミュレーション回数)にランダム抽出された d 次元のパラメータセットを $u = (u_1^j, ..., u_d^j)$ とする。実データから計算した m 個の集約統計量を $\bar{S} = (\bar{S}_1, ..., \bar{S}_m)$、$j$ 番目のパラメータが生成した人工データの集約統計量を $\bar{S}^j = (\bar{S}_1^j, ..., \bar{S}_m^j)$ とする。 これらを、人工データの平均と標準偏差で標準化したベクトルを $\bar{s}$, $\bar{s}^j$ とする。

このとき、k 番目 ($k = 1, ..., d$) のパラメータの事後分布の平均 $\bar{u}_k$ は、w_j というウェイトによるすべてのパラメータ $\{u_k^j\}$ のウェイト和として以下のように推定される。

$$\bar{u}_k = \sum_{j=1}^{n} w_j u_k^j$$

ここで、ウェイト w_j は

$$\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} = \left(\begin{pmatrix} k(\bar{s}^1,\bar{s}^1) & k(\bar{s}^1,\bar{s}^2) & \cdots & k(\bar{s}^1,\bar{s}^n) \\ k(\bar{s}^2,\bar{s}^1) & k(\bar{s}^2,\bar{s}^{21}) & \cdots & k(\bar{s}^2,\bar{s}^n) \\ \vdots & \vdots & \ddots & \vdots \\ k(\bar{s}^n,\bar{s}^1) & k(\bar{s}^n,\bar{s}^2) & \cdots & k(\bar{s}^n,\bar{s}^n) \end{pmatrix} + n\varepsilon I_n \right)^{-1} \begin{pmatrix} k(\bar{s}^1,\bar{s}\}) \\ k(\bar{s}^2,\bar{s}\}) \\ \vdots \\ k(\bar{s}^n,\bar{s}\}) \end{pmatrix} \qquad \text{(S6)}$$

で定める。ここにある $k(x, y)$ はカーネル関数で、よく使われるのは Gaussian kernel;

$$k(x,y) = \exp(-\sum_{i=1}^{m}(x_i - y_i)^2 / 2\sigma^2) \qquad \mathbf{x} = (x_1,...,x_m), \mathbf{y} = (y_1,...,y_m)$$

である。σ はカーネル関数のバンド幅と呼ばれる定数、I_n は n 次元の単位行列、ε は正の定数である。

概して σ の推定値に与える影響は小さく、簡便な決め方に

$$\sigma^2 = median(\sum_{i=1}^{m}(\bar{s}_i^{\,j} - \bar{s}_i^{\,h})^2; j,h = 1,...,n)$$

というものがある。

正の定数 ε を大きめに定めると生成された人工データに依らず同じくらいの推定値が得られ、ε を小さくすると、推定値は不安定になりがちである。最適な値は cross validation のような方法で決めることが望ましいが、おおまかに事後分布の平均を推定することが目的なら、推定値が比較的安定するなるべく小さい値くらいに決めてもあまり問題はなかろう。

4. 同調して繁殖に向かうモデル

同調のような個体間相互作用があると、個体ベースのデータがあっても、それらは互いに独立でないため、広く使われている統計手法による(同調の強さなどの)パラメータ推定ができない。そんなモデルの一例として、以下のようなものを考える。

ある動物集団のメスたちが繁殖に最も適する時点は、季節的に限定はされるが、個体差もあるため、平均 u, 標準偏差 s の正規分布に従って散らばっていると仮定する。ただ、最適時点前に他のメスが繁殖に向かうと、同調して最適時点前なのに繁殖に向かってしまうかもしれない、とする。

モデルとしては、まずメス t ($t = 1, \cdots, N$, N は個体数) の最適繁殖時点 u_t を、平均 u, 標準偏差 s の正規乱数で与える。それから、時点を変化させていく。最初は最も早い最適繁殖時点 u_1 のメスがその時点で繁殖に向かう。ところが、他のもっと遅い最適繁殖時点 u_i ($i = 2, \cdots, N, u_i \geq u_1$) のメスも、$u_i$ と u_1 の差に反比例する確率で同調して繁殖に向かってしまうとする。反比例は例えば

$$e^{-\frac{(u_i - u_1)^2}{2a^2}}$$

という式（a は未知パラメータ）で与えるとする。

以降、まだ繁殖に向かっていないメス t は、順次、u_t が訪れたら繁殖に向かうが、その前に他のメス j $(u_i > u_j)$ が繁殖に向かうと、式 $e^{-\frac{(u_i-u_j)^2}{2a^2}}$ に従って同調して繁殖に向かうとする。

この個体ベースモデルをシミュレーションで動かすことは容易である。しかし、実際の繁殖メス数に関する時系列データが与えられたとき、どのようにして未知パラメータ a（と u と s と N）を推定すればよいだろう。統計手法で最も普通に使われるのは最尤法である。そこでは、データが生成される確率(密度)に相当する尤度を未知パラメータの関数で表し、その最大化を図る。しかし、同調が入ると産卵数の時系列データは互いに独立でない(同調により大量のメスが繁殖に向かったら、直後の繁殖メス数は減る。つまり、隣接時点のデータ間に負の相関が出るため独立でない)。時系列データ全体の同時分布を求める必要があるが、これを解析的に導出することは難しい。近年よく使われる階層ベイズモデルにおいても、尤度式の導出は必要であり、解決にはならない。

こういう状況において、ABC 法は有効である。様々な未知パラメータの組み合わせでシミュレーションを行い、実データに近いデータを生成したパラメータだけ残せば、それが事後分布の近似になるからである。そして、上述したように、AKB 法では、「実データに近い人工データを生成した」パラメータを選ばなくても、事後分布の平均を推定できるのである。

論文 Koizumi and Shimatani (2016)では、北海道の河川で産卵するオショロコマの 30 集団の産卵床数データに、このモデルに観察過程やオショロコマの繁殖特性を加味したモデルを適用した。同時に、帰無モデルとして、同調のない、単なる正規分布モデルも適用し、両者の適合度を調べた。その結果、4 集団について、正規分布モデルでは産卵床数データを説明できないが同調を入れたモデルでは説明できた。このことから、このオショロコマのメスは同調して産卵に向かうと推察した。

参考文献

Fukumizu K, Song L, Gretton A. 2013. Kernel Bayes' rule: Bayesian inference with positive definite kernels. J Mach Learn Res. 14:3753-3783.

Koizumi, I. and I. K. Shimatani. (2016) Socially induced reproductive synchrony in a salmonid: an approximate Bayesian computation approach. Behavioral Ecology (in press) doi:10.1093/beheco/arw056.

Nakagome S, Fukumizu K, Mano S. 2013. Kernel approximate Bayesian computation in population genetic inferences. Stat Appl Genet Mol Biol. 12:667-678.