組織的カンニング手法としてのデータ同化

上野 玄太
情報・システム研究機構 統計数理研究所

平成16 年 11 月 30 日

1 はじめに

はじめに簡単な例として，最小二乗法による直線のあてはめを復習しよう．データ同化のエッセンスはすべてこの例に含まれている．t-y 平面上にデータ $(t, y_t), t = 1, \cdots, T$ が与えられているとして，このデータに直線

$$x_t = at + b \tag{1}$$

をあてはめることを考える．直線を規定するパラメータ a, b は，直線とデータの差の二乗和を最小にするように定める．残差の二乗和を

$$\begin{aligned} J(a,b) &= \sum_{t=1}^{T} (y_t - x_t)^2 \\ &= \sum_{t=1}^{T} (y_t - at - b)^2 \end{aligned} \tag{2}$$

と書くと，$J(a,b)$ が最小となる a, b では $J(a,b)$ は傾きがゼロ，すなわち

$$\frac{\partial J}{\partial a} = 0, \qquad \frac{\partial J}{\partial b} = 0 \tag{3}$$

が成り立つはずである．いま，式 (2) で与えられる $J(a,b)$ を具体的に代入すると，条件式 (3) は a, b に対して解析的に解けて，

$$a = \frac{\sum_{t=1}^{T} t y_t - \frac{T+1}{2} \sum_{t=1}^{T} y_t}{\frac{1}{3}(T-1)T(T+1)}, \tag{4}$$

$$b = \frac{1}{T} \sum_{t=1}^{T} y_t - \frac{T+1}{2} a \tag{5}$$

を得る．あとは右辺の値を計算することで，データにあてはまる直線を表す a, b が求められる．

さて，データ同化とは，要するに観測データにモデルをあてはめることである．上の例では，モデルとして直線を持ってきてデータへのあてはめをした．直線のかわりにシミュレーションモデルをデータにあてはめる手法のことをデータ同化という．

しかしそもそも，シミュレーションに関しては従来から脈々と受け継がれている基本理念がある．すなわち，シミュレーションとは，確立された物理学の基本法則にのっとって物理の素過程を明らかにするための道具であり，観測データとは独立に進めるべきものである，というものだ．さらに続けて，データとつき合わせるにしても，シミュレーションの計算がすべて完了してから行うべきものとも考えられている．この独立性を保っておかないと，どこが素過程として本質的な部分かがわからなくなってしまうというのがこの理念の根拠だ．ところが，データ同化とは，そういった理念とは反する姿勢の手法である．

理念に反してまでデータ同化を行うのにはそれなりの理由がある．その理由とは煎じ詰めると，シミュレーション結果は観測データを正確に再現していない点である．その原因は当然シミュレーションモデルの不備にある．シミュレーションを走らせる際の初期条件・境界条件，モデリングの際に無視した異スケールの物理，1,2 次元性などの仮定，グリッドの大きさ，経験的公式，などがモデル不備の内訳である．ところが，正確に再現しないと価値がないシミュレーションというものが存在する．天気予報が好例だ．そこで，従来のシミュレーションの理念には反するが，データを参考にしながらシミュレーションモデルを修正し，現象の正確な再現を図るというのがデータ同化のねらいである．理念上ではやってはいけないことではあるが，データのカンニングのプロセスをシミュレーションの実行に抱き合わせることで，現象の正確な再現という意味で価値がある「エセ」シミュレーションの結果を作り出すわけだ．

以後，シミュレーションモデルの雛形として，冒頭で例として挙げた直線モデル $x_t = at+b$ を考えることとする．ここで a は物理方程式に含まれるパラメータ，例えば拡散係数や比熱比などを想定してほしい．また b は，初期条件・境界条件を表すものと考える．シミュレーションとして，最も理念にかなっているのは，a も b もデータを見ないで決め，時間発展の計算結果を観測データ y_t とつき合わせるというものである．このとき，

$$x_t \;=\; at + b \tag{6}$$

をストイック・シミュレーションと呼ぼう．データをカンニングするなどもってのほか，自力で立ち上げたシミュレーションモデルである．

シミュレーションはデータとは独立とは言いながらも，実はこれまでにもデータのカンニングをしているシミュレーションは行われている．例えばシミュレーションモデルは立てたものの，初期条件や境界条件の設定値に自信がない場合は多い．そこで，初期条件や境界条件については，相当する観測データの値をそのまま代入してシミュレーションの計算を進める場合がある．磁気圏のグローバルシミュレーションを行う際に，太陽風のパラメータは衛星データを使うことがこれにあたる．ここで考えている直線モデルでいえば，切片 b の値はそれほど自信がないので，時刻 $t = 0$ でのデータ y_0 を使って b を置き換え，

$$x_t \;=\; at + y_0 \tag{7}$$

とするのである．

この程度ならばカンニングなどと大げさにいわなくてもよいかもしれない．ところが，データ同化が目指すカンニングでは，初期条件・境界条件はデータ y_0 をそのまま代入するだけではなく，全データ $y_1, \cdots, y_T$ に合うように切片 b を調整する．この状況は，仮に初期条件 x_0 が y_0 と一致していたとしても，シミュレーションモデルが与える $x_1, \cdots, x_T$ ではその後の $y_1, \cdots, y_T$ を再現しきれない場合に相当し，シミュレーションではもはや正確な x_t の時間発展が与えられないという「降参」状態の場合といえる．しかしそれでも，半ば悪あがきとして全データ $y_1, \cdots, y_T$ に合うように調整した $\hat{b}$ を用いて

$$x_t \quad = \quad at + \hat{b} \tag{8}$$

を使うならば，x_{T+1} 以降の計算値は改善される可能性があるし，調整された $\hat{b}$ はデータに適応的に設定した「適切な」初期条件と見なすこともできる．

さらに，パラメータ a も自信がない場合も起こりうる．この場合も同様に，a も全データに合うように設定するのである．

$$x_t \quad = \quad \hat{a}t + \hat{b} \tag{9}$$

式 (8) と (9) の $\hat{b}$ は一般に異なる値となるが，混乱の恐れはないので区別はしない．この場合も同様に，調整された $\hat{a}, \hat{b}$ は「適切な」パラメータ，初期条件と考えられる．

それでは，そのようなカンニングをどうやって行うのか．直線のあてはめなら容易なことだが，シミュレーションモデルのあてはめなど想像するだけでも大変そうだ．ここで急がずに，組織的なカンニングの作戦を立てることにしよう．まず，自分はシミュレーションコードを持っているとする．いくらシミュレーションが正確でないとはいえ，信頼できる方程式なども含まれているわけだし，データにもいろいろ誤差は入り得るだろうから丸のみするのもどうかと思う．そこで，データを参考にして，シミュレーションのどの部分を修正すべきかの計画を立てるのが第一歩である．次に，その計画を実行するための手法，すなわちどうやってカンニングをするのかについての吟味にうつる．そして最後に，どの程度までカンニングを認めるのが妥当か，という問題も生じてくる．

本稿では，このカンニング手法の基礎として，データ同化におけるモデル，アルゴリズムについて，ベイズ統計を基礎として述べることとする．

2　モデル

X を事象としたとき，X が起こる確率を $p(X)$ と表すことにする．さらに，X と Y が同時に起こる事象を X, Y ($X \cap Y$ と同じ意味)，Y が起こったもとで X が起こる事象を $X|Y$ と表すと，条件つき確率

$$p(X|Y) \quad = \quad \frac{p(X,Y)}{p(Y)} \tag{10}$$

が定義される．この定義から明らかに，

$$p(X,Y) \quad = \quad p(X|Y)\, p(Y) \tag{11}$$

が成り立つ (これを乗法定理という)．さらに，X, Y という事象は Y, X という事象と等しいので，

$$p(X,Y) \;=\; p(Y,X) \;=\; p(Y|X)\,p(X) \tag{12}$$

式 (11)，(12) を等値して，

$$p(X|Y) \;=\; \frac{p(Y|X)\,p(X)}{p(Y)} \tag{13}$$

を得る．これがベイズの定理である．

データへのモデルのあてはめを，このベイズの定理にもとづいて解釈することができる．X をモデルのパラメータ，Y をデータとすると，データが与えられたときのモデルのパラメータの条件つき確率は，ベイズの定理から

$$p(X|Y) \;=\; \frac{p(Y|X)\,p(X)}{p(Y)} \tag{14}$$

$$\propto\; p(Y|X)\,p(X) \tag{15}$$

となる．比例部分は，分母にあるデータの確率 $p(Y)$ は，手元にあるデータが確定している以上は定数とみなせることによる．ここで，左辺の $p(X|Y)$ を事後確率，右辺の $p(Y|X)$，$p(X)$ をそれぞれ尤度，事前確率という．事後確率とは，データが与えられたときのモデルの確からしさを表し，事前確率はどんなモデルを選ぶかを，尤度は選んだモデルはデータをどれくらい再現しうるかを表す．事後確率には，X の誤差などを含めた情報が含まれるが，さしあたって X のもっともらしい値を選ぶとすると，確率を最大とする X を選ぶことが考えられる．

この考え方は違和感があるかもしれないが，実はすでに多くの人が行っていることを仰々しく書いただけのものだ．冒頭に登場した，最小二乗法による直線のあてはめの例で見てみよう．モデルは直線なので

$$x_t \;=\; at + b \tag{16}$$

と書け，モデルとデータの間には

$$y_t \;=\; x_t + w_t \tag{17}$$

という関係があるとする．いま，全タイムステップでの y_t をまとめて $y_{1:T} = \{y_1, \cdots, y_T\}$ と書く．モデルの事前確率として，傾き a と切片 b は何を選んでもよいということで一様確率分布

$$p(a,b) \;=\; \text{const.} \tag{18}$$

とし，データとモデルの差 w_t は正規分布に従うと仮定すると，尤度は

$$p(y_{1:T}|a,b) \;=\; \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y_t - at - b)^2}{2\sigma^2}\right] \tag{19}$$

となる．式 (15) の比例関係式に (18)，(19) を代入すると，

$$p\,(a,b|y_{1:T}) \;\propto\; p\,(y_{1:T}|a,b)\,p\,(a,b) \;\;\rightarrow\;\; \max \tag{20}$$

$$\Longleftrightarrow \qquad \sum_{t=1}^{T}(y_t - at - b)^2 \;\;\rightarrow\;\; \min \tag{21}$$

であることが確認できる．これは取りも直さず，事後確率を最大とする a,b が残差の二乗和を最小にすること，すなわち最小二乗法により得られる a,b と一致することを示している．

次の例として，直線のあてはめに最小二乗法を使うのではなく，絶対偏差の最小化を使う場合を考えてみよう．これは，

$$\sum_{t=1}^{T}|y_t - at - b| \;\;\rightarrow\;\; \min \tag{22}$$

とする a,b を求めるもので，外れ値に大きく影響されずに直線を引ける方法として知られている．この場合は，事前確率は a と b の一様分布 (18) のままであるが，残差 w_t はラプラス分布に従うと仮定した場合に対応している．この仮定のもとでの尤度を実際に書いてみると，

$$p\,(y_{1:T}|a,b) \;\;=\;\; \prod_{t=1}^{T}\frac{1}{2\sigma}\exp\left[-\frac{|y_t - at - b|}{\sigma}\right] \tag{23}$$

となる．上の最小二乗法と比較すると，直線のモデルを仮定して a と b に一様分布を与える事前確率の仮定は共通だが，尤度の測り方の仮定が異なっている．

さらに第三の例として，これまで扱ってきた直線モデルについても信憑性が薄い場合を考えてみよう．このときは事前確率を変更して，モデルは基本的には直線だが，少々の折れ曲がりも許すものとして，

$$\begin{aligned} p\,(a,b,x_1,\cdots,x_T) \;\;=\;\; & \frac{1}{\sqrt{2\pi}\tau_1}\exp\left[-\frac{(x_1 - a - b)^2}{2\tau_1^2}\right]\cdot\frac{1}{\sqrt{2\pi}\tau_2}\exp\left[-\frac{(x_2 - x_1 - a)^2}{2\tau_2^2}\right] \\ & \cdot\prod_{t=3}^{T}\frac{1}{\sqrt{2\pi}\tau}\exp\left[-\frac{(x_t - 2x_{t-1} + x_{t-2})^2}{2\tau^2}\right] \end{aligned} \tag{24}$$

とする．ここで τ_1，τ_2，τ は折れ曲がりの程度を制御するパラメータである．折れ曲がりを許すことをシミュレーションとの対応でいえば，シミュレーションモデルを構成している方程式自身がそれほど絶対的なものでなく，データに応じて柔軟に変形するモデルを認めることといえる．

以上をまとめると，モデルのあてはめとは，事後確率を最大とするモデルを選ぶことといえる．事後確率は，事前確率と尤度の積に比例する．この事前確率と尤度は，ユーザーが好みに応じて仮定するため，この部分にユーザーの「味」が出る．

3　モデルとアルゴリズム

モデルが表すものが「考え方」「達成したいこと」とすると，アルゴリズムとはモデルの実行手段に相当する．この 2 つの概念は明らかに異なるものではあるのだが，現実には混同して使われる場合が極めて多い．そこで例を挙げてゆっくり説明しよう．

ドラえもんの第 1 話に，セワシがのび太にこんな話をする場面がある．

> セワシ「たとえば，きみが大阪へ行くとする．いろんな乗りものや道すじがある．だけど，どれを選んでも，方角さえ正しければ大阪へつけるんだ」．

これをモデル (M) とアルゴリズム (A) の 2 段構えの枠組みにあてはめると，

M　東京から大阪へ行く

A　{ 飛行機，自動車，新幹線，船 }

ということになる．モデルは一つだが，それを実行するための手段は 4 つあるということだ．

最小二乗法の例で考えてみよう．この場合は，

M　$J(a,b)$ を最小化する a,b を求める

A　$\partial J/\partial a = 0, \partial J/\partial b = 0$ を手で解き，数値を代入

というように切り分けられる．そして，アルゴリズムは他にも考えられて，例えば

A′　$\partial J/\partial a, \partial J/\partial b$ を数値的に計算して最急降下法

という方法でもモデルの内容を達成できよう．飛行機でも新幹線でも大阪には着けるということである．

複数のアルゴリズムがある場合には，アルゴリズム同士の比較をすることが多い．その際，大阪に行くための飛行機と新幹線の所要時間や運賃を比較することは意味がある．ところが，飛行機での到着地・大阪と，都営バスでの到着地・目黒を比較して，飛行機が都営バスよりもすぐれていると結論づけるのはおかしい．到着地で優劣を競うことは，大阪に行くというモデルと目黒に行くというモデルの別々のモデルのどちらがすぐれているかを判定することに等しい．飛行機も都営バスもそれぞれのモデル内容を実行すべく働いたのであって，優劣をつけられるべきはモデルの側である．

まとめると，モデルによって使用できるアルゴリズムの種類が変わってくる．モデルが同じならば，どのアルゴリズムを用いても同じ結果が得られる．また，モデルが違えば得られる結果は違う．

4　データ同化におけるモデルとアルゴリズム

4.1　正統派モデル

データ同化の分野で考えるモデルは，次の 2 つである．

M1　$p(x_1,\cdots,x_T|y_{1:T})$ を最大とする $x_1,\cdots,x_T$ を求める

M2　$p(x_t|y_{1:T})$ を最大とする x_t を求める $(t=1,\cdots,T)$

M1 では $x_1,\cdots,x_T$ の同時確率を，M2 では T 個の周辺確率を最大にするようにモデルのパラメータを推定する．得られるのは，いずれのモデルでも $x_1,\cdots,x_T$ の推定値である．だが，M1 と M2 は異なるモデルなので推定値は一般には異なったものとなる．ただ，推定値の「でき」としては，どちらのモデルから導かれたものも同等のものであると考えてよい[1]．仮に M1 を東京から大阪へ行くということで「大阪モデル」としたならば，M2 はさしづめ「新大阪モデル」くらいのイメージである．

[1] 厳密には，シミュレーションモデルと観測演算子がどちらも線型であり，事前分布と尤度がどちらも正

さて，モデルが与えられれば，それに応じて実行手段としてのアルゴリズムを用意しなければならない．M1 に対しては M1 向けのアルゴリズム A1，M2 に対してはアルゴリズム A2 を用意する．過去の研究において，

A1　[最適化型] {4 次元変分法，リプレゼンター法 }

A2　[分布推定型] { カルマンフィルター，またその発展型 }

が挙げられている．A1 の最適化型というのは，事後確率 $p(x_1, \cdots, x_T|y_{1:T})$ を最大とする $x_1, \cdots, x_T$ を求めることだけを行うアルゴリズムであることを意味している．一方で A2 は，直接的には $p(x_t|y_{1:T})$ の最適化ではなく，$p(x_t|y_{1:T})$ の確率分布そのものを推定するアルゴリズムである．分布自体がわかれば，当然最大値を与える x_t もわかるはずである．

A1 に属するアルゴリズムとしては，次の 2 つがある．一つは，微分係数 $\nabla p(x_1, \cdots, x_T|y_{1:T})$ を数値的に求め，降下法を行うというもので，これは **4** 次元変分法と呼ばれている．もう一つのアルゴリズムはリプレゼンター法と呼ばれ，数値的に微分をとるのではなくて，$\nabla p(x_1, \cdots, x_T|y_{1:T}) = 0$ という方程式をニュートン法的に反復代入により解くものである．

A2 に属するアルゴリズムでは，予測分布 $p(x_t|y_{1:t-1})$, フィルター分布 $p(x_t|y_{1:t})$，平滑化分布 $p(x_t|y_{1:T})$ についての漸化式を利用する．この漸化式にもとづき，カルマンフィルター・平滑化，拡張カルマンフィルター，アンサンブルカルマンフィルター，粒子フィルターなどのアルゴリズムが提案されている．

4.2　妥協モデル

アルゴリズムの実行には，現実的な計算負荷のやりくりが伴う．上で述べたモデル M1 もしくは M2 を採用する場合，アルゴリズム A1, A2 の計算負荷が大きすぎることが起こり得る．しかし，x_t の推定値は求めなければならないとき，モデルの妥協をして対処する．これまでに使われている妥協モデルは，

M3　$p(x_t|y_t)$ を最大とする x_t を求める $(t = 1, \cdots, T)$

である．このモデルは M2 とよく似ているが，x_t の推定に全データ $y_{1:T} = \{y_1, \cdots, y_T\}$ を使うのではなく，同時刻のデータ y_t のみを使う点が M3 の独自の点である．M3 でも同じく $x_1, \cdots, x_T$ の推定値を求めることはできるが，前の 2 つのモデルと比べると，推定値の品質は明らかに劣る．M1, M2 を「大阪モデル」「新大阪モデル」に例えたことに倣うと，M3 は「都内でやりくりモデル」といえよう．ゴールの設定を近くにしたのだから安価で達成できるが，満足度は低くなる．例えば美味しいお好み焼きを食べたいとする．そこでぼてぢゅう「梅田店」「新大阪駅店」に出向けば満足できるが，そこまで行くには交通費が馬鹿にならない．でもお好み焼きはぜひとも食べたい．そこで「目黒店」で済ませるという案が浮上するが，味は当然それなりである．

モデル M3 についても「最適化型」「分布推定型」の 2 流派のアルゴリズムが提唱されている．前者のことを **3** 次元変分法，後者のことを最適内挿法と呼んでいる．

A3　{[最適化型] 3 次元変分法; [分布推定型] 最適内挿法 }

規分布である場合に限って両推定値が一致する．この場合には，以下のアルゴリズム A1, A2 のどちらも共通に利用できる状況になる．

ただし，最適内挿法で扱えるモデルは観測演算子が線型，尤度が正規分布の場合に限るため，3 次元変分法のほうがより広いモデルを扱える．

5　まとめ

データ同化とは，データを取り込んだシミュレーションである．初期条件・境界条件，入力パラメータ（拡散係数など）のチューニングが自動でできる．シミュレーションコードがあれば，事前分布・尤度を仮定することによりモデルが完成する．モデルに応じて，アルゴリズムはいろいろある．予算が限られているときは，まずアルゴリズムを検討し，それでもダメならばモデルを妥協する．

データ同化の実行に際しては，シミュレーションの 100 倍を超えるメモリが必要となる．そのため，データ同化の実行には，低分解能のシミュレーションモデルをベースに同化モデルを構成していくことになる．従来のシミュレーション研究がより高分解能のモデルを追求していることを考えると，この「低分解モデル＋データ」のデータ同化手法による研究は，今後の数値モデルによる研究分野の新たな方向性を示すものと考えられる．