数理解析研究所講究録
第2057巻 2017年 59-66

データ同化による不確かさを持つ現象の理解と予測ならびにモデリングへの展開

明治大学・総合数理学部　中村　和幸
Kazuyuki Nakamura
School of Interdisciplinary Mathematical Sciences,
Meiji University

1　序論

データ同化 [1, 2] とは，気象学・海洋学の分野で発展してきた手法であり，計算機シミュレーションモデルと実観測データを組み合わせ，よりよいシミュレーション予測や未知物理観測量の推定を志向した手法である．近年では，感染症 [3]，生命科学 [4]，地盤工学 [5]，分子動力学 [6] といった，従来適用されてきた分野以外の分野にも応用範囲が広がっている．応用範囲の広がりの結果，予測精度向上のための観測データを適切に利用した初期値・状態推定という範囲も超えてきている．一方で，統計数理的な観点からの整理もされるようになってきている．このような統計数理的な整理は，新たな推定手法の導入につながるという観点でも重要である．

そこで本稿では，データ同化の統計数理としての側面について，これまでの知見を整理する．この整理を通じて，データ同化における不確かさの取り扱いについて例も交えて説明し，データ同化を通じた不確かな現象の理解と予測のためには，どのようなモデリングが求められるかについて議論する．

2　データ同化と状態空間モデル

2.1　実システムのモデル化と状態空間モデル

離散時間 $t(1 \leq t \leq T)$ の各時刻において，系の状態を表す変数群 $\tilde{\boldsymbol{x}}_t$ を考える．また，$\boldsymbol{x}_0$ は系の初期状態である．この系は，いわゆる力学系を想定しており，時刻 t における状態は直前の時刻 $t-1$ における状態 $\tilde{\boldsymbol{x}}_{t-1}$ に依存して決定的に決まるとする．すると，

$$\tilde{\boldsymbol{x}}_t = \tilde{\boldsymbol{f}}_t(\tilde{\boldsymbol{x}}_{t-1}) \tag{1}$$

と書かれることになる．ここで $\tilde{\boldsymbol{f}}_t$ は，系の状態更新を表現する関数である．

現象を理解する上での上記の定式化の問題点は，実際に解かれるべき多くの問題において，系の時間発展を与える式である $\tilde{\boldsymbol{f}}_t$ が完全な形ではわかっていないこと，また，初

期状態 $\tilde{x}_0$ も正確な値としては得られない場合がほとんどであることにある．予測精度を下げるこれらの要因は，Lorenz によって示された決定論的カオス・初期値鋭敏性も含め，実問題での予測や発見において，適切なモデル化か推定が必要となり，上記の定式化だけでは不十分なことがわかる．

そこで，基本的に上記の状態変数と同等なものであるが，その不確実性を内包する変数として，確率変数として取り扱うことを考える．確率変数としての面を強調し，式 (1) の $\tilde{x}_t$ と区別するために，確率変数として取り扱うすなわち，実現象における状態変数そのものとは区別して，状態変数を x_t と表記しなおす．これを状態ベクトルと呼ぶ．すると，状態ベクトル x_t は，直前時点での状態に依存して確率的に決まるとできる．以上を踏まえると，系の発展については

$$x_t = f_t(x_{t-1}, v_t) \tag{2}$$

と定式化できることになる．ただし，v_t は状態変数の時間発展を定式化した際に，時間発展において含まれる確率的な項であり，システムノイズと呼ぶ．実問題においては，この項はモデル化誤差や数値解析上の誤差など，さまざまな誤差が重畳されて含まれていると考えられる．関数 f_t をうまくとることで，一般性を失わずに v_t を正規分布 $N(\mathbf{0}, \Sigma_{v,t})$ と仮定できる．

一方で，実観測データを得るプロセスでは，系の状態そのものの値を得られず，何らかの計測誤差が含まれることになる．そこで，時刻 t において得られる観測を，状態ベクトルとは区別して y_t と表記することとし，観測ベクトルと呼ぶ．すると，観測のプロセスについては

$$y_t = h_t(x_t, w_t) \tag{3}$$

と定式化できる．ここで w_t は観測時の誤差を表し，観測ノイズと呼ぶ．観測ノイズ w_t も，一般性を失わずに正規分布 $N(\mathbf{0}, \Sigma_{w,t})$ に従うと仮定できる．これらの式 (2), (3) と初期状態ベクトル x_0 についての分布を与えたものを総称して，非線形・非ガウス状態空間モデルと呼ぶ．

なお，より一般的には

$$x_t \sim Q_t(\cdot|x_{t-1}, \theta_{\text{sys}}), \tag{4}$$
$$y_t \sim R_t(\cdot|x_t.\theta_{\text{obs}}), \tag{5}$$
$$x_0 \sim Q_0(\cdot) \qquad (t = 1, \ldots, T)$$

に従うと考えることができる．これは，一般化状態空間モデルと呼ばれる．非線形・非ガウス状態空間モデルの f_t,h_t をうまくとると，一般化状態空間モデルと等価なモデルにすることができるが，陽に表現するのが困難な場合があるため，両者の区別がある．一般化状態空間モデルは，音声認識などで用いられる隠れマルコフモデルも含んでおり，この後説明するフィルタリングも適用可能である．

非線形・非ガウス状態空間モデルや一般化状態空間モデルは，次のようなマルコフ性を持っていることが本質的である：

$$p(x_t|x_{1:t-1}, y_{1:t-1}) = p(x_t|x_{t-1}). \tag{6}$$
$$p(y_t|x_{1:t}, y_{1:t-1}) = p(y_t|x_t). \tag{7}$$

ただし，

$$\boldsymbol{y}_{1:j} \equiv \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_j\}$$

という表記を用いており，$\boldsymbol{x}_t$ についても同様である．以下では，式 (6), (7) を満たす構造を持つ時系列モデルを総称して，状態空間モデルと呼ぶこととする．

2.2 状態空間モデルと逐次ベイズフィルタ

状態空間モデルにおけるある時点 t の状態ベクトルについて，ある一定区間 $1 \leq t \leq j$ の間の観測ベクトル $\boldsymbol{y}_t$ を得た下での推定について考える．これは，$p(\boldsymbol{x}_t|\boldsymbol{y}_{1:j})$ という条件付き周辺分布の推定を得ることに対応するが，さらに，t と j の関係により，次の3種類に分類される：

- $t > j$:予測分布
- $t = j$:フィルタ分布
- $t < j$:平滑化分布

特に，$t = j+1$ のときを一期先予測分布と呼ぶ．これらの分布を与えるのに，式 (6) ならびに式 (7) を用いて導出される次の関係式を用いることができる：

$$p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1}) = \int p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) p(\boldsymbol{x}_{t-1}|\boldsymbol{y}_{1:t-1}) \mathrm{d}\boldsymbol{x}_{t-1}, \tag{8}$$

$$p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t}) = \frac{p(\boldsymbol{y}_t|\boldsymbol{x}_t) p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1})}{\int p(\boldsymbol{y}_t|\boldsymbol{x}_t) p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1}) \mathrm{d}\boldsymbol{x}_t}, \tag{9}$$

$$p(\boldsymbol{x}_t|\boldsymbol{y}_{1:j}) = p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t}) \int \frac{p(\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:j}) p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)}{p(\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:t})} \mathrm{d}\boldsymbol{x}_{t+1}. \tag{10}$$

ここで，式 (8) の $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ は状態空間モデルのシステムモデルから，(9) の $p(\boldsymbol{y}_t|\boldsymbol{x}_t)$ は観測モデルからそれぞれ導出ができる．一期先予測の式 (8) にはフィルタ分布が，フィルタの式 (9) には一期先予測分布が含まれることから，これらを交互に実行することで，フィルタ分布を次々と得ることができる．この手続きを逐次ベイズフィルタと呼ぶ．また，ある時点までのフィルタ分布とそれまでの一期先予測分布が与えられると，式 (10) を用いることで，それより過去の平滑化分布を求めることができる．これを固定区間平滑化と呼ぶ．逐次的に得たデータをシミュレーションモデルに反映する逐次データ同化は，数理的には上記の一期先予測とフィルタリングの繰り返しを行っていることに対応する．これらの分布は，拡張カルマンフィルタ・アンサンブルカルマンフィルタ・粒子フィルタ等によって計算される．

2.3 各種のフィルタリングアルゴリズム

2.3.1 カルマンフィルタ

線形・ガウス状態空間モデル

$$\boldsymbol{x}_t = F_t \boldsymbol{x}_{t-1} + \boldsymbol{v}_t,$$
$$\boldsymbol{u}_t = H_t \boldsymbol{x}_t + \boldsymbol{w}_t$$

は，システムモデルと観測モデルがともに線形で，全てのノイズがガウス分布に従うものである．この仮定より，一期先予測分布ならびにフィルタ分布は，その条件付き平均ならびに分散共分散行列を求めればよい．データ同化で直接用いられることは少ないが，非線形システムを線形化することによって得られたシステムに，カルマンフィルタが適用されることがあり，このときのアルゴリズムを拡張カルマンフィルタと呼ばれる．

カルマンフィルタは，以下の手続きによって計算される．

[一期先予測]

$$\boldsymbol{x}_{t|t-1} = F_t \boldsymbol{x}_{t-1|t-1},$$
$$V_{t|t-1} = F_t V_{t|t-1} F_t^T + G_t Q_t G_t^T.$$

[フィルタリング]

$$K_t = V_{t|t-1} H_t^T (H_t V_{t|t-1} H_t^T + R_t)^{-1},$$
$$\boldsymbol{x}_{t|t} = \boldsymbol{x}_{t|t-1} + K_t(\boldsymbol{y}_t - H_t x_{t|t-1}),$$
$$V_{t|t} = (I - K_t H_t) V_{t|t-1}.$$

ただし，$\boldsymbol{x}_{\cdot|\cdot}$, $V_{\cdot|\cdot}$ は，それぞれ，条件付き平均 $\boldsymbol{x}_{j|k} = E[\boldsymbol{x}_j|\boldsymbol{y}_{1:k}]$ ならびに分散共分散行列 $V_{j|k} = E[(\boldsymbol{x}_j - \boldsymbol{x}_{j|k})(\boldsymbol{x}_j - \boldsymbol{x}_{j|k})^T]$ である．

2.3.2 拡張カルマンフィルタ

拡張カルマンフィルタ (EKF, Extended Kalman filter) は，非線形状態空間モデルに従う場合の推定アルゴリズムである [7]．EKF では，その時点での平均推定値周りでモデルを線形化して，線形の状態空間モデルを構成した上で，分散共分散行列の更新にカルマンフィルタの更新式を適用する．すなわち，

$$\hat{F}_t = \left.\frac{\partial \boldsymbol{f}_t}{\partial \boldsymbol{x}_{t-1}}\right|_{(\boldsymbol{x}_{t-1|t-1}, \boldsymbol{0})},$$
$$\hat{G}_t = \left.\frac{\partial \boldsymbol{f}_t}{\partial \boldsymbol{v}_t}\right|_{(\boldsymbol{x}_{t-1|t-1}, \boldsymbol{0})},$$
$$\hat{H}_t = \left.\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{x}_t}\right|_{(\boldsymbol{x}_{t|t-1}, \boldsymbol{0})}$$

とし，

$$\boldsymbol{x}_{t|t-1} = \boldsymbol{f}_t(\boldsymbol{x}_{t-1|t-1}, \boldsymbol{0})$$

の他はカルマンフィルタの更新式によって更新する．

2.3.3 アンサンブルカルマンフィルタ

アンサンブルカルマンフィルタならびに粒子フィルタは，確率密度関数 $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:j})$ をディラックのデルタ関数を用いた実現値集合（アンサンブル）による近似

$$p(\boldsymbol{x}_t|\boldsymbol{y}_{1:j}) \approx \frac{1}{N}\sum_{i=1}^{N}\delta(\boldsymbol{x}-\boldsymbol{x}_{t|j}^{(i)}) \tag{11}$$

としたものである．ここで，$\delta(\cdot)$ はディラックのデルタ関数，$\{x_{t|j}^{(i)}\}_{i=1}^{N}$ は分布 $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:j})$ に従う実現値の集合で，これによって密度関数の近似を行っている．ここでのディラックのデルタ関数は，多次元空間上に定義されるもので，全ての要素が 0 となる 1 点にのみ point mass が存在するものである．なお，密度関数が存在しない離散確率変数の場合にも粒子フィルタは有効である．これは，累積分布関数でいえばステップ関数近似を行っていることに相当するためである．

アンサンブルカルマンフィルタは，一期先予測にはモンテカルロシミュレーションを用い，フィルタリングではカルマンフィルタの更新式を用いる方法である．フィルタリングにおいては，状態ベクトルの線形和が出てくることから，状態ベクトルが本来持っている物理学的な連続性を破壊する可能性がある．しかし，弱非線形の系の場合には，データ同化によって得られる利点の方が大きいため，アンサンブルカルマンフィルタによる推定がうまくいく傾向にある．アンサンブルカルマンフィルタのアルゴリズムは次の通りである．

1. 初期集合 $\{\boldsymbol{x}_{0|0}^{(i)}\}_{i=1}^{N}$ を初期分布 $p(\boldsymbol{x})$ から抽出して作成する．$t \leftarrow 1$ とする．
2. 一期先予測として，$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}=\boldsymbol{x}_{t-1|t-1}^{(i)})$ に従って $\boldsymbol{x}_{t|t-1}^{(i)}$ をサンプリングにより生成．
3. フィルタリングとして，以下の (a),(b),(c) を実行する．
   (a) 標本分散共分散行列 $\hat{V}_{t|t-1}$ をアンサンブル集合から計算する．
   (b) カルマンゲイン K_t をカルマンフィルタの式により計算する．
   (c) $\boldsymbol{x}_{t|t}^{(i)}=\boldsymbol{x}_{t|t-1}^{(i)}+K_t(\boldsymbol{y}_t-H_t\boldsymbol{x}_{t|t-1}^{(i)}+\boldsymbol{w}_t^{(i)})$ を各 i に適用し，フィルタ分布のアンサンブルを得る．
4. $t=T$（終点）なら停止．それ以外は，$t \leftarrow t+1$ として 2) に戻る．

2.3.4 粒子フィルタ

粒子フィルタは，一期先予測にモンテカルロシミュレーションを，フィルタリングにおいては，尤度関数によるアンサンブルメンバーの重みづけによって推論する手法である．フィルタリングにおいて，さらにリサンプリングすることにより粒子間の重みの違いを解消するのが，SIR(Sequential Importance Resampling) 法であり，リサンプリングせずに単に重みづけを行う手法が SIS(Sequential Importance Sampling) 法である．SIR 法の場合のアルゴリズムは次の通りである．

1. 初期集合 $\{\boldsymbol{x}_{0|0}^{(i)}\}_{i=1}^{N}$ を初期分布 $p(\boldsymbol{x})$ から抽出して作成する．$t \leftarrow 1$ とする．

2. 一期先予測として，$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1} = \boldsymbol{x}_{t-1|t-1}^{(i)})$ に従って $\boldsymbol{x}_{t|t-1}^{(i)}$ をサンプリングにより生成．

3. フィルタリングとして，以下の (a),(b),(c) を実行する．

   (a) 各粒子の尤度 $l_t^{(i)} = p(\boldsymbol{y}_t|\boldsymbol{x}_t = \boldsymbol{x}_{t|t-1}^{(i)})$ を計算する．

   (b) $w_t^{(i)} = l_t^{(i)} / \sum_{j=1}^{N} l_t^{(j)}$ と規格化して，重みの集合 $\{w_t^{(i)}\}$ を得る．

   (c) 各粒子 $\boldsymbol{x}_{t|t-1}^{(i)}$ の確率を $w_t^{(i)}$ として，N 回復元抽出し，その集合を $\{\boldsymbol{x}_{t|t}^{(i)}\}$ とする．

4. $t = T$（終点）なら停止．それ以外は，$t \leftarrow t+1$ として 2) に戻る．

SIS 法では，復元抽出の代わりに，重みの集合もあわせて保持して，次の時点の重みとの積を取るということを繰り返すことになる．

3 モデル構築の方法としてのデータ同化

3.1 システムモデルと状態ベクトルの構成について

データ同化の枠組みでは，シミュレーションモデルに含まれる変数が状態ベクトルの主な要素となる．通常，シミュレーションモデルは時間発展について離散化後ものを想定することから，有限で離散時間のシステムとしてよい．このことから，最も単純な状態ベクトルの構成として，t 時点の状態ベクトルを，それまでの全てのシミュレーション変数を含めるということが可能である．この方法は，フィルタリングを行うだけで，固定区間平滑化を得ることが出来る [7]．一方で，徐々に状態ベクトルの次元が増えてしまうことから，実際の問題では現実的でない．通常は，マルコフ性の仮定が満たされる範囲となるように，適当な範囲の過去データを含めるようにして構成する．例えば，Leapfrog 法により離散化された場合には，システムの持つ状態の次元以上のデータを持っておく必要がある．

一方システムノイズによって定式化される誤差として，離散化誤差，モデル化誤差の両方が考えられる．これらは，実際のシステムを完全に完成できない，あるいは模倣が不可能であることに対応する項であり，状態に影響を与えているとみなすことができるためである．しかし，シミュレーションに対して，このシステムノイズが小さくとも，適切な結果を得られない場合がある．例えば，地盤工学分野において適用した例 [5] では，システムノイズによる状態ベクトルの擾乱，アンサンブルカルマンフィルタを回避し，SIS 法を用いている．

3.2 逐次データ同化アルゴリズムの選択

一般には，非線形モデル時の保証のある粒子フィルタが有利なようであるが，必ずしもそうでないことがわかっている．粒子フィルタには，標本点の尤度が極端に小さくなって

しまうことにより，分布を表現するには十分でない点の集合になってしまうような退化の問題が存在する．そのため，アンサンブル数が少ない時のモンテカルロ誤差が大きくなる傾向がある．これに対して，非線形システムであってもそれが強くない場合には，状態の加法・平均演算に妥当性があり，その結果として，アンサンブルカルマンフィルタによって得られる推定は安定的になる傾向がある．さらに，高次元であっても問題なく適用できる．

以上のことから，大まかに高次元で弱非線形である場合，例えば気象の問題などでは，アンサンブルカルマンフィルタが，これから比較的次元が低いが非線形性が高い場合には粒子フィルタを使うと良いということになる．一方で，非線形性が強いシステム，例えば生命科学の多くのモデルの場合には，分布の非対称性などが見受けられ，また次元も高くないため，粒子フィルタが適当と考えられる．実際に，Nakamura *et al.*[4] では，アンサンブルメンバー数を多くした SIR 法により，適切なパラメータの分布推定を与えることに成功している．また，特に非線形性が強く，状態ベクトルへの擾乱に問題がある地盤変形システムのようなシステムの場合には，システムノイズを少なくして SIS 法を用いると良いと判断できる．

3.3 モデル構築における適用に向けて

データ同化をモデル構築に適用する一つの利点として，実データと比較した尤度ならびに予測精度を導出できる点が挙げられる．また，アンサンブルカルマンフィルタや粒子フィルタを用いることで，フィルタ分布や予測分布をアンサンブル集合として得ることができることも挙げられる．これらにより，予測分布の形状や予測精度の観点からモデルを選ぶことが可能になるためである．Ohya *et al.*[8] では，逐次データ同化の枠組みではないが，予測分布の構成に関しては，シミュレーションモデルの結果と組み合わせたジャックナイフ法を用いた枠組みによって，予測誤差を推定し，津波の遡上問題におけるパラメータ推定を与えている．この問題では，予測分布の構成がパラメータの空間分布を与える一つの指針となっているが，データ同化を用いて予測分布のアンサンブルを構成することができるので，これをもとに複数モデルの選択の指針として使えると考えられる．

また，別の利点としては，システムノイズと観測ノイズという形でノイズのモデリングを明示的に行っていることが挙げられる．特に，システムノイズに含まれるべきノイズとして様々な誤差があるが，応用研究においては，これらの間でバランスを取る必要がある．数理的には峻別すべきこれらのノイズについて，統一的に扱うことができることが一つの利点となると考えている．実際に，粒子法流体解析においてはさまざまな誤差の表出が実用上の問題となっており，数値解析上の誤差とモデル化誤差を峻別しつつも比較する手法を，データ同化の枠組を基礎として開発している．

4 結論

本稿では，データ同化の枠組みとアルゴリズムについて，統計数理的観点から整理した．さらに，モデル構築の方法としてのデータ同化について，従来の研究例も交えなが

ら，手法の選択ならびにモデル選択の可能性について議論した．

データ同化は，近年では工学分野も含めて実システムへの適用が広がっている分野である一方，数理的な整理も広がりつつあり，特に海外において書籍も含めたリソースが増えてきている分野である．一方で工学応用の観点では，最適設計やシステム同定といった類似手法もある．今後は，これらの知見も含めながら，数理科学的・統計科学的整理を進めていくことが課題となる．

5 謝辞

本研究の成果の一部は，JSPS 科研費 JP15H05303, JP26289162, JP26280006 の助成を受けたものである．

参考文献

[1] 中村和幸，上野玄太，樋口知之，データ同化：その概念と計算アルゴリズム，統計数理，2005.

[2] K. Nakamura, T. Higuchi, and N. Hirose, “Sequential Data Assimilation: Information Fusion of a Numerical Simulation and Large Scale Observation Data,” J.UCS, Vol. 12, pp. 608-626, 2006.

[3] Masaya M. Saito, Seiya Imoto, Rui Yamaguchi, Masahiro Kami, Haruka Nakada, Hiroki Sato, Satoru Miyano, Tomoyuki Higuchi, Extension and verification of the SEIR model on the 2009 influenza A (H1N1) pandemic in Japan, Mathematical Biosciences, 246 (1), pp. 47-54, 2013.

[4] K. Nakamura, R. Yoshida, M. Nagasaki, S. Miyano, and T. Higuchi, ”Parameter estimation of in silico biological pathways with particle filtering towards a petascale computing,” The Proceedings of 14th Pacific Symposium on Biocomputing, pp. 227-238, 2009.

[5] A. Murakami, T. Shuku, S. Nishimura, K. Fujisawa, and K. Nakamura, ”Data assimilation using the particle filter for identifying the elasto-plastic material properties of geomaterials,” International Journal for Numerical and Analytical Methods in Geomechanics, Vol.37, No.11, Doi: 10.1002/nag.2125, 2013.

[6] Y. Matsunaga, A. Kidera, and Y. Sugita, ”Sequential data assimilation for single-molecule FRET photon-counting data,” Journal of Chemical Physics 142, 2015.

[7] 樋口知之，上野玄太，中野慎也，中村和幸，吉田亮，データ同化入門，朝倉書店，2011.

[8] Y. Ohya and K. Nakamura, ”A New Setting Method of Friction Parameter for Real-Time Tsunami Run-Up Simulations Based on Inundation Observation,” Theoretical and Applied Mechanics Japan, Vol. 62, pp. 167-178, 2014.