

データ同化：その概念と計算アルゴリズム

中村 和幸^{1,3}・上野 玄太^{2,3}・樋口 知之^{2,3}

(受付 2005 年 3 月 31 日；改訂 2005 年 8 月 31 日)

要 旨

気象学・海洋学において活用されているデータ同化について、その概念と状態空間モデルによる表現について紹介する。また、データ同化の中でよく用いられている手法であるアンサンブルカルマンフィルタについて、粒子フィルタとの関連に注目して議論する。小規模なシステムに対する数値実験により、非線形観測をもつ問題に対して粒子フィルタを適用した場合に、アンサンブルカルマンフィルタに対して優位性を持つことを示し、さらに実際の非線形観測をもつ問題に対する適用可能性について議論する。

キーワード：データ同化，粒子フィルタ，アンサンブルカルマンフィルタ，状態空間モデル。

1. 序論

データ同化(Wunsch, 1996; Daley, 1991; Bennett, 2002; 多田・村上, 1997)は気象学や海洋学の分野において用いられている手法である。これらの分野においては、物理法則などの法則に基づいて時空間シミュレーションモデルを構成し、これの計算を行なうことで、実際の地球システムなどの再現、解明、予測を行なうことになる。その際に、システムの持つ非線形性が原因となり、初期条件や境界条件さらにはモデルに含まれるパラメータの与え方によって、シミュレーション結果が現実の現象から乖離した、著しく不適切なものとなることが問題となる。データ同化の目的は、これらシミュレーションにおける初期条件、境界条件ならびにシミュレーションモデルに含まれるパラメータを、実際の観測に基づいて適切なものに構成することである。データ同化には逐次型のものとは非逐次型のものがある。逐次型データ同化は、実観測を得るたびにシステムに含まれる状態の変数を修正し、適切なものに収束させていくものであり、本論文ではこの逐次型を議論することにする。なお、逐次型でないものには4次元変分法(four-dimensional variational (4DVAR))(Talagrand and Courtier, 1987; Courtier et al., 1994)などがある。

逐次型データ同化の分野において広く用いられている手法に、アンサンブルカルマンフィルタ(Ensemble Kalman Filter (以下 EnKF))(Evensen, 1994, 2003)がある。これは、シミュレーションモデル内の状態を表す確率変数について、その分布を実現値集合であるアンサンブルによって保持し、観測を得る毎に、観測モデルをもとにしたカルマンフィルタ(Kalman, 1960)による推定により、2次モーメントまでが一致するように、アンサンブルを修正することを繰り返

¹ 総合研究大学院大学 複合科学研究科統計科学専攻：〒106-8569 東京都港区南麻布 4-6-7

² 統計数理研究所：〒106-8569 東京都港区南麻布 4-6-7

³ 科学技術振興機構(JST), 戦略的基礎研究推進事業(CREST)

す手法である。

一方状態空間モデルの視点から見ると、データ同化はシステムモデルとしてシミュレーションモデルが用いられているもとでの、状態推定ならびにパラメータ推定の問題とみなすことができる(Higuchi, 2003)。加えて、アルゴリズム的な観点から見れば、EnKF は粒子フィルタ(Particle Filter(以下 PF))(Kitagawa, 1996; Gordon et al., 1993; Liu, 2001; Doucet et al., 2001; 樋口, 1996; 北川, 1996)を用いた推定手法と類似している。後の章で見るように、EnKF と PF のアルゴリズム上の違いは、フィルタリングを行なう際にカルマンゲインをもとにしたゲインを用いるか、重み関数によるリサンプリングによるかの違いのみである。本論文ではこの類似点と相違点に注目しながら、関連づけを行なう。

論文の構成は次の通りである。第 2 章では、データ同化の背景について述べる。第 3 章では、データ同化の状態空間モデルでの定式化を行なう。第 4 章では、逐次データ同化手法でよく用いられている EnKF のアルゴリズムについて説明する。第 5 章では、EnKF の適用範囲について触れ、観測モデルが非線形である場合の拡張について述べる。第 6 章では、PF と EnKF の相違点を比較・検討し、その内容を踏まえた上で、システムモデル、観測モデルが共に非線形である時の、PF と EnKF に関する数値実験の結果を示す。第 7 章で結論を述べる。

2. データ同化の背景

この章では、データ同化分野の背景と適用手法、研究の変遷について触れる。

そもそも地球システムのような複雑で規模の大きいシステムを理解するためには、その内部でのフィードバック的な性質と非線形的な過程を再現する必要がある。例えば気象現象においては、水蒸気量や風向・風速、そして地形などのさまざまな要因が互いに影響しあい、しかもその関係は線形の関係とは限らない。このような現象を再現するには、計算機シミュレーションの方法に依る必要がある。しかしながら、初期値の微細な誤差が予測の結果を大幅に変えてしまうことを意味する「バタフライ効果(Lorenz, 1963)に代表される、非線形性が持つ予測不可能性によって、実際のシステムを再現したり、その中に含まれる諸変数を推定する、あるいは予測のための適切な初期値を設定するのは容易ではない。

一方、人工衛星などの観測システムの発達により、シミュレーションモデルよりも解像度が粗いながらも、数多くの観測データが得られるようになってきた。そこで、この観測データとシミュレーションを統合し、より精度の高いシステムの再現、適切な諸変数の推定や初期値の構成を行なうことを目指すようになったのが、データ同化である。予測においては、このようにして得られた諸変数を初期値として用いることにより、予測の精度が向上することが見込まれ、また推定においては、実際のシステムの理解に役立つことになる。

データ同化のうち、逐次データ同化では観測が得られるたびに、その時刻での状態変数を推定する。これには、EnKF が用いられるまではカルマンフィルタや拡張カルマンフィルタ(片山, 1983; 有本, 1977)が用いられてきていた。しかし、拡張カルマンフィルタにおける問題として、システムモデルの分散共分散行列が安定しない点、およびオンラインでシステムモデルのヤコビ行列を計算しなければならない点の 2 点があった。EnKF はこの両者の欠点を克服するために Evensen(1994)によって導入されたものである。

実際に EnKF は、推定されるべき状態変数ベクトルの次元が大きい実問題に対して適用されてきた。一例を挙げると、Lisæter et al.(2003)による北極域における氷の密接度のデータ同化に使われ、そのシステムの状態変数ベクトルの大きさは 2×10^6 であった。ただし、適用する際には計算の制約上から、格子点上毎にデータ同化を行なうというモデルの近似を行なうことが多い。この Lisæter et al. によるモデルでも、観測モデルについては格子点ごとに部分観測

モデルを設定し、それに対してフィルタリングを行なうという近似をしており、格子について 1 点、113 変数に対してのアンサンブルカルマンフィルタを行ない、その並列という形で実装されている。つまり、実際は 113 次元の問題の適用にとどまっている。その一方で、理論的な部分についても研究され、EnKF の変種である EAKF (Ensemble Adjustment Kalman Filter) (Anderson, 2001) などが研究されている。また、EnKF における固定ラグ平滑化であるアンサンブルカルマンスムーザ (EnKS) も Evensen and van Leeuwen (2000) により導入されている。

他方、EnKF と類似の関係にある PF あるいはその変種の適用については、Pham (2001) による 3 次元 Lorenz モデル、van Leeuwen (2003) による南アフリカ海域のモデル、Manda et al. (2003) による混合層モデルに対する適用例以外はない。いずれの同化実験においても、線形観測のシステムを想定して同化が行なわれている。

3. 状態空間モデルにおける定式化

3.1 データ同化の設定

本節では、データ同化の設定について述べる。データ同化において登場するものは、実システム、シミュレーションモデル、観測モデルの 3 つである。実システムはここでは、大気や海洋などの実際の解析対象である。実システムの構造を明らかにするために、実システムはまず偏微分方程式の形で表現され、さらにコンピュータで計算するために空間は格子に切れ、時間についても離散化される。偏微分方程式によって近似表現された段階で、すでに実システムのうちモデル化されない挙動が残る。また、離散化されたシミュレーションモデルにおいては、実システムでは連続な空間上に存在する風速などの変数が、各格子点上においてのみ定義される事になる。

これらシミュレーションモデルに変形する際の誤差により、シミュレーションモデルは本質的に実システムに対して誤差を持つ。また、それとは別に、もともと実システムとして注目している部分の範囲外からの影響による擾乱成分、確率的な挙動も存在する。

一方、観測データに注目すると、次のような事が言える。一般に、得られる観測データの空間の解像度はシミュレーションモデルでの格子点よりも粗く、サンプリング時間間隔もシミュレーションモデルにおける時間更新間隔よりも非常に長い。また、全ての状態変数が観測されることもない。すなわち、逆問題となる。その上、観測データはなんらかの非線形も含む変換を受けていることがある。

データ同化において興味があるのは、このように離散化されたシミュレーションモデルに含まれる状態変数をデータに合わせて推定すること、あるいはその中のパラメータを推定することである。このことにより、予報のための初期条件の精度あるいはパラメータの推定精度を上げ、予測精度を向上させることが可能になる。

3.2 状態空間表現

前節の事項を踏まえ、データ同化の問題を状態空間モデルを用いて位置づける。まず状態変数ベクトル x_n を定義する。状態変数ベクトルは、シミュレーションモデルの全格子点上で定義される値を全て並べたものとする。例えば、図 1 のように M 個の格子点 $\{1, 2, \dots, M\}$ 上で、温度 T_i 、水蒸気量 S_i 、風向風速を表す 2 次元ベクトル (U_i, V_i) がシミュレーションモデルで定義されていたとする。

このとき状態変数ベクトル x は、

$$x = (T_1, S_1, U_1, V_1, T_2, S_2, U_2, V_2, \dots, T_M, S_M, U_M, V_M)$$

と定義され、時刻 n での x を状態変数ベクトル x_n と定義する。一般に、 M 個の各格子点

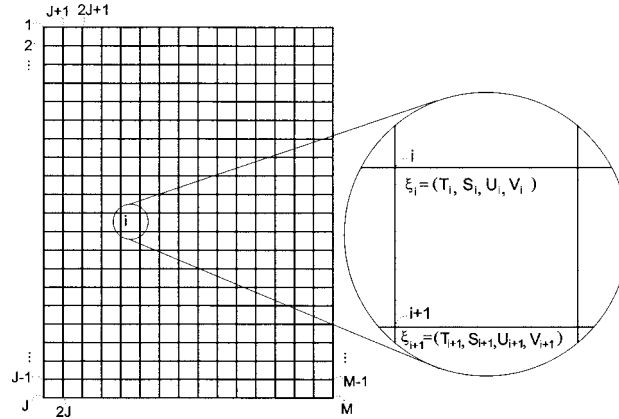


図 1. 格子点上でのシミュレーションモデルの例. 各格子点 $i (i = 1, 2, \dots, M)$ 上で温度 T_i , 水蒸気量 S_i , 風向風速 (U_i, V_i) が与えられる.

$i \in \{1, 2, \dots, M\}$ 上において, 物理量を並べたベクトル ξ_i が与えられた時, x を

$$x = (\xi_1^T, \xi_2^T, \dots, \xi_M^T)^T$$

とし, 時刻 n での x を状態変数ベクトル x_n とする. ただし T は転置を表す.

このときシミュレーションモデルは, x_{n-1} と時刻 n での擾乱 v_n を与えれば x_n が決まる形をしているので,

$$x_n = f_n(x_{n-1}, v_n)$$

という形に書くことができる. v_n は実システムにおいて本質的に存在する誤差成分と, シミュレーションモデルを設定した事により存在する擾乱を表現する役割を担っている. 便宜上, 通常 v_n はガウス分布に従うものとする.

また, 実際の観測においては, 格子点上の物理量の一部が観測される. 観測時に非線形変換などを受ける場合も考えて, 観測モデルは

$$y_n = h_n(x_n, w_n)$$

と書かれる. ここで, w_n は何らかの分布に従うものとする. 以上から, シミュレーションモデルは形式的に非線形・非ガウス型状態空間モデル(北川, 2005)で書き表す事ができる. 以下その枠組で扱うが, このような枠組の場合には

$$\begin{cases} \text{予測分布: } p(x_n | y_{1:n-1}) = p(x_n | y_1, y_2, \dots, y_{n-1}) \\ \text{フィルタ分布: } p(x_n | y_{1:n}) = p(x_n | y_1, y_2, \dots, y_n) \\ \text{平滑化分布: } p(x_n | y_{1:T}) = p(x_n | y_1, y_2, \dots, y_T) \text{ (ただし } T > n \text{)} \end{cases}$$

という確率分布が重要な役割を果たす事になり, これを求める問題となる.

なお, 実際のデータ同化においては, 観測モデルについては加法的なガウシアンノイズが加わると考え,

$$y_n = h_n(x_n) + w_n$$

とする場合が多い。さらに EnKF においては、後述するように直接的には非線形観測のシステムを扱う事が出来ないので、観測モデルについては

$$(3.1) \quad y_n = H_n x_n + w_n$$

とすることが普通である。

データ同化を実際のシミュレーションモデルに対して適用する際の、最大の特徴は、状態変数ベクトルの次元が非常に大きいものとなることである。これは、対象とする物理現象を記述するシミュレーションモデルが、格子状の各点に変数を持つ、という設定になるためである。仮に、格子点の数を 2 次元方向に 100×100 だけ取り、各格子点が 10 変数を持つとする。すると、状態変数ベクトルの次元は 10^5 となり、統計の問題において通常考慮する状態変数ベクトルの次元に比べ、非常に大きいものとなる。観測変数ベクトルについてはこれより低い次元となるが、それでもなお 10^2 から 10^4 程度の次元となる。

逐次データ同化の目的は、状態変数ベクトル x_n の時刻 n における $y_{1:n} = \{y_1, \dots, y_n\}$ での条件付き推定値 $\hat{x}_{n|n}$ 、パラメータ θ の推定値 $\hat{\theta}$ 、あるいはそれぞれの条件付き分布 $p(x_n|y_{1:n}), p(\theta|y_{1:n})$ を、各観測 y_n 毎に得る事にある。次章以降において、この $p(x_n|y_{1:n})$ のアンサンブルによる近似を逐次的に得る手法であり、逐次データ同化の分野で広く用いられている EnKF を、PF と関連づけながら確認する。

4. アンサンブルカルマンフィルタ

本章では、EnKF のアルゴリズムと成立する性質について議論する。

4.1 アンサンブルカルマンフィルタ

EnKF の手順は、通常のカルマンフィルタと同様一期先予測とフィルタリングの 2 つの手順に分れる。この際の基本となる更新則を以下に示す。4.1 節では、観測モデルは線形の(3.1)に従い、 w_n は $N(0, R_n)$ に従うものとする。

時刻 $n-1$ 時点での $p(x_{n-1}|y_{1:n-1})$ の N 個の実現値集合(以下、アンサンブルと呼ぶ) $\{x_{n-1|n-1}^{(i)}\}_{i=1}^N$ が準備されているとする。すなわち、

$$p(x_{n-1}|y_{1:n-1}) \cong \frac{1}{N} \sum_{i=1}^N \delta(x_{n-1} - x_{n-1|n-1}^{(i)})$$

となっているとする。これは PF での粒子による近似と同様である。このアンサンブルと観測 y_n から

$$p(x_n|y_{1:n}) \cong \frac{1}{N} \sum_{i=1}^N \delta(x_n - x_n^{(i)})$$

となる $\{x_n^{(i)}\}_{i=1}^N$ を求める以下の一期先予測とフィルタリングの手続きが、EnKF である。

4.1.1 一期先予測

一期先予測においては、各粒子 $x_{n-1|n-1}^{(i)}$ をシステムモデルに基づいて更新し、予測分布のアンサンブル $\{x_{n|n-1}^{(i)}\}_{i=1}^N$ を得る:

$$x_{n|n-1}^{(i)} = f_n(x_{n-1|n-1}^{(i)}, v_n^{(i)}), \quad v_n^{(i)} \sim N(0, Q_n).$$

この更新則は PF の場合と全く同じである。すなわち、EnKF と PF の間で一期先予測の部分でのアルゴリズム上での違いは存在しない。

この状態変数ベクトルの一期先予測において、データ同化の場面では $x_{n|n-1}^{(i)}$ はシミュレーションモデルにもとづいて計算される。このときの $v_n^{(i)}$ は、前章で触れたシミュレーションモデルに課せられた誤差や範囲外からの影響による擾乱である。

4.1.2 フィルタリング

フィルタリング、すなわち一期先予測分布のアンサンブルからフィルタ分布のアンサンブルを得る手続きは、以下の通りである。

まず、観測ノイズ $\{w_n^{(i)}\}_{i=1}^N$ を $N(0, R_n)$ に従って生成し、状態変数ベクトルと観測ノイズのサンプル分散共分散行列を計算する：

$$\begin{aligned} x_{n|n-1}'^{(i)} &= x_{n|n-1}^{(i)} - \frac{1}{N} \sum_{j=1}^N x_{n|n-1}^{(j)}, \\ \hat{V}_{n|n-1} &= \frac{1}{N-1} \sum_{j=1}^N x_{n|n-1}'^{(j)} x_{n|n-1}'^{(j)T}, \\ \hat{R}_n &= \frac{1}{N-1} \sum_{j=1}^N w_n^{(j)} w_n^{(j)T}. \end{aligned}$$

次に、サンプル分散共分散行列からカルマンゲインを求める：

$$(4.1) \quad \hat{K}_n = \hat{V}_{n|n-1} H_n^T (H_n \hat{V}_{n|n-1} H_n^T + \hat{R}_n)^{-1}.$$

最後にカルマンフィルタの更新則によって、フィルタ分布のアンサンブル粒子を得る：

$$(4.2) \quad x_{n|n}^{(i)} = x_{n|n-1}^{(i)} + \hat{K}_n (y_n + w_n^{(i)} - H_n x_{n|n-1}^{(i)}).$$

3 章でも述べた通り、観測のサンプリング時間間隔は一期先予測の時間幅、すなわちシミュレーションモデルにおける更新時間間隔よりも長いので、新たな観測がない場合がある。その時はフィルタリングのステップを省略する。この点もまた、通常のカルマンフィルタや PF における更新と同様である。その一方で (4.2) 式で $w_n^{(i)}$ を加えているように、EnKF によるフィルタリングにおいては、一期先予測の場合と同様に各粒子の状態変数に確率的な要素を加えることによって行なわれることがわかる。この点は、既存のアンサンブル粒子からのリサンプリングを行なう PF とは異なる点である。後述するように、この相違が原因となって両手法間の特徴の相違が現れる。

なお、発生させた観測ノイズの実現値 $w_n^{(i)}$ は、式(4.2)内の \hat{K}_n と、式(4.2)括弧内の $w_n^{(i)}$ を通じて、フィルタ分布のアンサンブル粒子の構成に影響を与える。

4.1.3 初期分布のアンサンブルの設定

EnKF はこのように通常 PF と同様に行なわれるので、設定をしなければいけないものとして $n=0$ での初期分布のアンサンブル $\{x_{0|0}^{(i)}\}_{i=1}^N$ がある。初期分布のアンサンブルの生成法は以下の通りである。まず、 $n=0$ の時点の観測から x_0 として適切と思われる値を補間するなどの何らかの方法により決める。次に、格子点の隣同士との滑らかさは保つという意図で、同じ物理量の変数同士(例えば、温度同士、あるいは風速ベクトルの同方向成分同士等)について近い格子間で相関を持たせた、状態変数ベクトルの実現値を N 個生成し、前に決めた x_0 に摂動的に加えたものを初期分布のアンサンブルとする(Evensen, 2003)。また、実際の問題においては、スピンアップという期間をとる場合がある。これは変数同士が同期するまで、シミュレーションプロセスを行ない、同期した時点での分布を初期分布とするものである。

4.2 サンプル更新の行列表現

本節では、EnKF の分野において導入されている行列表現 (Evensen, 2003) を、統計分野での表現に直す。その上で、次節以降の説明において見通しを良くするために、Evensen (2003) において示されている変形を行なう。

まず、次の行列および表現を導入する：

$$\begin{aligned} X_{n|\cdot} &= [x_{n|\cdot}^{(1)}, x_{n|\cdot}^{(2)}, \dots, x_{n|\cdot}^{(N)}], \\ \frac{1}{N} X_{n|\cdot} \mathbf{1}_N &= [\hat{x}_{n|\cdot}, \hat{x}_{n|\cdot}, \dots, \hat{x}_{n|\cdot}], \\ X'_{n|\cdot} &= X_{n|\cdot} - \frac{1}{N} X_{n|\cdot} \mathbf{1}_N, \\ W_n &= [w_n^{(1)}, w_n^{(2)}, \dots, w_n^{(N)}], \\ Y_n &= [y_n, \dots, y_n] \quad (y_n \text{ を } N \text{ 列並べた行列}). \end{aligned}$$

ただし、 $\mathbf{1}_N$ は全要素 1 の $N \times N$ 行列、 $\hat{x}_{n|\cdot}$ は、 $\{x_{n|\cdot}^{(i)}\}_{i=1}^N$ のサンプル平均である。このとき、次の関係が成立する：

$$\hat{V}_{n|\cdot} = \frac{1}{N-1} X'_{n|\cdot} X'_{n|\cdot T}, \quad \hat{R}_n = \frac{1}{N-1} W_n W_n^T.$$

すると、フィルタリングの式 (4.2) は

$$X_{n|n} = X_{n|n-1} + \hat{V}_{n|n-1} H_n^T (H_n \hat{V}_{n|n-1} H_n^T + \hat{R}_n)^{-1} (Y_n + W_n - H_n X_{n|n-1})$$

と行列表現することが出来る。さらにこれは

$$\begin{aligned} (4.3) \quad X_{n|n} &= X_{n|n-1} + \frac{1}{N-1} X_{n|n-1} \left(I - \frac{1}{N} \mathbf{1}_N \right) X'_{n|n-1 T} H_n^T (H_n \hat{V}_{n|n-1} H_n^T + \hat{R}_n)^{-1} \\ &\quad \times (Y_n + W_n - H_n X_{n|n-1}) \\ &= X_{n|n-1} (I + X'_{n|n-1 T} H_n^T (H_n X'_{n|n-1} X'_{n|n-1 T} H_n^T + W_n W_n^T)^{-1} (Y_n + W_n - H_n X_{n|n-1})) \end{aligned}$$

と変形可能である。ただし、

$$\mathbf{1}_N X'_{n|n-1 T} = O$$

の関係を用いた。従って、式 (4.3) の括弧内を $Z_{n|n-1}$ とすると、

$$(4.4) \quad X_{n|n} = X_{n|n-1} Z_{n|n-1}$$

となる。ここで、 $Z_{n|n-1}$ は $N \times N$ の行列であり、単位行列とそれ以外の 2 つの行列の和となっており、しかも各列毎での行和が 1 となる。以上のことから、EnKF では予測分布のアンサンブルの加重和を用いて、フィルタ分布のアンサンブルを構成している事が分かる。

PF との関連について検討する。PF においては、フィルタ分布のアンサンブルは、尤度によって重み付けられた粒子のリサンプリングによって、与えられるのであった。ここで、EnKF にて実現値を書き並べた行列 $X_{n|\cdot}$ の表現を PF にも用いると、この場合も形式的に式 (4.4) のように

$$X_{n|n} = X_{n|n-1} Z_{n|n-1}$$

と書くことができる．PF の場合ではこの $Z_{n|n-1}$ には、各列毎にいずれかの行に 1 つだけ 1 が入り、残りの行が 0 という行列になっている．この行列 $Z_{n|n-1}$ の第 i 行目に入る 1 の数は、 $x_{n|n-1}^{(i)}$ の全体に対する重みにほぼ比例することになる．

4.3 アンサンブルカルマン smoother

アンサンブルカルマン smoother (Ensemble Kalman Smoother, 以下 EnKS) は PF に対する固定ラグ平滑化 (Particle Smoother, 以下 PS) (北川, 1996; Kitagawa, 1996) と同じ関係にある．平滑化を行なうラグ幅を L とする．ここで、時刻 n' ($n-L \leq n' \leq n-1$) について

$$\hat{J}_{n'|n} \equiv \frac{1}{N-1} \left(\sum_{i=1}^N (x_{n'|n-1}^{(i)} - \hat{x}_{n'|n-1})(x_{n|n-1}^{(i)} - \hat{x}_{n|n-1})^T \right) H_n^T (H_n \hat{V}_{n|n-1} H_n^T + \hat{R}_n)^{-1}$$

$$x_{n'|n}^{(i)} \equiv x_{n'|n-1}^{(i)} + \hat{J}_{n'|n} (y_n + w_n^{(i)} - H_n x_{n|n-1}^{(i)})$$

とすると、この $x_{n'|n}^{(i)}$ が時刻 n 時点で平滑化されたアンサンブルの粒子となる．この式を行列表現すると

$$(4.5) \quad X_{n'|n} = X_{n'|n-1} Z_{n|n-1}$$

と書き換えることができる．この $Z_{n|n-1}$ は EnKF のフィルタリングステップにて用いられた式 (4.4) の $Z_{n|n-1}$ である．

従って、EnKS は過去の粒子を保存しておき、現時点でのゲインによって保存した粒子を修正する手続きとなる．一方、PS の固定ラグ平滑化では、過去の粒子を保存し、現時点でのフィルタリングステップで求める重みに合わせて保存した粒子を逐次リサンプルした．このことから、前節で述べた EnKF と PF の間に成立した関係と同様の関係が、式 (4.5) を通じて EnKS と PS の間にも成立していることがわかる．

5. EnKF の適用範囲

5.1 モーメントの一致性

以下では特定の条件の下、EnKF により得られたアンサンブルのアンサンブル平均とアンサンブル分散共分散行列が、カルマンフィルタにより推定された平均値および分散共分散行列に一致する事を示す (Burgers et al., 1998) ．

以下 $x_{n|n-1}^{(i)}$ および $w_n^{(i)}$ の 2 次までのモーメントが、サンプル数増加にともなって、それぞれ x_n の $y_{1:n-1}$ での条件つき平均と分散、 w_n の平均と分散に収束すると仮定する．すなわち、 $N \rightarrow \infty$ に従って

$$\frac{1}{N} \sum_{i=1}^N x_{n|n-1}^{(i)} \rightarrow x_{n|n-1}, \quad \hat{V}_{n|n-1} \rightarrow V_{n|n-1},$$

$$\frac{1}{N} \sum_{i=1}^N w_n^{(i)} \rightarrow 0, \quad \hat{R}_n \rightarrow R_n$$

となるとする．ただしここで $x_{n|n-1}, V_{n|n-1}$ は

$$x_{n|n-1} = E[x_n | y_{1:n-1}], \quad V_{n|n-1} = E[(x_n - x_{n|n-1})(x_n - x_{n|n-1})^T]$$

である．カルマンフィルタによって得られるカルマンゲインを K_n とすると、 $N \rightarrow \infty$ で式 (4.1) の \hat{K}_n は

$$\hat{K}_n \rightarrow K_n$$

となるから、フィルタ分布のアンサンブルのサンプル平均 $\hat{x}_{n|n}$ は

$$\hat{x}_{n|n} \longrightarrow x_{n|n-1} + K_n(y_n - H_n x_{n|n-1}) = x_{n|n}$$

と漸近的にカルマンフィルタによって得られるフィルタ分布の平均に一致する。

以下ではさらに、フィルタ分布のアンサンブルのサンプル分散共分散行列も近似的に一致することを見る。今フィルタ分布のアンサンブルのサンプル分散共分散行列 $\hat{V}_{n|n}$ は、

$$\begin{aligned} \hat{V}_{n|n} &= \frac{1}{N-1} \sum_{i=1}^N \{(x_{n|n}^{(i)} - \hat{x}_{n|n})(x_{n|n}^{(i)} - \hat{x}_{n|n})^T\} \\ &= (I - \hat{K}_n H_n) \hat{V}_{n|n-1} + \frac{1}{N-1} \sum_{i=1}^N \{(I - \hat{K}_n H_n)(x_{n|n-1}^{(i)} - \hat{x}_{n|n-1}) w_n^{(i)T} \hat{K}_n^T \\ &\quad + \hat{K}_n w_n^{(i)} (x_{n|n-1}^{(i)} - \hat{x}_{n|n-1})^T (I - \hat{K}_n H_n)^T\} \\ &\quad (\because \hat{K}_n = \hat{V}_{n|n-1} H_n^T (H_n \hat{V}_{n|n-1} H_n^T + \hat{R}_n)^{-1}) \\ &\approx (I - \hat{K}_n H_n) \hat{V}_{n|n-1} \\ &\quad \left(\because \frac{1}{N-1} \sum_{i=1}^N w_n^{(i)} (x_{n|n-1}^{(i)} - \hat{x}_{n|n-1})^T \approx O \right) \\ &\longrightarrow (I - K_n H_n) V_{n|n-1} \quad (N \longrightarrow \infty) \end{aligned}$$

となるので、 w_n と x_n が無相関であれば、 $\hat{V}_{n|n}$ はカルマンフィルタによって得られるフィルタ分布の分散共分散行列 $V_{n|n}$ に漸近的に一致する。なお、式(4.1)において、発生させた観測ノイズアンサンブルのサンプル分散共分散行列 \hat{R}_n のかわりに観測ノイズのもとの分散共分散行列 R_n を用いた場合、上記の式変形の上から 2 番目の等号は、厳密に等しくはならず近似的に成立する。

なお、Evensen (1994) によって当初導入された際には、式(4.2)内に $w_n^{(i)}$ の項が含まれていなかったが、このときには $\hat{V}_{n|n}$ は

$$\hat{V}_{n|n} \longrightarrow (I - K_n H_n) V_{n|n-1} (I - K_n H_n)^T \quad (N \longrightarrow \infty)$$

となり、分散共分散行列がカルマンフィルタによるものと一致しなくなり、誤りである (Burgers et al., 1998)。

以上から、EnKF は状態変数と観測ノイズの 2 次モーメントまでを利用してフィルタ分布のアンサンブルを構成する手法であることが確認される。このことは、予測分布およびフィルタ分布が正規分布に従う場合は、EnKF がサンプル数 $N \rightarrow \infty$ の時に、近似的に正しい推定を行なう事を示している。

5.2 非線形観測モデルへの拡張法

PF では直接的に扱う事が可能な非線形観測の場合について、EnKF では直接的に扱う事は不可能である。一つの解決策としては、拡張カルマンフィルタと同様、非線形観測モデル内の h_n のアンサンブル平均周りでのヤコビ行列を計算し、観測モデル(3.1)の H_n として適用する方法がある。しかし、アンサンブルカルマンフィルタの利点であったオンラインでの微分演算が不要というメリットがなくなるため、使われていない。これに代わる解決策として、以下のような方法が取られている (Evensen, 2003)。まず、状態ベクトルを次のように拡張する：

$$(5.1) \quad \tilde{x}_n = [x_n^T, h_n(x_n)^T]^T.$$

このとき、システムモデルおよび観測モデルは

$$\begin{aligned}\tilde{x}_n &= [f_n(x_{n-1}, v_n)^T, h_n(f_n(x_{n-1}))^T]^T \equiv \tilde{f}_n(\tilde{x}_{n-1}, v_n), \\ y_n &= [O_{l \times k}, I_{l \times l}] \tilde{x}_n + w_n \equiv \tilde{H}_n \tilde{x}_n + w_n\end{aligned}$$

と形式的に線形の形に書き表すことができる。ただし k, l はそれぞれ x_n, y_n の次元、 $I_{l \times l}$ は l 次元単位行列、 $O_{l \times k}$ は l 行 k 列の零行列である。この拡張された観測モデルに対してフィルタ分布のアンサンブルを更新する式を考える。

今、予測誤差を $\varepsilon_n = y_n - \tilde{H}_n \tilde{x}_{n|n-1}$ とすると、

$$\begin{aligned}\text{Cov}(x_n, \varepsilon_n) &= \text{Cov}(x_n, \tilde{H}_n(\tilde{x}_n - \tilde{x}_{n|n-1}) + w_n) \\ &= \text{Cov}([I_{k \times k}, O_{k \times l}] \tilde{x}_n, \tilde{H}_n(\tilde{x}_n - \tilde{x}_{n|n-1}) + w_n) \\ &\approx \frac{1}{N-1} [I_{k \times k}, O_{k \times l}] \tilde{X}'_{n|n-1} \tilde{X}'_{n|n-1 T} \tilde{H}_n^T \\ &= \frac{1}{N-1} X'_{n|n-1} \tilde{X}'_{n|n-1 T} \tilde{H}_n^T, \\ \text{Var}(\varepsilon_n) &\approx \frac{1}{N-1} (\tilde{H}_n \tilde{X}'_{n|n-1} \tilde{X}'_{n|n-1 T} \tilde{H}_n^T + W_n W_n^T)\end{aligned}$$

となるので、フィルタリングの式は

$$\begin{aligned}\tilde{X}_{n|n} &= \tilde{X}_{n|n-1} + X'_{n|n-1} \tilde{X}'_{n|n-1 T} \tilde{H}_n^T (\tilde{H}_n \tilde{X}'_{n|n-1} \tilde{X}'_{n|n-1 T} \tilde{H}_n^T + W_n W_n^T)^{-1} \\ &\quad \times (Y_n + W_n - \tilde{H}_n \tilde{X}_{n|n-1})\end{aligned}$$

となる。このように、非線形観測が含まれる際の EnKF は、状態変数ベクトルを拡張し、変形したフィルタリングの式を用いることによって行なわれる。ここでも、非線形観測の 2 次までのモーメントを利用して、高次モーメントの情報までは含まれない。

6. PF との比較

6.1 相違点

本節では、EnKF と PF の間での相違点について議論する。両手法をデータ同化に用いた場合、予測のステップに関しては同じ手続きを踏むことになるため、比較すべき点はフィルタリングのステップでの相違点である。PF と比較した場合、EnKF には

- フィルタ分布のアンサンブル構成に 2 次モーメントまでを利用
- 計算負荷の大きい逆行列計算が必要
- アンサンブルはフィルタリングステップにおいて非退化

という相違点がある。このうち前 2 点は PF の方が有利な点、後の 1 点が EnKF の有利な点である。

EnKF ではフィルタ分布のアンサンブル構成時、2 次モーメントまでを利用する。よって、EnKF で推定されたフィルタ分布はカルマンフィルタによって推定した場合の結果と漸近的に一致する。しかし、シミュレーションモデルは非線形性があるため、フィルタ分布は正規分布に従わない。従って、EnKF においてはアンサンブル数を増加させても、必ず分布の近似精度、あるいはパラメータの推定精度が上がるとは限らない。一方、PF の場合はアンサンブル数を増加させれば、フィルタ分布が正規分布でなくとも近似精度が上がるので、この点は PF の方が有利である。

2 点目の EnKF が計算負荷の大きい逆行列を必要とする点について議論する。PF においては通常、予測分布のアンサンブルから重みの比によってリサンプルすることによりフィルタ分

布のアンサンブルを構成する．従って，予測分布のアンサンブルからフィルタ分布のアンサンブルを構成するためには，重みの計算をする際の逆行列計算を除いて，逆行列の計算は必要無い．重み計算時に必要な逆行列の計算は，通常計算負荷が小さい．例えば観測ノイズがガウス分布であったとする．要素間で無相関ということ仮定してよい（観測間のノイズが無相関の場合，観測ノイズの分散共分散行列は対角行列でありその逆行列の計算量は非常に少ない．一方，EnKFにおいてフィルタ分布のアンサンブルを構成する際には，式(4.1)から分かるように，計算負荷の大きい逆行列計算が必要である．従って，計算量としてはPFに比べて不利である．加えて，問題が大規模であることから並列計算機で計算する機会が多いが，この逆行列の計算を並列化するのは難しいのに対し，リサンプルを並列化する事は難しくない．またリサンプルそのものも，行列演算が入らないので，その分だけPFの方がより速くなる．

3点目のアンサンブルの非退化性について議論する．退化とは，アンサンブル内に同一の実現値が存在することである．退化の度合いが大きい場合，すなわち同一の実現値が多い場合，本来の連続分布を近似できないことが問題となる．式(4.3)内の括弧内に単位行列が入っていることからわかるように，EnKFでは，フィルタ分布のアンサンブル構成時に同一の実現値が出来る事はほとんどなく，基本的には退化しない．一方PFは，予測分布のアンサンブルよりリサンプルを行なうことから，特殊な場合を除き必ず退化する．予測分布のアンサンブルを多数発生させておき，それよりリサンプルする方法も考えられるが，実問題のオーダ（少なくとも $10^3 \sim 10^4$ ）を考慮すると現実的でない．その上，状態空間内における粒子密度を考慮に入れると，事前分布のアンサンブルを多数発生させても少数粒子の尤度が大きくなり過ぎるので，結局同一の粒子がリサンプルされる可能性が高いと考えられる．Rejection Control (Liu et al., 1998)のように，重みが一定以下のときに確率的に棄却する方法により，退化しないように粒子を発生させることも理論的には可能だが，この場合には棄却される確率が極めて高くなり，アンサンブルを発生させる回数が増えてしまい，やはり現実的でない．以上はPFの方が不利な点であり，実際にデータ同化の問題にPFを適用する上では，アルゴリズム上かモデル上で何らかの対処を行なう必要がある．

6.2 数値実験

本節では，非線形システム・非線形観測の小規模なシステムに対する状態推定ならびにパラメータ推定を，EnKF, EnKS, PF および PS によって行ない，その結果を比較する．

データ y_n は，非線形状態同定の問題としてよく用いられるシステム (Carlin et al., 1992; Gordon et al., 1993; Kitagawa and Gersch, 1996; Kitagawa, 1998)

$$(6.1) \quad \begin{cases} x_n = \frac{1}{2}x_{n-1} + \frac{25x_{n-1}}{1+x_{n-1}^2} + 8\cos(1.2n) + v_n \\ y_n = \frac{x_n^2}{20} + w_n \\ x_0 \sim N(0,5), \quad v_n \sim N(0,1), \quad w_n \sim N(0,10), \quad n = 1, \dots, 100 \end{cases}$$

によって生成した．発生系列の一例が，図2である．

このシステムに対し，次の2通りの数値実験を行なった：

- ・ 式(6.1)に従い， x_n を推定．
- ・ 式(6.1)に従うが，さらに v_n の分散を未知として自己組織化モデル (Kitagawa, 1998) によってシステムモデルの分散も推定．

EnKF については，式(5.1)の非線形観測時の状態変数ベクトルの拡張を利用した．以下この2通りの数値実験についてその概要と結果を述べる．

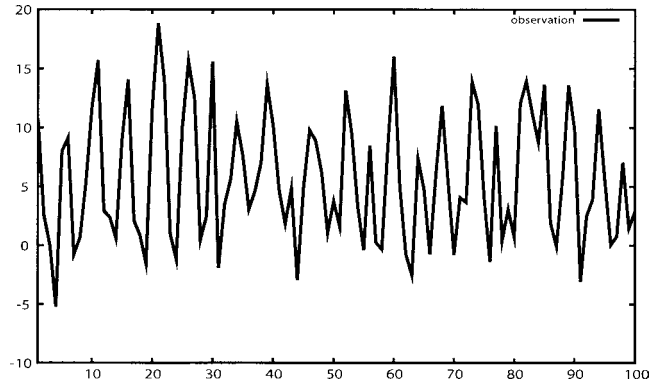


図 2. 観測 y_n の例．横軸が時刻 n ，縦軸が y_n である．

表 1. 手法別状態推定誤差．100 回の実験間の平均である．

アンサンブル数	PF	EnKF	PS	EnKS
100	1841.76	2853.11	567.84	1618.73
1000	1710.01	2779.09	404.90	1470.93
2500	1701.90	2771.28	397.03	1447.65

6.2.1 状態変数のみ推定する場合

データ系列 y_n を発生させ、それに対し EnKF, EnKS, PF 及び PS の各推定法を適用した．アンサンブル数 N が 100, 1000, 2500 の各場合について、実験を行なった．これを 100 系列の y_n に対して行ない、元のデータ系列における x_n との差の 2 乗和

$$\sum_{n=1}^{100} (\hat{x}_n - x_n)^2$$

の実験間平均により、その推定の良さを評価する．ここで、 \hat{x}_n は各時刻毎のアンサンブルの平均値である．なお、EnKS および PS におけるラグ幅 L は 20 とした．

数値実験結果は表 1 である．

これにより、以下の 2 点が確認される．1 点目は、この非線形観測を含むシステムに対する推定は、同サンプル数で比較した場合に、EnKF よりも PF, EnKS よりも PS の方が性能が良いという事である．2 点目は、粒子数増加に対応した誤差の減少量が、EnKF よりも PF, EnKS よりも PS の方が大きいことである．すなわち、粒子数を増やすことのメリットはどの手法に対しても存在するが、PF および PS の方がより大きいということになる．このことは、より正確な推定をするという観点では、PF および PS にして粒子数を増やすことがこれらの間で最も有効であるということの意味する．その一方で、実システムに適用する場合、シミュレーションでの時間およびメモリの負荷が大きいので、粒子数をあまり多くはできないという事情があり、この部分のトレードオフが問題となる．

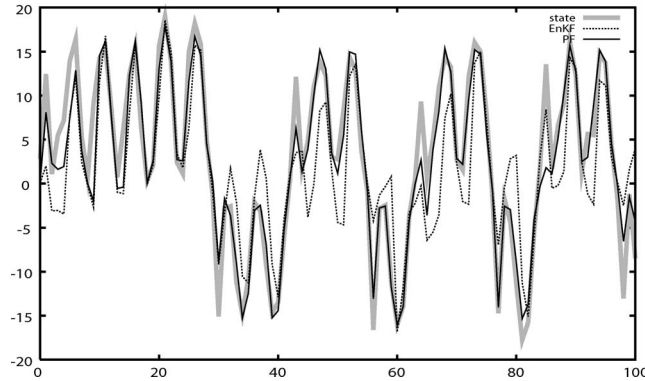


図 3. 状態 x_n の推定例．灰色太線が真の x_n ，点線が EnKF による推定，実線が PF による推定である．

また，EnKF と PF それぞれにて推定された x_n のプロットの一例が図 3 である．ただしアンサンブル数は 100 であり，灰色太線が真値 x_n ，点線，実線がそれぞれ EnKF，PF による \hat{x}_n である．

この図からも，PF による推定の方が EnKF よりもより良く追従していることが確認できる．さらに，EnKF による推定値 \hat{x}_n が状態 x_n から外れる場合として， $n = 45$ 付近のように x_n の絶対値が減少から増加に転じる部分で起こっている場合が多いことが確認できる．これは，各時刻での x_n の分布が，0 を挟んでの双峰形 (Doucet et al., 2001) になっていることによる．EnKF においては，正規分布を仮定しているのと同様であり，単峰の分布を仮定していることになる．これは，EnKF においては，本来の分布に含まれる二つの峰の内の確率の低い峰の影響を受けて，修正が十分に効かないためである．

6.2.2 システムノイズの分散も推定する場合

数値実験の設定は，当てはめるシステムモデルが異なる点を除き状態変数のみ推定する場合と同様である．当てはめるシステムモデルは

$$\begin{bmatrix} x_n \\ \theta_n \end{bmatrix} = \begin{bmatrix} \frac{1}{2}x_{n-1} + \frac{25x_{n-1}}{1+x_{n-1}^2} + 8\cos(1.2n) \\ \theta_{n-1} \end{bmatrix} + \begin{bmatrix} v_n \\ u_n \end{bmatrix}$$

$$x_0 \sim N(0, 5), v_n \sim N(0, \exp(\theta_n)), \theta_0 \sim U(-2, 6)$$

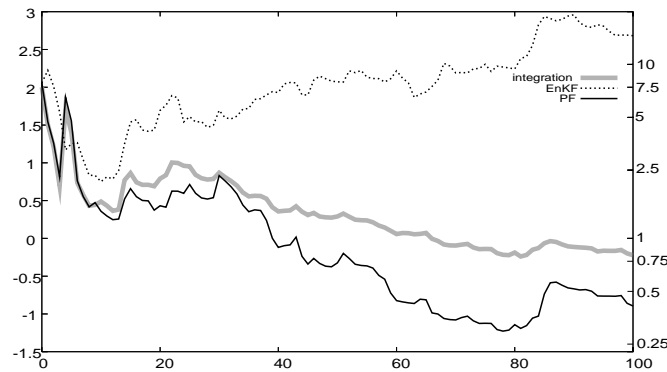
である． $\exp(\theta_n)$ が v_n の未知分散となる．このシステムモデルの θ_n を推定することにより， $\hat{\theta}_n$ を得る．このときに $\exp(\hat{\theta}_n)$ とすることで v_n の分散が推定される．ここで，

$$\hat{\theta}_n = \frac{1}{N} \sum_{i=1}^N \theta_n^{(i)}.$$

とする．元の系列の分散は 1 であったから， $\exp(\hat{\theta}_n)$ が 1，すなわち $\hat{\theta}_n$ が 0 に近付けば， v_n の分散として 1 が正しく推定されたことになる．ここで，分散が時不変というモデルに対応するのは $u_n = 0$ とした場合である．よって，EnKF では $u_n = 0$ とした．PF においてこのように設定すると，推定パラメータ $\hat{\theta}_n$ を表すアンサンブルが退化してしまう．そのため，逐次的に推定している途中に本来推定されるべき値を表現するアンサンブルが無くなってしまったときに，

表 2. 手法別のパラメータ $\hat{\theta}_{100}$ 推定誤差 . 100 回の実験間の平均と標準偏差である .

	PF	EnKF	数値積分
平均	0.021	1.334	-0.020
標準偏差	0.662	0.933	0.568

図 4. EnKF 及び PF による $\hat{\theta}_n$ 推定例 . 灰色太線が数値積分による推定, 点線が EnKF による推定, 実線が PF による推定であり, 縦軸は左側が $\hat{\theta}_n$, 右側が $\exp(\hat{\theta}_n)$ である .

その本来存在すべきアンサンブルを, 以降の逐次推定の過程でもとに戻すことが出来なくなってしまふ . この退化を避けるために PF では $N(0, 0.0025)$ に従う微小分散のノイズを加えた .

時不変分散のモデルの場合, $u_n = 0$ なので, $y_{1:100}$ を用いた推定値は $\hat{\theta}_{100}$ で与えられる . すなわち, 固定区間平滑化による推定の結果と等しくなる . 以下でフィルタリングのみを行ない, 平滑化を行っていないのはこの理由による .

このシステムに従い, アンサンブル数 500 の条件下でシステム分散を推定する実験を 100 系列の y_n に対して行なった . また, EnKF の場合と同様に $u_n = 0$ とし, Kitagawa (1987) に従い数値積分することにより推定する実験も行なった . この推定による結果に近ければ, より正確な手法であるということになる . その結果が表 2 である . これは, $\hat{\theta}_{100}$ の実験間での平均と標準偏差をとったものである .

この結果から, 数値積分による推定により近い推定は PF による場合であり, しかも推定値として期待される 0 に近い値が得られていることがわかる . このことは, 実験結果のプロット例である図 4 を見ればはっきりする .

また, 各時刻 n での $\{\theta_n^{(j)}\}_{j=1}^N$ のアンサンブル数の変動を表したものが図 5 および 6 である .

同時刻 $n = 100$ での EnKF と PF のアンサンブル数の分布を比べると, EnKF の方がより広い範囲に存在し, しかも期待される 0 から離れた 2 付近で最大となっているのに対し, PF では範囲が狭くしかも比較的 0 に近い部分に集中していることが確認できる . さらに, PF の方は早い時刻から 0 周辺に集中していることが確認できる . この理由については次項において触れる .

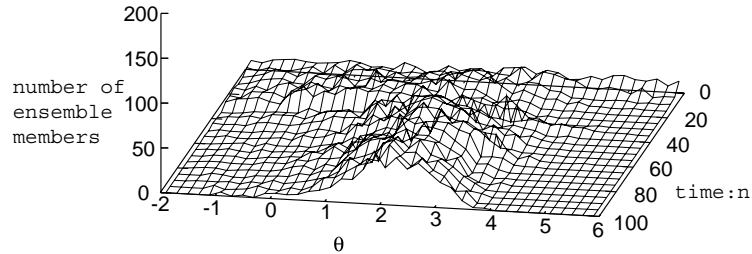


図 5. EnKF における $\{\theta_n^{(j)}\}_{j=1}^N$ の分布．各軸は時刻 n ，パラメータ θ_n ，アンサンブルメンバー数を表す．

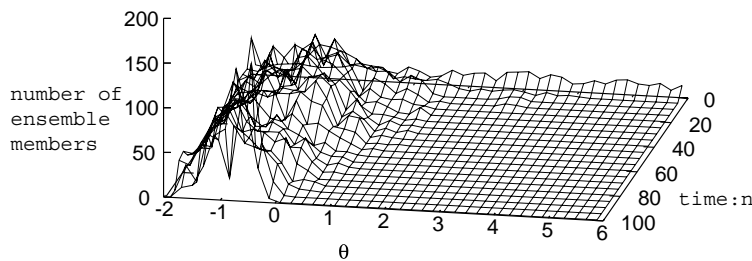


図 6. PF における $\{\theta_n^{(j)}\}_{j=1}^N$ の分布．各軸は時刻 n ，パラメータ θ_n ，アンサンブルメンバー数を表す．

6.2.3 考察

状態変数ベクトルの推定，ならびにパラメータの推定の 2 つの結果において，非線形システムにおける PF の推定に 2 つの意味での優位性が見られる．1 点目は，状態変数ベクトルおよびパラメータの各時点での近似精度の良さである．その理由は，フィルタ分布が正規分布でないという条件の下，EnKF の方は正規分布を仮定するカルマンフィルタによる推定値と 2 次モーメントまでがほぼ一致するようなフィルタ分布のアンサンブルを構成しているのに対し，PF はフィルタ分布に正規分布を仮定せずに観測モデルそのものによって自然に導かれるフィルタ分布と漸近的に一致するアンサンブルを構成するので，フィルタ分布の近似精度が著しく異なるためである．

2 点目は，パラメータの初期推定値が本来のパラメータから外れている場合に，その収束速度が PF の方は数値積分によるものとほぼ同じであるのに対し，EnKF はそうはなっていない点である．これも，先ほど触れた分布を表現する柔軟性の違いによるものであると考えられる．すなわち， $\{\theta_n^{(j)}\}_{i=1}^N$ を構成する際に，PF の方は正規分布を仮定せず，近似精度が EnKF より高く，数値積分により得られる θ_n に近いとめと考えられる．

その一方で，PF と自己組織化モデルを用いて時不変パラメータを推定する場合，ノイズを加えることは本質的に必要であり，欠点となる．この問題への対処は，パラメータが高次元の場合には難しい．実際の問題における物理定数のような，低次元の時不変パラメータを推定する場合においては，モンテカルロ誤差が存在するものの，尤度の計算を通じて最尤推定を行なうことが考えられる．

7. 結論

本論文では、逐次データ同化の非線形・非ガウス型状態空間表現での定式化とその中での変数の意味を考察した。また、手法として広く用いられている EnKF についてその性質を考察した。さらに、PF との類似点および相違点を議論し、特にシステムモデルが非線形でかつ観測モデルも非線形の場合において、状態推定ならびにパラメータ推定を行なったときの PF の優位性を、状態ベクトルの次元が低いモデルに対する数値実験により示した。ただし、本論文における数値実験において用いられたシミュレーションモデルとデータの次元は、実際にデータ同化を行なう場合のシミュレーションモデルに比べて、非常に小さいものであることに触れておく。このことは、PF を実際のデータ同化に用いる上での問題点、すなわち推定すべき状態変数ベクトルの次元が大きい場合、リサンプルに耐えるのに十分なアンサンブル数を確保することが難しいという問題点(van Leeuwen, 2003)を避けていることになる。

今後この分野において進むべき方向は、システムモデルの非線形性が強い場合の計算手法、ならびにシステムモデルだけでなく観測モデルも非線形であり、かつ状態変数ベクトルや観測変数ベクトルの次元がそこそこ大きい場合(数十次元から数百次元程度)の計算手法を与えることである。EnKF の文脈においてもこの点に関するより深い研究が必要である事は言われている(Evensen, 2003)。本論文内で見たとおり、PF と EnKF については計算手法的に関連づけが可能であり、両者の欠点を補う形、すなわち

- ・ 正規分布の仮定を置かない
- ・ 大きな行列の逆行列計算が不要
- ・ アンサンブルがほとんど退化しない

といった方向での、アンサンブルベースの方法論が必要になると考えられる。このことにより、最終的な目標である、「非線形性をもつ大規模なモデルで書かれる実問題に対する、効果的なデータ同化の方法」を与えることになると考えられる。

従って、PF と EnKF 両者の欠点を補う方法を考えることになるが、Pham(2001)によって EnKF の smooth bootstrap(Efron and Tibshirani, 1993; Stavropoulos and Titterton, 2001)との関連が触れられており、その方向からのアプローチなどが両者の欠点を補う形での解決策の候補として考えられる。

謝 辞

本論文の成果の一部は科学技術振興機構・戦略的基礎研究推進事業(JST/CREST)の援助を受けて行なわれました。また、査読者には建設的かつ貴重なコメントを頂きました。この場を借りて感謝致します。

参 考 文 献

- Anderson, J. L.(2001). An ensemble adjustment Kalman filter for data assimilation, *Monthly Weather Review*, **129**, 2884-2903.
- 有本 卓(1977). 『カルマン・フィルター』, 産業図書, 東京.
- Bennett, A. F.(2002). *Inverse Modeling of the Ocean and Atmosphere*, Cambridge University Press, Cambridge.
- Burgers, G., van Leeuwen, P. J. and Evensen, G.(1998). Analysis scheme in the ensemble Kalman filter, *Monthly Weather Review*, **126**, 1719-1724.

- Carlin, B. P., Polson, N. G. and Stoffer, D. S. (1992). A Monte Carlo approach to nonnormal and nonlinear state-space modeling, *Journal of the American Statistical Association*, **87**, 439–500.
- Courtier, P., Thepaut, T. and Hollingsworth, A. (1994). A strategy for operational implementation of 4DVAR, using an incremental approach, *Quarterly Journal of the Royal Meteorological Society*, **120**, 1367–1387.
- Daley, R. (1991). *Atmospheric Data Analysis*, Cambridge University Press, Cambridge.
- Doucet, A., de Freitas, N. and Gordon, N. (2001). An introduction to sequential Monte Carlo methods, *Sequential Monte Carlo Methods in Practice* (eds. A. Doucet, N. de Freitas, and N. Gordon), 3–14, Springer-Verlag, New York.
- Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, New York.
- Evensen, G. (1994). Sequential data assimilation with a non-linear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, *Journal of Geophysical Research*, **99**, 10143–10162.
- Evensen, G. (2003). The ensemble Kalman filter: Theoretical formulation and practical implementation, *Ocean Dynamics*, **53**, 343–367.
- Evensen, G. and van Leeuwen, P. J. (2000). An ensemble Kalman smoother for nonlinear dynamics, *Monthly Weather Review*, **128**, 1852–1867.
- Gordon, N. J., Salmond, D. J. and Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation, *IEE Proceedings F*, **140**, 107–113.
- 樋口知之 (1996). 遺伝的アルゴリズムとモンテカルロフィルタ, *統計数理*, **44**, 19–30.
- Higuchi, T. (2003). Data assimilation with Monte Carlo mixture Kalman filter toward space weather forecasting, *Proceedings of International Symposium on Information Science and Electrical Engineering 2003*, 122–125.
- Kalman, R. E. (1960). A new approach to linear filtering and prediction problems, *Journal of Basic Engineering*, **82**, 35–45.
- 片山 徹 (1983). 『応用カルマンフィルタ』, 朝倉書店, 東京.
- Kitagawa, G. (1987). Non-Gaussian state-space modeling of nonstationary time series (with discussion), *Journal of the American Statistical Association*, **82**, 1032–1063.
- 北川源四郎 (1996). モンテカルロ・フィルタおよび平滑化について, *統計数理*, **44**, 31–48.
- Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models, *Journal of Computational and Graphical Statistics*, **5**, 1–25.
- Kitagawa, G. (1998). Self-organizing state space model, *Journal of the American Statistical Association*, **93**, 1203–1215.
- 北川源四郎 (2005). 『時系列解析入門』, 岩波書店, 東京.
- Kitagawa, G. and Gersch, W. (1996). *Smoothness Priors Analysis of Time Series*, Lecture Notes in Statistics, No. 116, Springer-Verlag, New York.
- Lisæter, K. A., Rosanova, J. and Evensen, G. (2003). Assimilation of ice concentration in a coupled ice-ocean model, using the ensemble Kalman filter, *Ocean Dynamica*, **53**, 368–388.
- Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*, Springer-Verlag, New York.
- Liu, J. S., Chen, R. and Wong, W. H. (1998). Rejection control and sequential importance sampling, *Journal of the American Statistical Association*, **93**, 1022–1031.
- Lorenz, E. N. (1963). Deterministic nonperiodic flow, *Quarterly Journal of Atmospheric Science*, **20**, 130–141.
- Manda, A., Hirose, N. and Yanagi, T. (2003). Application of a nonlinear and non-Gaussian sequential estimation method for an ocean mixed layer model, *Engineering Sciences Reports, Kyushu University*, **25**, 285–289.

- Pham, D. T. (2001). Stochastic methods for sequential data assimilation in strongly nonlinear systems, *Monthly Weather Review*, **129**, 1194–1207.
- Stavropoulos, P., Titterton, D. M. (2001). Improved particle filters and smoothing, *Sequential Monte Carlo Methods in Practice* (eds. A. Doucet, N. de Freitas and N. Gordon), 295–317, Springer-Verlag, New York.
- 多田英夫, 村上茂教(1997). データ同化の現状と展望・第1章データ同化概論, 数値予報課報告・別冊, **43**, 1–16.
- Talagrand, O. and Courtier, P. (1987). Variational assimilation of meteorological observations with the adjoint vorticity equation I: Theory, *Quarterly Journal of the Royal Meteorological Society*, **113**, 1311–1328.
- van Leeuwen, P. J. (2003). A variance-minimizing filter for large-scale applications, *Monthly Weather Review*, **131**, 2071–2084.
- Wunsch, C. (1996). *The Ocean Circulation Inverse Problem*, Cambridge University Press, Cambridge.

Data Assimilation: Concept and Algorithm

Kazuyuki Nakamura,¹ Genta Ueno² and Tomoyuki Higuchi²

¹Department of Statistical Science, The Graduate University for Advanced Studies; JST CREST

²The Institute of Statistical Mathematics; JST CREST

The data assimilation technique, which was developed for meteorology and oceanography, aims at accommodating states of a physical simulation model to observations. It is motivated to provide a good initial condition for the nonlinear simulation model and to realize an online model parameter estimation. To attain these purposes, sequential data assimilation such as the Ensemble Kalman Filter is used. Compared to the Particle Filter, the Ensemble Kalman Filter has a similar methodology in view of the ensemble-based method, but it has different structures. For example, linear calculation is done instead of resampling at the filtering step and approximated probability density does not converge to real probability density even if the number of ensemble members goes to infinity in almost all cases. This paper reviews the concept of data assimilation, and formulates using the nonlinear state space model. On the basis of this formulation, the Ensemble Kalman Filter is reviewed, focusing on the difference and similarity between the Ensemble Kalman Filter and the Particle Filter. After the review, we demonstrate assimilation experiments of nonlinear observations with small scale simulation and the superiority of the Particle Filter to the Ensemble Kalman Filter under these conditions. Applicability to actual systems including nonlinear observations is also suggested.