2D-04

異常値を考慮したアンサンブルフィルタの設計

佐藤 哲 †

NHN テコラス株式会社 データサイエンスチーム †

1. はじめに

確率密度関数を実現値で近似するアンサンブルフィルタは，任意の非線形確率密度関数を扱える汎用性の高い手法である．しかし，通常は想定できない大きな外れ値が瞬間的ではなく一定の期間中観測された場合，アンサンブルフィルタの枠組みの中では異常値としての検出や状態推定精度の維持が難しくなる．そこで本発表では，実現値の集合が低次元空間の点群とみなせることを利用し，点群の大域的な情報を抽出する位相的データ解析を用いて異常値検出やモデルパラメータの適切さを推定するする手法を検討する．

2. 状態空間モデルにおける異常値を考慮したアンサンブルフィルタ

本発表では，状態空間モデルにおいてモデルの確率密度関数を複数の実現値により近似し状態を推定する手法全般をアンサンブルフィルタと呼ぶ．また，簡単のためにデータは 1 次元の時系列データであるとする．

研究対象とする状態空間モデルは，以下のようなシステムモデルと観測モデルから構成される：

$$\begin{cases} x_k &= F(x_{k-1}, v_k) \\ y_k &= H(x_k, \omega_k) \end{cases} \tag{1}$$

ここで，x_k 及び y_k は，時刻 k での状態ベクトル及び観測ベクトル，v_k 及び ω_k はシステムノイズ及び観測ノイズと呼ばれる量である．状態ベクトルは観測できない隠れ変数であり，1 ステップ前の状態ベクトルとシステムノイズによって決まると仮定し，観測可能な観測ベクトルは状態ベクトル及び観測ノイズによって決まると仮定する．さらにアンサンブルフィルタにおいては，アンサンブルカルマンフィルタ [1] では線形性・システムモデルのガウス性を，粒子フィルタ [2] では実現値の数が十分に多いことを，アンサンブル変換法 [3] ではモデルが十分に良い精度であることを仮定する．その仮定の下で，(1) システムモデルによる予測値の計算，(2) 観測モデルに基づく予測値の修正，を繰り返すことで状態ベクトルを推定することがアンサンブルフィルタの一般的なアルゴリズムである．いずれの手法も，仮定したモデルから外れた入力値に対しては対処が難しいが，カルマンフィルタとは異なりアンサンブルフィルタでは非ガウス性のノイズを利用できるため，予め外れ値を考慮したシステムノイズを仮定する手法が考えられる．システムノイズがパラメータ θ による確率密度関数 $r(\theta)$ に従う場合を考えると，

$$v \sim r(\theta) \tag{2}$$

異常値が混入する場合，ノイズの確率密度関数として，通常とは異なるパラメータ $\hat{\theta}$ によるノイズが確率 $1-\alpha$ で混入するような混合分布の導入が考えられる：

$$v \sim \alpha r(\theta) + (1-\alpha) r(\hat{\theta}) \tag{3}$$

このモデルは異常値を考慮しているが，観測データが異常値であるかどうか判定する手法が必要である．例えば観測ノイズに異常値を捉える項を加え，EM アルゴリズムを用いてその項の寄与率を推定する手法も考えられるが，本研究では状態ベクトルに対するノイズ，異常値，継続的な異常値などスケールの異なる非連続的な変化を捉えるため，順序保存符号化と位相的データ解析（TDA）を組み合わせた手法を利用する [4]．TDA はフィルトレーションと呼ばれるスケールを変化せて幾何的な特徴量を抽出する操作があるため，色々な様相を取り得るノイズや異常値に対するデータ解析が可能であると期待される．

3. アンサンブルフィルタの適用及びアンサンブル集合の解析実験

アンサンブルフィルタや TDA は任意次元のデータに適用可能であるが，簡単のために時系列でスカラー値が入力される 1 次元時系列データに対する実験結果を紹介する．

入力データは 4 時間分のネットワークのパケットキャプチャデータであり，タイムスタンプを元に計算する 5 分単位のパケットサイズの合計バイト数をフィルタリング対象としている．また，データの中に連続した 1 時間の Denial of Service 攻撃を模した，通常の揺れ幅の約 100 倍の大きさのノイズを異常値として人為的に加えている．この入力データを観測値とし，アンサンブルフィルタにより状態を推定する．パケットデータの場合は，電気的なノイズ等によりデータが破損しても誤り訂正や検出が行われるため，観測ノイズはゼロに近い．従って状態推定結果としては，観測値に近い結果が得られることが望ましい．

本発表では，システムモデル，観測モデルともに状態ベクトル及び観測ベクトルとノイズの線形和であると仮定し，システムノイズは混合コーシー分布，観測ノイズはガウス分布に従うとした．すなわち状態空間モデルは以下である：

$$\begin{cases} x_k &= x_{k-1} + v_k \\ y_k &= x_k + \omega_k \end{cases} \tag{4}$$

Designing Ensemble Filter Considering Anomalies
†Tetsu R. Satoh, NHN Techous Corp.

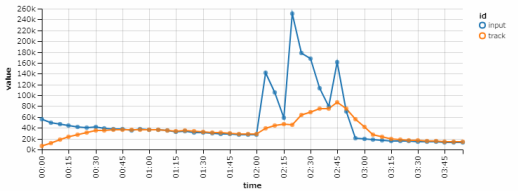

図 1: アンサンブルフィルタ結果 ($\alpha = 1.0$)

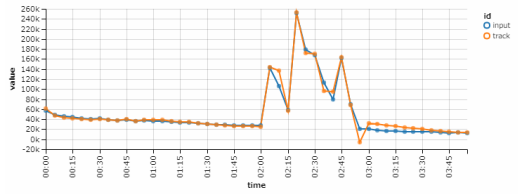

図 2: アンサンブルフィルタ結果 ($\alpha = 0.5$)

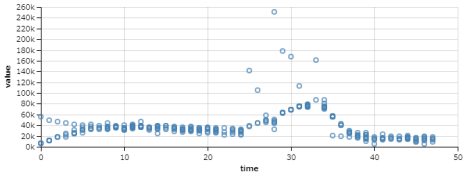

図 3: アンサンブル集合 ($\alpha = 1.0$)

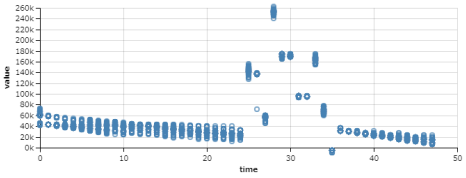

図 4: アンサンブル集合 ($\alpha = 0.5$)

ここで，$v_k \sim \alpha\tau\{\pi(t^2+\tau^2)\}^{-1} + (1-\alpha)\hat{\tau}\{\pi(t^2+\hat{\tau}^2)\}^{-1}$，$\omega_k \sim N(\mu, \sigma^2)$ である．アンサンブルフィルタは，実装が最も簡単な粒子フィルタを用いる．パラメータは，粒子数は 500，システムノイズのパラメータは $\tau = 100.0, \hat{\tau} = 400.0\tau$，観測ノイズのパラメータは $\mu = 0, \sigma = 200.0$ とした．状態ベクトルの初期値は $x_0 = 0.0$ とした．

これらの条件の元で，データに対しフィルタを適用した結果を図 1 及び図 2 に示す．青いマーカー付き折れ線が観測値を，オレンジが状態推定値を表す．縦軸は 5 分当たりのキャプチャパケットサイズ合計バイト数，横軸は観測時刻である．図 1 は，式 (3) での通常のノイズの割合を表すパラメータが $\alpha = 1.0$ の場合である．時刻 00:00 や，異常値が観測された 02:00 から異常値が消失した 03:00 の区間で，観測値と推定値の乖離が発生している．一方，図 2 では $\alpha = 0.5$ であり，大きな変動を許容しているため，異常値が観測されても追従できている．

次に，2 つの結果を出した観測ベクトル及び状態ベクトルのアンサンブル集合を図 3 及び図 4 に示す．図から分かるように 2 次元平面上の点群が構成されており，TDA に適するデータである．ただし縦軸と横軸のスケールが異なると TDA による特徴抽出精度が落ちるため，順序保存符号化によるある種の正規化を実施する．その後，TDA によるパーシステント図（PD）を作成した結果が図 5 及び図 6 である．PD において，まず y 軸上の点は主に孤立点の情報を表す．孤立点の情報から，図 5 では y=35 に特

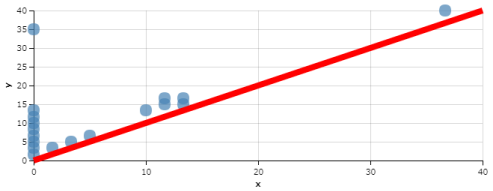

図 5: アンサンブル集合の PD($\alpha = 1.0$)

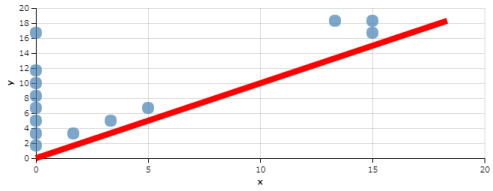

図 6: アンサンブル集合の PD($\alpha = 0.5$)

徴的な点があり，点群の中に外れ値が存在することを示している．一方で図 6 では y 軸上に一定間隔で点があり，異常値の観測を示唆しつつもフィルタが追従できていることを表している．次に対角線近辺にある点は，点群の中に円状の幾何特徴が存在することを表している．ただし対角線に近い点ほど本質的な特徴ではないノイズを表すとされる．図 5 では x=10 近辺に $y = x$ の対角線から距離 3 以上離れた点が存在し，小さな円構造の存在を表している．一方で図 6 では x=13 近辺に対角線から距離 5 ほど離れた点があり，大きな円構造の存在を表している．観測データが直線や変化の少ない滑らかな曲線の場合は円構造は発生し得ないので，これらの特徴量から，TDA により異常値の存在が検知できていることや，$\alpha = 0.5$ のモデルの方が適切な結果が出ていることなどが分かる．

以上の実験は Apache Spark-2.3.0 クラスタ上で実施し，プログラミング言語は主に Scala-2.11(Java-1.8) を用いた．TDA の計算には JavaPlex[†] を使った．クラスタのマシンのうち，主要な計算処理を実施するワーカノードは 9 台で，スペックは CPU は Intel Xeon X5690，メモリ 72G バイトである．

4. おわりに

本発表では，異常値が観測されることを考慮した混合分布型のシステムモデルを用いてアンサンブルフィルタを実行し，その結果を点群とみなして位相的データ解析を行うことにより異常値検出やパラメータの適切さを推定する手法を検討した．

参考文献

[1] G. Evensen, Sequential Data Assimilation With a Nonlinear Quansi-Geostrophic Model Using Monte-Carlo Methods to Forecast Error Statistics, J. Geo. Res. Atmos., Vol. 99, pp. 10143–10162, 1994.

[2] G. Kitagawa, A Monte Carlo Filtering and Smoothing Method for Non-Gaussian Nonlinear State Space Models, Proc. 2nd U.S.-Japan Joint Sem. Stat. Time Series Analysis, pp. 110-131, 1993.

[3] S. Reich, A Nonparametric Ensemble Transform Method for Bayesian Inference, SIAM J. Sci. Comp., Vol. 35, No. 4, pp. A2013-A2024, 2013.

[4] 佐藤哲，長期的な外れ値データ列を含む時系列データのクラスタリング手法の検討，第 16 回情報科学技術フォーラム，CF-003，2018.

[†]http://appliedtopology.github.io/javaplex/