ウェーブレット変換によるデータ圧縮の最適モデル*

松 山 佐 和¹

Suitable Model for Data Compression on Wavelets Transform

Sawa MATSUYAMA

ABSTRACT

A suitable selection of base functions for discrete wavelets transform is a difficult problem for data analysis. The grouping frequencies derived from the analysis are often dependent on the base functions and a selection of the base functions has been made mostly by trial and error. On the selection of the most suitable model from several statistical models, the model of which AIC (Akaike Information Criterion) is minimum, is selected as the suitable model. In this study, we propose the application of the AIC as the methodology for selecting the most suitable model for data compression on wavelets transform, that is, the compressions of one and two dimensional data are made by wavelets transform, and we select the minimum AIC model as the suitable one. The same method is applied to the analysis of the global wind, and we see that it can be also valid for an evaluation of the efficiency of data compression. The suitable selection of the base functions which reflects a characteristic of the data is constructed by the relation between the order of the base functions and the suitable model selected by AIC.

Keywords : Akaike information criterion, Wavelets transform, Multi resolution analysis

1. はじめに

ウェーブレット変換は周波数解析の手法として注目されている比較的新しいデータ解析法である。ウェーブレット変換は時間または位置とともに振動数や振幅が変化するようなデータの解析において、時間（位置）・周波数空間のスペクトラム分布を示すことができる¹⁾。また、離散値系ウェーブレット変換は、ウェーブレットスペクトラムのマザーウェーブレット近傍要素にデータの特徴を集約する特徴を持ち、これを利用したデータ圧縮、および、多重解像度解析による時間・周波数領域にまたがった解析が可能である²⁾。そのため、工学・理学の分野で広汎な応用が期待され、多くの研究が発表されている。特に可視化の分野では多重解像度解析を用いた噴流画像の解析等に使用され成果を挙げている³⁾。しかし、ウェーブレット変換は採用する基底関数によって抽出されるグルーピング周波数が異なるため、解析対象のデー

タの特性に適した基底関数の選択が難しく、現時点では、基底関数の選択は試行錯誤の段階である事は否めない。基底関数の選択法としては、齊藤らがデータ圧縮の圧縮率の評価法として、データ圧縮後のデータと圧縮前の元のデータとの相関係数を用いることを提案し、相関係数により最適な基底関数および圧縮率を決定している^{2),4)}。

ウェーブレット変換によるデータの圧縮は、データの特徴がウェーブレットスペクトラムのマザーウェーブレット近傍に集中することを利用して、その近傍のスペクトラムの要素のみ残し、他をゼロとすることで行われる⁵⁾。本論文では、データ圧縮の最適モデルの選択法として AIC (Akaike Information Criterion) の導入を提案する。データ圧縮のモデルを統計モデルと考え、各モデルの AIC を求め AIC が最小となるモデルを最適モデルとする手法である。圧縮後のデータと元のデータとの相関や 2 乗誤差平均による圧縮率の決定は閾値の判断が難しいが、この AIC モデルにより圧縮率決定の判断基準を示すことができる。ウェーブレット変換においても AIC の値が最小となるモデルを示すことができ、データ圧縮の圧縮率の評価に有用であることを示す。ここでは、選択したある基底関数における AIC モデルを示しているが、

* 原稿受付 2004 年 2 月 9 日

¹ 正会員 法政大学計算科学研究センター (〒184-8584 東京都小金井市 梶野町3-7-2)
お茶の水女子大学大学院人間文化研究科

今後、基底関数の次数と AIC による最適モデルとの関連を示すことにより、データの特徴をより顕著に抽出する最適な基底関数の選択法の確立に役だてると考えている。

2. 赤池情報量規準 (AIC)

AIC (Akaike Information Criterion) とはデータを説明する統計モデルが複数あれば、それらのどのモデルが最適なモデルであるかを選択する規準として赤池により導入された統計量である。E をモデルの最大対数尤度、 m をモデルの自由パラメータ数として、

$$AIC = -2E + 2m \quad (1)$$

与えられる。これは、「平均 2 乗誤差」、「尤度」の延長線上にある評価基準で、モデルの真の分布に対する期待平均対数尤度の推定量にあたる。したがって、AIC の値が小さいほど良いモデルであると考えられる⁶⁾。

我々が本論文で扱うモデルはいわゆる回帰モデルである。以下に回帰モデルにおいて AIC が一般にどのように計算されるかを述べよう。

等間隔で観測されたデータ

$$\{x_1, x_2, \dots, x_n\} \quad (2)$$

与えられたとする。このデータを $x(i) = x_i$ のように離散格子点で値を取る関数 x と考える。関数 φ_j を関数 x と同じ範囲の離散格子点で値を取る関数とする。この関数 φ_j の族を基底関数とし、関数 x をそれらの一次結合で最小 2 乗近似することを考える。すなわち、

$$x(i) = \alpha_1 \varphi_1(i) + \alpha_2 \varphi_2(i) + \dots + \alpha_m \varphi_m(i) + \varepsilon_i \quad (i=1, 2, \dots, n) \quad (3)$$

である。このとき ε_i は独立に正規分布 $N(0, \sigma^2)$ に従う確率変数と考えるモデルが、我々の統計モデルである。各回帰係数 α_j の最小 2 乗推定値を $\hat{\alpha}_j$ とすると、2 乗誤差平均は

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \{x(i) - (\hat{\alpha}_1 \varphi_1(i) + \hat{\alpha}_2 \varphi_2(i) + \dots + \hat{\alpha}_m \varphi_m(i))\}^2 \quad (4)$$

となり、これは誤差の分布が正規分布なので、誤差分散 σ^2 の最尤推定値である。このモデルの最大対数尤度は

$$E = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \hat{\sigma}^2 - \frac{n}{2} \quad (5)$$

となる。またモデルの自由パラメータ数は m 個の回帰係数と誤差の分散 σ^2 で $m+1$ 個になる。したがってこのモデルの AIC は

$$AIC = n \log 2\pi + n \log \hat{\sigma}^2 + n + 2(m+1) \quad (6)$$

与えられる。

さらに、モデルの違いは基底関数達の選び方で与えられることになる。AIC の式に含まれる $n \log 2\pi + n + 2$ はモデルの違いに依存しないので、結果的に AIC が最小になるモデル、すなわち最適なモデルは

$$n \log \hat{\sigma}^2 + 2m \quad (7)$$

が最小となるモデルを選択することになる。

一般に、1 次元の離散ウェーブレット変換は、1 次変換

$$s = Wx \quad (8)$$

で表される。ここで、 x は n 次の列ベクトル、 s はウェーブレット変換スペクトラム、 W は $n \times n$ のウェーブレット変換マトリックスである。 x はウェーブレット逆変換により、

$$x = W^T s \quad (9)$$

となる。ここでは直交基底関数を用いているため、

$$W^T = W^{-1} \quad (10)$$

である。ここで、ウェーブレット変換マトリックスを

$$W = (\varphi_1, \varphi_2, \dots, \varphi_n)^T \quad (11)$$

と書き直すと、ウェーブレット逆変換の式(9)は、

$$x = s_1 \varphi_1 + s_2 \varphi_2 + \dots + s_n \varphi_n \quad (12)$$

となる。ただし、 s_i はベクトル $s = (s_1, s_2, \dots, s_n)^T$ の各成分である。ここで、我々は基底関数 φ_j の数を減らすことでデータ圧縮を行うことにする。したがって、

$$x(i) = s_1 \varphi_1(i) + s_2 \varphi_2(i) + \dots + s_m \varphi_m(i) + \varepsilon_i \quad (i=1, 2, \dots, n) \quad (13)$$

がデータ圧縮の統計モデルとなる。

ウェーブレット変換の基底関数 $\varphi_1, \varphi_2, \dots, \varphi_n$ は、多重解像度解析のレベルにグルーピングされているため、AIC はスペクトラム数 m をレベルごとに増加させて求める。AIC を最小とするスペクトラム数がウェーブレット変換によるデータ圧縮の最適な圧縮率を与える。

3. モデルデータのウェーブレット変換

モデルデータをウェーブレット変換し、逆変換してデータを復元する。復元時に使用する多重解像度解析のレベル数 (スペクトラム数) による AIC の変化を求め、データの復元時に最適なレベル数を調べる。モデルデータには $\sin$ 関数の和で表わされる 1 次元データと 2 次元データを使用する。

3.1 1 次元モデルデータ

まず、1 次元モデルデータを $\sin$ 関数の和で表わされるものとし、データの個数を n としたときの i 番目のデータを

$$x_i = \sin\left(\frac{3\pi}{n}i\right) + 2\sin\left(\frac{7\pi}{n}i\right), \quad i=1, 2, \dots, n \quad (14)$$

とする。また、 n 個のノイズデータを正規分布 $N(0, 1)$ に従う乱数で作成し、式(14)との和をノイズを含むモデルデータとする。Fig. 1 にモデルデータを示す。ここでは

Fig. 1(a) Model data calculated by Eq. (14)

Fig. 1(b) Random noise data for model data

Fig. 1(c) Model noisy data

Fig. 1 Model data.

データ数 $n = 128$ とする。(a)は式(14)により計算されたモデルデータ, (b)は正規分布乱数で作られたノイズデータ, (c)は(a)と(b)の和でノイズを含むモデルデータである。

Fig. 1(c) のノイズを含むモデルデータをウェーブレット変換する。ウェーブレットスペクトラムをウェーブレット変換による多重解像度解析の各レベルまでに圧縮し逆変換して AIC を求める。Daubechies の異なる次数の基底関数を使用して AIC を求めた結果のうち, AIC の最小値とその次のレベルの値との変化の割合が大きなものを選び **Fig. 2** に示す。**Fig. 2** は基底関数に Daubechies の 4 次基底を使用した場合である。以下, 本論文で示す AIC の図は同様の理由で選択したものである。

各レベル 0 から 5 までの逆変換で使用されたスペクトラム数は 2, 4, 8, 16, 32, 64 である。**Fig. 2** に示されるように AIC の最小値はレベル 3 である。AIC が最小であることは, そのモデルの復元に必要なパラメータ数はレベル 3 までであり, それ以上のパラメータは必要ないこと, すなわち全体の 8 分の 1 のスペクトラムを逆変換

Fig. 2 AIC calculated by the 4th-order Daubechies base function on model data shown in Fig. 1 (c).

Fig. 3 Model noisy data recovered from level 0 to 3.

Fig. 4 Model noisy data recovered from level 0 to 4.

して得られたデータでモデルデータを十分に復元できることを示している。レベル 0 からレベル 3 で復元されるデータを **Fig. 3** に, レベル 0 からレベル 4 で復元されるデータを **Fig. 4** に示す。**Fig. 1(c)** のモデルデータは **Fig. 3** に示されるデータの情報量で十分であり, **Fig. 4** のデータでは冗長であることを示している。

Fig. 1(c) に示されるモデルデータの S/N 比は 2.26 である。S/N 比は信号データとノイズの分散の比で,

$$\frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|^2 / \sigma^2 \quad (15)$$

で与えられる。ここで n はデータの個数, $\bar{x}$ は x の平均, σ^2 はノイズの分散である。ここで S/N 比 0.56, 1.23, 2.26, 4.56, 9.03 について AIC を計算した結果を **Fig. 5** に示す。

S/N 比が, 大きい場合, すなわちノイズがデータに比べて小さい場合には, AIC の値はレベル 3 から 5 につ

Fig. 5 AIC on several signal to noise ratios.

あまり大きくならない。この結果はノイズとデータとの判別が難しくなることによると思われる。逆にノイズが大きくなると、レベル0から3にかけて AIC の値は小さくなり、レベル3から5にかけては、顕著に大きくなる。この場合では、S/N 比 2.26 が、レベル3における AIC の最小値を明確に示している。

3.2 2次元モデルデータ

2次元のモデルデータを $\sin$ 関数の和で表わされるものとし、データマトリックスの大きさを $m \times n$ としたときの i 行 j 列の要素を

$$x_{ij} = \sin\left(\frac{3\pi}{m}i\right) + 2\sin\left(\frac{7\pi}{n}j\right),$$

$$i = 1, 2, \dots, m \quad j = 1, 2, \dots, n \quad (16)$$

とする。また、 $m \times n$ 個のノイズデータを正規分布

Fig. 6(a) Model data calculated by Eq. (16)

Fig. 6(b) Random noise data for model data

Fig. 6(c) Two dimensional model noisy data

Fig. 6 Two dimensional model data.

$N(0,1)$ に従う乱数で作成し、式(16)との和をノイズを含むモデルデータとする。Fig. 6 に 2次元のモデルデータを示す。ここではデータ数を $m = 64$, $n = 128$ とする。(a)は式(16)によるモデルデータ、(b)は正規分布乱数で作られたノイズデータ、(c)は(a)と(b)の和でノイズを含むモデルデータである。この場合の S/N 比は 2.39 である。

Fig. 6(c)のノイズを含むモデルデータを 2次元ウェーブレット変換する。ウェーブレットスペクトラムをウェーブレット変換による多重解像度解析の各レベルまでに圧縮逆変換してそのときの AIC を求める。基底関数に Daubechies の 4 次基底を使用した場合の AIC の変化を Fig. 7 に示す。

各レベル0から5までの逆変換で使用されたスペクトラム数は 4, 8, 32, 8×16 , 16×32 , 32×64 である。Fig. 7 に示されるように AIC の最小値はレベル4である。すなわち全体の 16 分の 1 のスペクトラムを逆変換して得られたデータでモデルデータを十分に復元できることを示している。

レベル0からレベル4で復元されるデータを Fig. 8 に、レベル0からレベル5で復元されるデータを Fig. 9 に示す。

Fig. 7 AIC calculated by the 4th-order Daubechies base function on model data shown in Fig. 6 (c).

Fig. 8 Two dimensional model noisy data recovered from level 0 to 4.

Fig. 9 Two dimensional model noisy data recovered from level 0 to 5.

Fig. 7 に示される AIC の値によれば, Fig. 6(c) のモデルデータは Fig. 8 に示されるデータで十分表されており, Fig. 9 ではノイズが含まれ冗長であることを示している。

4. 風データへの応用

ここでは 250 hPa 高度の風データ (NCEP/NCAR 再解析データ月平均, $2.5^\circ \times 2.5^\circ$) を使用して, ウェーブレット変換を用いた圧縮について AIC で評価する. このデータは東西成分, 南北成分の 2 成分を持つベクトルデータである. Fig. 10 は 1993 年 7 月 (冷夏) の例である. このデータの特徴はジェット気流が例年に比べて南寄り (日本上空の北緯 $37^\circ \sim 47^\circ$ 付近) である. 例年での日本上空での夏のジェット気流の風速は約 30m/sec. であり, 冬は 50 ~ 60m/sec. である. データ数を 64×128 個とするためウェーブレット変換の対象とする範囲を南緯 75° から北緯 82.5° , 東経 0° から西経 42.5° とする.

Fig. 10 Monthly mean wind data on July 1993.

4.1 1次元の風データ

Fig. 11 は日本の上空の東経 140° , 北緯 35° の地点での 1989 年 5 月から 1994 年 8 月の月平均風速 (64 ヶ月) を図示したものである. u は東西成分, v は南北成分である. 風速の変化が周期 1 年であることがよくわかる. また, Fig. 10 に表われているように, 南北成分に比べて東西成分が非常に大きい.

この風のベクトルデータのベクトルの各成分 u , v についてそれぞれウェーブレット変換し, 多重解像度解析の各レベルまでに圧縮してデータを復元し, 3.1 節と同様に AIC を求める. 実際の風のデータでは, 3 章に示したモデルデータの場合とは異なり, AIC の値の最小値をレベル 0 からスペクトル数 2 分の 1 のレベルまでに見つけら

Fig. 11 Wind Data (140E, 35N).

Fig. 12 AIC calculated by the 8th-order Daubechies base function on wind data shown in Fig. 11.

Fig. 13 Wind Data (140E, 35N) recovered from level 0 to 3.

れないため, データ数を 4 分の 3 に圧縮した場合の AIC を計算し, 2 分の 1 のレベルが最小値であることを確かめている. 基底関数に Daubechies の 8 次基底を使用した場合の AIC の値を Fig. 12 に示す.

データ数 4 分の 3 のときの AIC の値をレベル 4 として示している. レベル 0 からレベル 3 で復元されるデータを Fig. 13 に示す.

Fig. 11 に示す元の風のデータと比較すると, 細かな変動が削減されていることがわかる. AIC モデルによる圧縮率の判断基準によれば, このような気象データの季節の変わり目の小さな変動は冗長な情報として切り捨てられてしまう. このモデルを使用するに当たっては, 実際のデータ解析の目的により十分な配慮が必要である.

4.2 2次元の風データ

Fig. 10 に示した 1993 年 7 月の風のデータを使用して, 平面ベクトルである風データを 3.2 節と同様にベクトル成分 u , v についてそれぞれ 2 次元ウェーブレット変換して圧縮し AIC を求める. 平面ベクトルの場合も 1 次元の風の場合と同様に, AIC の値の最小値をレベル 0 からスペクトル数 4 分の 1 のレベルまでに見つけられないため, データ数を 16 分の 9 に圧縮した場合の AIC を計算し, 4 分の 1 のレベルが最小値であることを確かめている. 基底関数に Daubechies の 4 次基底を使用した場合の AIC の値を Fig. 14 に示す. データ数 16 分の 9 の場合の AIC の値をレベル 6 として示している.

レベル 0 からレベル 5 で復元されるデータを Fig. 15 に示す. Fig. 10 に示す元の風のデータと比較してもほとんど差がなくよく復元されていることがわかる.

Fig. 14 AIC calculated by the 4th-order Daubechies base function on wind data shown in Fig.10.

Fig. 15 Monthly mean wind data on July 1993 recovered from level 0 to 5.

4.3 3次元の風のデータ

1989年5月から1994年8月の64ヶ月分の風データを使用し、4.2節の風データに時間軸を加えた3次元データとみなして、3次元ウェーブレット変換を行い、圧縮して復元しAICを求める。4.1, 4.2節と同様に、AICの値の最小値をレベル0から8分の1のレベルまでに見つけれないため、データ数を64分の27に圧縮した場合のAICを計算し、8分の1のレベルが最小値であることを確かめている。基底関数にDaubechiesの4次基底を使用した場合のAICの値をFig. 16に示す。データ数64分の27のときのAICの値をレベル6として示している。

ここでは、3次元風データを図示しないが2次元同様に元のデータをよく復元している。実データにおいてもAICによるモデルの評価が有用であることを示している。

Fig. 16 AIC calculated by the 4th-order Daubechies base function on three dimensional wind data.

5. おわりに

ウェーブレット変換によるデータ圧縮の最適モデルの選択法としてAICを導入した。1次元および2次元のモデルデータをウェーブレット変換しAICの値が最小となる最適モデルを選択できることを示した。また、風のデータにも同様の手法を適用しデータ圧縮の圧縮率の評価に有用であることを示した。本論文では、選択したある基底関数においてAICモデルが有効であることを示したが、今後、基底関数の次数とAICによる最適モデルとの関連を示すことにより、データの特徴をより顕著に抽出する最適な基底関数の選択法を確立できると考えている。

最後に風データを提供していただいた(株)ウェザーマップの松山志保氏、および本論文作成にあたり、ご討論、ご助言いただいたお茶の水女子大学の佐藤浩史教授、吉田裕亮助教授、法政大学の斉藤兆古教授にお礼申し上げます。

参考文献

- 1) 榊原進：ウェーブレットビギナーズガイド，東京電機大学出版局（1995）。
- 2) 斉藤兆古：Mathematicaによるウェーブレット変換，朝倉書店（1996）。
- 3) 木村元昭，武居昌宏，斎藤兆古，堀井清之，斎間厚：ウェーブレット多重解像度を用いた凝縮噴流画像の分離，可視化情報学会論文集，Vol. 23, No. 2（2003）pp. 9-16。
- 4) Matsuyama, S., Oguchi, Y., Saito, Y.: Applications of the Wavelet Transform to the Meteorological Vector Data, Proceedings of FEDSM'98 1998 ASME Fluids Engineering Division Summer Meeting.
- 5) 松山佐和，小口雄康，松山志保，斉藤兆古，國井利泰：ウェーブレット変換によるベクトル動画像の生成，可視化情報，Vol. 20, Suppl., No. 1（2000）pp. 145-148。
- 6) 坂元慶行，石黒真木夫，北川源四郎：情報量統計学，共立出版株式会社，情報科学講座（1993）。