EM アルゴリズムを用いた折れ線回帰モデルの推定

加茂下　茜　(指導教官：吉田 裕亮)

1　はじめに

一般に, 時系列の折れ線近似の問題では, 折れ点を含む回帰直線モデルを作り, 折れ点を推定し, モデル選択を行って最適化をおこなう手法がとられる. しかしこれは, 莫大な計算量を必要とするため, 大変手間のかかる作業といえ, あまり効率の良い推定法とはいえない.

そこで, 本研究では, 等間隔時系列データの折れ線近似を行うにあたって, EM アルゴリズムを用い, 系列の階差をクラスタリングし, AIC(情報量基準) を用い, 最適なモデル選択を行うことにより, 計算量を軽減する手法を提案する.

2　EM アルゴリズム

世の中には, 欠損値を含む不完全データが多く存在する. このような不完全データの解析に有効な統計的学習法のひとつに, EM アルゴリズムがあげられる.

2.1　混合分布問題

K 個のクラスからなる混合分布とは, j 番目のクラスの確率密度関数を $f_j(x|\theta_j)$, 混合比を $p_j(j=1,\cdots,K)$ とするとき, 確率密度関数

$$f(x)=\sum_{j=1}^{K}p_jf_j(x|\theta_j),\quad \sum_{j=1}^{K}p_j=1$$

と, 与えられるような分布である.

混合分布問題とは, 標本 $\{y_1,\cdots,y_n\}$ が与えられたとき, 各分布のパラメータ $\{\theta_1,\ldots,\theta_K\}$, 混合比率 p_j, および分布数 K を推定する問題である. 各 y_i はどのクラスに所属するかを表す変数 z_{ij} を欠損値として持つ, 不完全データである. この欠損値 $\vec{z_i}=(z_{i1},\cdots,z_{iK})$ を EM アルゴリズムを用いて補い, 各パラメータの最尤推定を行う.

本研究では, 各分布が $\theta_j=(\mu_j,\sigma_j^2)$ の場合を扱う. すなわち, 1 次元正規分布の確率密度関数が

$$f(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

と定義されているので, これを用い, 潜在変数 z_{ij} を期待値で補う. このとき, $\sum_{i=1}^{K}p_i=1$ なので, モデルのパラメータ数 k は分布数 K を用い, $k=3K-1$ と表される.

2.1.1　EM アルゴリズムと混合分布問題

まず, クラス j の確率密度関数 f_i のパラメータと, 各分布の混合比の初期値 $\mu_j^{(0)},\sigma_j^{(0)^2},p_j^{(0)}(j=1,2,\cdots,K)$ を適当に与える.

EM アルゴリズムの期待値段階

(Expectation Step)

$$\text{期待値：}\quad z_{ij}=\frac{p_j^{(m)}f_j(y_i;\mu_j^{(m)},\sigma_j^{(m)^2})}{\sum_{k=1}^{K}p_k^{(m)}f_k(y_i;\mu_k^{(m)},\sigma_k^{(m)^2})}$$

次に, 前段階で求めた期待値 z_{ij}, 観測値 y_i を用いて, 各分布のパラメータの最尤推定を行う.

EM アルゴリズムの最大化段階

(Maximization Step)

$$\text{平均：}\quad \mu_j^{(m+1)^2}=\frac{\sum_{i=1}^{n}z_{ij}y_i}{\sum_{i=1}^{n}z_{ij}}$$

$$\text{分散：}\quad \sigma_j^{(m+1)^2}=\frac{\sum_{i=1}^{n}z_{ij}y_i^2}{\sum_{i=1}^{n}z_{ij}}-(\mu_j^{(m+1)})^2$$

$$\text{混合比：}\quad p_j^{(m+1)}=\frac{1}{n}\sum_{i=1}^{n}z_{ij}$$

これらの 2 つの Step を収束するまで繰り返すことにより, 欠損値 $\vec{z_i}$ を補い, 各パラメータの最尤推定を行うことが可能となる.

3　AIC(赤池情報量基準)

AIC は, 最適なモデルを選択するひとつの基準として, 最大対数尤度 $l(\hat{\theta})$, モデルのパラメータ数 k により, 一般的に以下のような式で導入されている.

$$AIC=-2\{l(\hat{\theta})-k\}$$

この AIC の値が最小となるモデルが, より最適なモデルとして選択される. 先の述べたように, 本研究の混合分布数が K のときは, $k=3K-1$ となる.

4　提案方法

本研究では, EM アルゴリズムを用い, 等間隔時系列データの折れ線近似をするにあたって, まず, データの折れ点の有無, またその範囲を推定することを目的とする. 推定の手順は以下の通りである.

1. 等間隔時系列データに対し, その階差を求める.
2. 適当な平均 μ_j, 分散 σ_j^2, 混合比 p_j, 分布数 K の初期値を与える.
3. EM アルゴリズムの期待値段階 (Expectation Step), 最大化段階 (Maximization Step) を, 収束するまで繰り返し, それぞれの最尤推定値 $\hat{\mu_j}$, $\hat{\sigma_j^2}$, $\hat{p_j}$ を求める.
4. AIC を算出する.

5. 最適なモデルで推定を行うため, 分布数 K とそれに応じた初期値を変化させ, 手順 3, 4 を繰り返し行う.
6. AIC が最小となるモデルを最適なモデルとして採用し, 分布数 K を決定する.

（このとき, EM アルゴリズムを再度収束するまで何回か繰り返し, 期待値 z_{ij} の移動平均を適当な値で区切って算出し, その平均のデータを扱う.）

7. 期待値 z_j をそれぞれのクラスに対してグラフ表示する.
8. z_j の変化はグループに属する確率の変化に等しいので, その変動点を求めるために, それぞれのグラフを重ね合わせる.
9. 交点は, グラフが変化している点であるので, これを折れ点を含む範囲として推定する.

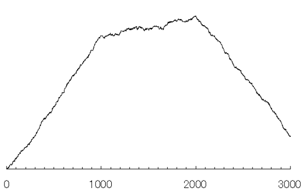
5　推定実験

5.1　シミュレーション実験

正規乱数によるノイズ $\epsilon \sim N(0,4)$ をもつ, 以下のようなシミュレーションデータを用意する.

$$\begin{cases} x_{n+1} &= x_n + 1 + \epsilon \quad (0 \le n < 1000) \\ x_{n+1} &= x_n \qquad + \epsilon \quad (1000 \le n < 2000) \\ x_{n+1} &= x_n - 1 + \epsilon \quad (2000 \le n < 3000) \end{cases}$$

これらの階差に, EM アルゴリズムを用いてクラスタリングを行う. それぞれのパラメータが収束したら, 各クラスごとに, 期待値 z_j の 100 ずつの移動平均を算出し, グラフ化し, その交点を求める.



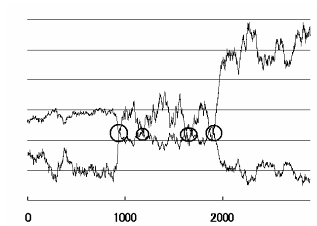


図 1: ノイズ $\epsilon \sim N(0,4)$ を入れた等間隔な時系列データ (左)
図 2: 期待値 z_{ij} の 100 ずつの移動平均の重ね合わせ (右)

5.1.1　結果

図 2(右) より, グラフの交点は 930, 1900 と読み取れる. これらは 100 ずつの移動平均の始点なので, 折れ点の範囲は 930〜1030, 1900〜2000 となり, 折れ点を含むと推定される.

5.1.2　考察

AIC によるモデル選択では, 階差による分類が行われるため, 異なるグループとして設定しても, 階差がほぼ等しいものは, ひとつのグループとしてまとめられてしまう場合がある. よって AIC で求められる分布数 K と, 最適なクラス数は必ずしも同じとは限らないと考えられる.

5.2　実データへの応用

5.2.1　為替データ

実データへの応用例として, 98.10.28〜09.1.30 の JPY-USD 相場の変動データ (図 3) を本研究で提案方法で, 折れ線回帰近似を与える折れ点の区間を推定する.

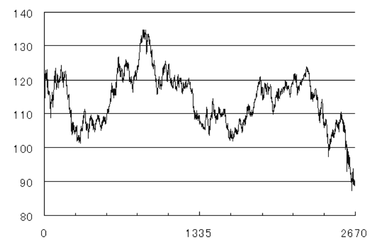
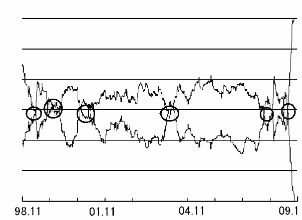


図 3: 98.10.28〜09.1.30 の JPY-USD 相場の変動データ

まず AIC を用いてモデル選択を行ったところ, $K=3$ のときに最小となり, 最適な分布数を $K=3$ と判断した. そこで, EM アルゴリズムを用いて求まった, いくつかのクラスに対する z_{ij} の平均値の移動平均をグラフ化し, 重ね合わせ (図 4), 折れ点の範囲を推定した (図 5).



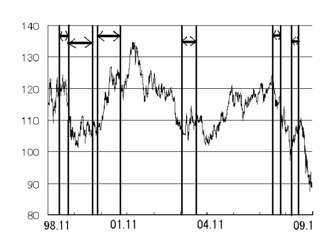


図 4: JPY-USD 相場の z_{ij} の移動平均の重ね合わせ (左)
図 5: 為替グラフの大まかな折れ点範囲 (右)

5.2.2　結果

図 5 の結果より, 折れ点を含むであろう範囲が推定できた. 実際のデータと目視して比較してみても, 折れ点を含むであろうことが予測される.

6　まとめ

等間隔時系列データの折れ線近似問題において, 本研究で提案した手法を用いて, 折れ点を含む範囲を推定することができた. しかし, 推定された範囲が広いので, さらに狭い範囲を推定する場合にも, 本手法が適用可能なのか, 今後の課題として検討したい.

参考文献

1. 坂元慶行, 石黒真木夫, 北川源四郎, 情報量統計学, 共立出版 (1983)