社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－MPS－67 (4)
2007－BIO－11 (4)
2007／12／20

外れ値データの発生を含む回帰モデルに対するベイズ予測アルゴリズム

須子 統太 †，松嶋 敏泰 †，平澤 茂一 ††

† 早稲田大学基幹理工学部 †† 早稲田大学創造理工学部

統計解析を行う際，得られたデータの中に外れ値が含まれることが多々ある．外れ値は少量であっても解析結果に大きく影響を与えることがあるため，従来から外れ値の取り扱いについて広く研究が行われている．従来，Box らにより線形回帰モデルに対し混合分布を用いて外れ値の発生をモデル化する研究が行われている．同様のモデルに対し様々な研究が行われているが，いずれも外れ値の検出やパラメータの推定を目的としている．そこで本研究では，外れ値データの発生を含む回帰モデルに対する予測法について扱う．まず，このモデルに対しベイズ基準のもとで最適な予測法を示す．次に EM アルゴリズムを用いて計算量を削減した近似アルゴリズムを提案し，シミュレーションにより有効性を検証する．

A Bayes Prediction Algorithm for Regression models with Outliers

Tota SUKO † Toshiyasu MATSUSHIMA † Shigeichi HIRASAWA ††

† School of Fundamental Science and Engineering, Waseda University

†† School of Creative Science and Engineering, Waseda University

Outliers are often included in statistical data. The statistics analysis result is influenced from outliers. Therefore, there are many researches for handling of outliers. Box modeled outliers using mixture distribution. There are many researches that aim parameter estimation or outlier detection about this model. In this paper, we treat prediction problem about this model. First, we present an optimal prediction method with reference to the Bayes criterion in this model. The computational complexity of this method grows exponentially. Next, we propose an approximation algorithm reducing the computational complexity using EM algorithm , and evaluate this algorithm through some simulations.

1 はじめに

統計解析を行う際，得られたデータの中に外れ値が含まれることが多々ある．外れ値の取り扱いについては主に，外れ値の発生に確率モデルを仮定する場合と仮定しない場合とにわけることができる．本研究では前者の外れ値の発生に確率モデルを仮定する場合を扱う．

従来，線形回帰モデルに対し混合分布を用いて外れ値の発生をモデル化する研究が行われている．Box らは，正常値の発生する分布と外れ値の発生する分布の混合分布を用いることで外れ値の発生をモデル化した．[1] 同様のモデルに対し様々な研究が行われているが，いずれの研究も外れ値の検出やパラメータの推定を目的としている．[2, 3, 4] そこで本研究では，外れ値データの発生を含む回帰モデルに対して，外れ値の検出ではなく予測を解析の目的とする．この場合，外れ値を検出し取り除いたデータを用いて予測を行う事が必ずしも精度の良い予測方法になるとは限らない．

本研究ではまず，外れ値データの発生を含む回帰モデルに対し，ベイズ基準のもとで最適な予測法を示す．しかし，この予測法はデータ数に対して指数的に計算量が増えてしまうという問題点がある．そこで，本研究ではデータ数が多い場合にも適用可能な計算量を削減した近似アルゴリズムを提案する．また，近似アルゴリズムの有効性についてシミュレーションにより検証する．

2 混合分布による外れ値データのモデル化 [1, 2]

Box らは混合分布を用いることで外れ値データの発生を含む回帰モデルを提案した．[1] 今，i 番目のデータの説明変数を $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip-1})^t$，目的変数を y_i とする．但し，t は行列の転置を表す事とし，$x_{ij}, y_i \in \mathcal{R}$ とする．また，β_0, σ_0^2 をそれぞれ正常値の従う p 次元回帰係数ベクトルおよび分散パラ

メータ，β_1, σ_1^2 をそれぞれ外れ値の従う p 次元回帰係数ベクトルおよび分散パラメータとする．さらに，外れ値の発生する確率を α とした時，y_i は確率 $1-\alpha$ で正常値の分布から発生し，確率 α で外れ値の分布から発生すると仮定する．以降，正常値の分布のパラメータを $\boldsymbol{\theta}_0 = (\beta_0, \sigma_0^2) \in \Theta_0$，外れ値の分布のパラメータを $\boldsymbol{\theta}_1 = (\beta_1, \sigma_1^2) \in \Theta_1$ で表記し，全体の分布のパラメータを $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma_0^2, \sigma_1^2) \in \Theta$ と表記する．この時，y_i の確率分布を以下で定義する．

$$
\begin{aligned}
& p(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) \\
&= (1-\alpha) p_0(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}_0) + \alpha p_1(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}_1) \\
&= (1-\alpha) \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{ -\frac{1}{2\sigma_0^2}(y_i - \boldsymbol{x}_i^t \beta_0)^2 \right\} \\
&+ \alpha \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left\{ -\frac{1}{2\sigma_1^2}(y_i - \boldsymbol{x}_i^t \beta_1)^2 \right\}. \qquad (1)
\end{aligned}
$$

また，$\boldsymbol{x}^n = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n), y^n = (y_1, y_2, \cdots, y_n)$ と表記し，各データは独立に生起するものと仮定する．上記モデルは，パラメータ $\boldsymbol{\theta}$ に制約を置くことで，外れ値データの発生する構造を表現する事ができる．本研究では外れ値データが，正常値データとは全く別の母集団から発生した場合を考え，$\boldsymbol{\theta}_0$ と $\boldsymbol{\theta}_1$ が独立である事を仮定する．また，従来研究同様，外れ値の発生する確率 α は既知，パラメータ $\boldsymbol{\theta}$ は未知として扱う．

3 ベイズ基準に基づく最適な予測法

3.1 外れ値データの発生を含む回帰モデルにおける予測問題

本研究では予測問題として，説明変数と目的変数の n 個の組 $(\boldsymbol{x}^n, y^n)$ がデータとして得られたもとで，$\boldsymbol{x}_{n+1}$ が与えられたときの y_{n+1} の予測値 $\hat{y}_{n+1}$ を求める問題を扱う．上記モデルを仮定すると，予測問題として，予測対象となる y_{n+1} が $p(y_{n+1}|\boldsymbol{x}_{n+1}, \boldsymbol{\theta})$ に従って発生する問題と，$p_0(y_{n+1}|\boldsymbol{x}_{n+1}, \boldsymbol{\theta}_0)$ から発生する問題の二種類の問題が考えられる．前者の場合，予測対象となる y_{n+1} が外れ値の分布から発生する可能性があり，外れ値も予測する必要が出てくる．そのため，多くの場合では後者の，y_{n+1} は正常値の分布からのみ出現すると仮定する事が考えられる．本件研究ではどちらの場合でも同様の議論ができるが，本稿では以下，後者の場合のみについて示す．

3.2 パラメトリックな確率モデルにおけるベイズ最適な予測 [6]

まず，予測に対する損失関数を定義する．本研究では以下の二乗誤差損失を用いる．

$$
Loss = (y_{n+1} - \hat{y}_{n+1})^2. \qquad (2)
$$

この損失関数にデータの出現確率とパラメータの事前分布 $w(\boldsymbol{\theta})$ で期待値をとったベイズリスクを考える．ベイズ基準のもとで最適な予測（以下，ベイズ最適な予測と呼ぶ）とは，このベイズリスクを最小にする予測を行うことである．二乗誤差損失を仮定したもとでのベイズ最適な予測は次式で得られる．

$$
\begin{aligned}
& \hat{y}^*_{n+1} = \\
& \int_{\mathcal{R}^n} y_{n+1} \times \\
& \int_{\Theta_0} p_0(y_{n+1}|\boldsymbol{x}_{n+1}, \boldsymbol{\theta}_0) w_0(\boldsymbol{\theta}_0|\boldsymbol{x}^n, y^n) d\boldsymbol{\theta}_0 dy_{n+1}.
\end{aligned} \qquad (3)
$$

但し，$w_0(\boldsymbol{\theta}_0|\boldsymbol{x}^n, y^n)$ は $\boldsymbol{\theta}_0$ の事後分布とする．この時，$\int_\Theta p(y_{n+1}|\boldsymbol{x}_{n+1}, \boldsymbol{\theta}) w(\boldsymbol{\theta}|\boldsymbol{x}^n, y^n) d\boldsymbol{\theta}$，を予測分布と呼ぶ．つまり，ベイズ最適な予測は予測分布の期待値を求めることで得られる．

予測分布の計算にはパラメータ空間上での積分計算が必要となり，一般には解析的に解くことができない．しかし，y_i の分布が指数型の分布の時，パラメータの事前分布に自然共役な事前分布が存在し，予測分布を解析的に求めることができることが知られている．[5] 前述の外れ値データを含む回帰モデルは指数型の分布ではないので，自然共役な事前分布を構成できなが，Box らによって提案されたパラメータの事後分布を求める手法 [1] を応用することで，解析的な予測分布の計算法が簡単に導出できる．

3.3 ベイズ最適な予測法の導出

まず，隠れ変数 z_i を導入する．z_i は i 番目のデータが正常値であれば 0，外れ値であれば 1 をとる変数である．また，$z^n = (z_1, z_2, \cdots, z_n) \in Z^n$ とすると，z^n は n 個のデータの中でどのデータが外れ値であるかの出現パターンを表している．この z^n を用いる事で，(3) 式は次式で計算される．

$$
\begin{aligned}
& \hat{y}^*_{n+1} \\
&= \sum_{z^n \in Z^n} q(z^n|\boldsymbol{x}^n, y^n) \int_{\mathcal{R}^n} y_{n+1} \int_{\Theta_0} \\
&\times p_0(y_{n+1}|\boldsymbol{x}_{n+1}, \boldsymbol{\theta}_0) w_0(\boldsymbol{\theta}_0|\boldsymbol{x}^{\Gamma_0(z^n)}, y^{\Gamma_0(z^n)}) d\boldsymbol{\theta}_0 dy_{n+1}.
\end{aligned} \qquad (4)
$$

但し，$\Gamma_0(z^n) = \{i | z_i = 0, i = 1, 2, \cdots, n\}$，$y^{\Gamma_0(z^n)} = (y_i | i \in \Gamma_0(z^n))$，同様に $\Gamma_1(z^n) = \{i | z_i = 1, i = 1, 2, \cdots, n\}$，$y^{\Gamma_1(z^n)} = (y_i | i \in \Gamma_1(z^n))$，とする．

この時，$\boldsymbol{\theta}_0$ の事前分布に，通常の線形回帰モデル

に対する自然共役事前分布，

$$w_0(\boldsymbol{\theta}_0) \propto (\sigma_0^2)^{-\frac{\nu_0'}{2}} \exp\{-\tfrac{1}{2}[\lambda_0' + (\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0')^t C_0'(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0')]\}. \quad (5)$$

を仮定すると，
$\int_{\Theta_0} p_0(y_{n+1}|\boldsymbol{x}_{n+1}, \boldsymbol{\theta}_0) w_0(\boldsymbol{\theta}_0|\boldsymbol{x}^{\Gamma_0(z^n)}, y^{\Gamma_0(z^n)}) d\boldsymbol{\theta}_0$，
はt分布となり解析的に計算できる．また，z^n の事後確率 $q(z^n|\boldsymbol{x}^n, y^n)$ も解析的に求める事ができる．但し，$q(z^n)$ は以下で定義される事とする．

$$q(z^n) = (1-\alpha)^{|\Gamma_0(z^n)|}\, \alpha^{|\Gamma_1(z^n)|}. \quad (6)$$

つまり，ベイズ最適な予測値は (5) 式の事前分布を仮定する事で，t分布の期待値を全ての $z^n \in Z^n$ の事後確率で重み付ける事で得られることが分かる．

4 計算量を削減した近似アルゴリズム

(4) 式において，$q(z^n|\boldsymbol{x}^n, y^n)$ での重み付け和計算部分が計算量の主要項となる．z^n のとり得る値は全部で $|Z^n| = 2^n$ 個あるため，和計算の回数も 2^n 回必要となる．そのため，(4) 式の計算は n に対し指数オーダの計算量が必要となる．そこで次に，計算量を削減した近似予測アルゴリズムを提案する．

和計算を減らす方法として，重み付ける z^n の集合を，Z^n からそれより小さい何らかの集合に制限するという方法が考えられる．最も単純な方法としては，$q(z^n|\boldsymbol{x}^n, y^n)$ の高い順にいくつかの z^n だけを重み付けて近似する方法が考えられる．しかし，$q(z^n|\boldsymbol{x}^n, y^n)$ を求めるには，結局のところ 2^n 回の和計算を行わなくてはならず計算量を削減できない．そこでEMアルゴリズムを用いて，$q(z^n|\boldsymbol{x}^n, y^n)$ が高い z^n の集合を近似的に求める方法を提案する．EMアルゴリズムより求められたパラメータの推定値 $\hat{\boldsymbol{\theta}}$ を利用し，$q(z_i|\boldsymbol{x}_i, y_i, \hat{\boldsymbol{\theta}})$ という分布を計算する．これは，パラメータの推定値が与えられたもとで，i 番目のデータが外れ値であるか，正常値であるかを表した分布になっており，良いパラメータの推定値が与えられれば，$q(z^n|\boldsymbol{x}^n, y^n)$ の高い z^n を見つける指標になると考えられる．また，$q(z_i|\boldsymbol{x}_i, y_i, \hat{\boldsymbol{\theta}})$ は，EMアルゴリズムの反復計算の中で計算され，簡単に求める事ができる．

近似アルゴリズム

step1: EMアルゴリズムを用いて $q(z_i|\boldsymbol{x}_i, y_i, \hat{\boldsymbol{\theta}})$ を求める．

step2: $\hat{z}_i$ を $i = 1, 2, \cdots, n$ について次式で求め，

$$\hat{z}_i = \begin{cases} 0 & q(z_i = 0|\boldsymbol{x}_i, y_i, \hat{\boldsymbol{\theta}}) > 0.5, \\ 1 & otherwise, \end{cases} \quad (7)$$

確信度 $r_i = |0.5 - q(\hat{z}_i|\boldsymbol{x}_i, y_i, \hat{\boldsymbol{\theta}})|$ を計算する．

step3: $i = 1, 2, \cdots, n$ について，r_i が何番目に小さいかを表す関数を $\eta(r_i)$ とし，$\Omega^A = \{i|\eta(r_i) \leq A, i = 1, 2, \cdots, n\}$ とする．この時，以下の集合を求める．

$$\tilde{Z}^n(A) = \{(\tilde{z}_1, \tilde{z}_2, \cdots, \tilde{z}_n)|\tilde{z}_i \in \tilde{Z}_i(A), i = 1, 2, \cdots, n\}. \quad (8)$$

但し，

$$\tilde{Z}_i(A) = \begin{cases} \{0,1\} & i \in \Omega^A, \\ \{\hat{z}_i\} & otherwise. \end{cases} \quad (9)$$

step4: 次式で予測値を計算．

$$\tilde{y}_{n+1} = \sum_{\tilde{z}^n \in \tilde{Z}^n(A)} \tilde{q}(\tilde{z}^n|\boldsymbol{x}^n, y^n) \int_{\mathcal{R}^n} y_{n+1} \int_{\Theta_0} \times p_0(y_{n+1}|\boldsymbol{x}_{n+1}, \boldsymbol{\theta}_0) w_0(\boldsymbol{\theta}|\boldsymbol{x}^{\Gamma_0(\tilde{z}^n)}, y^{\Gamma_0(\tilde{z}^n)}) d\boldsymbol{\theta}_0 dy_{n+1}. \quad (10)$$

但し，$\tilde{q}(\tilde{z}^n|\boldsymbol{x}^n, y^n)$ は，$\tilde{Z}_i(A)$ に含まれる z^n についてのみ事後確率を計算し正規化した値である．□

上記近似アルゴリズムは $\tilde{Z}^n(A)$ を求める所のみにEMアルゴリズムを用いており，予測値自体は $\hat{\boldsymbol{\theta}}$ を使わず計算している．また，ベイズ最適な予測値との違いは，重み付ける z^n の集合を Z^n から $\tilde{Z}^n(A)$ に制限している部分である．$|\tilde{Z}^n(A)| = 2^A$ となるので，近似アルゴリズムでは 2^A 個の z^n を重み付けている事になる．そのため，仮に $A = n$ とすると，近似アルゴリズムによる予測値とベイズ最適な予測値は一致する．

5 シミュレーションによる評価

近似アルゴリズムの性能をシミュレーションにより評価した．

データ数が少ない場合には，事後確率 $q(z^n|\boldsymbol{x}^n, y^n)$ を正確に計算する事ができる．そこで，$q(z^n|\boldsymbol{x}^n, y^n)$ の高い順に z^n の重み付け数を増やしていった場合と，近似アルゴリズムにより $A = 0, 1, 2, \cdots$ と増やしていった場合の予測誤差を fig.1 に示す．また，データ数が増えた場合，データ数の変化による近似アルゴリズムの予測誤差について fig.2 に示す．

データは $\boldsymbol{\theta}$ を (5) 式に従いランダムに発生させたもとで，$\alpha = 0.1$ とし $(\boldsymbol{x}^n, y^n)$ と $(\boldsymbol{x}_{n+1}, y_{n+1})$ を発生させた．実験は30000回行い，その平均値を示した．また比較のため，外れ値が全て正確に分かってい

たもとでのベイズ予測の二乗誤差と，EM アルゴリズムによって求めた推定値 $\hat{\boldsymbol{\theta}}$ を用いて，$\hat{y} = \boldsymbol{x}_{n+1}^t \hat{\boldsymbol{\beta}}_0$ と予測した場合の二乗誤差を載せた．

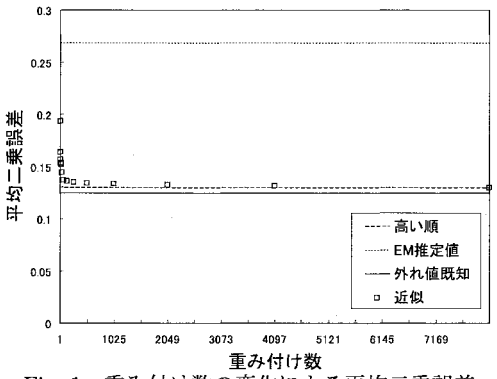

Fig. 1　重み付け数の変化による平均二乗誤差

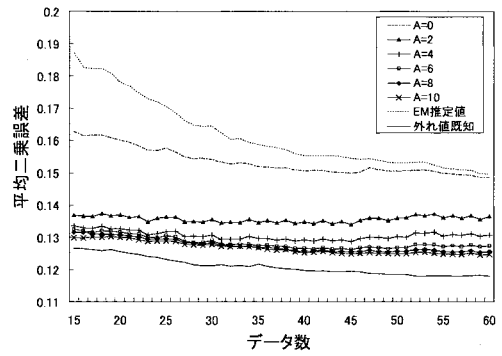

Fig. 2　データ数の変化による平均二乗誤差

6　考察

図 1, の結果から正確な事後確率の高い順番で重み付けを行う場合，重み付け数が非常に少ない段階で平均二乗誤差が収束し，ベイズ最適な予測とほぼ同じ値になる事が分かった．そのため，事後確率の大きい z^n を上手く見つけることができれば，全ての系列の重み付けをしなくても充分ベイズ最適に近い予測が可能であると考えられる．また，近似アルゴリズムは A がある一定の値になると平均二乗誤差は収束し，ベイズ最適な予測に近づいている事がわかる．この事から，A の値を上手く調整する事で近似アルゴリズムでも充分精度の高い予測が可能であると考えられる．図 2 より，近似アルゴリズムは実験した全て A について，データ数に関わらず，EM アルゴリズムの推定値を用いた予測よりも常に平均二乗誤差が少なくなっている．また，近似アルゴリズムはデータ数が少ない段階から，比較的安定して高精度の予測が可能であるのに対し，EM アルゴリズムはデータ数によって予測の精度が大きく変わる事がわかる．そのため，近似アルゴリズムはデータ数が比較的少ない時に，より有効な手法であると考えられる．

7　まとめ

本研究では，外れ値データの発生を含む回帰モデルに対し，ベイズ基準のもとで最適な予測法を示した．また，ベイズ最適な予測法はデータ数が増えると計算量が指数的に増えてしまうため，EM アルゴリズムを用いた近似アルゴリズムを提案し，その性能をシミュレーションによって示した．シミュレーションの結果，近似アルゴリズムは計算量を削減しているものの，充分に精度の高い予測が可能である事が分かった．

謝辞

本研究を行うにあたり数多くの御助言，御支援を賜りました松嶋研究室・平澤研究室の各氏に感謝致します．なお，本研究の一部は学術振興会科学研究費基盤研究 (C)18560391 の援助による．

参考文献

1) G.E.P.Box and G.C.Tiao, "A Baysian approach to some outlier problems," *Biometrika*, pp.119-129, 1968.

2) B. Abraham and G.E.P.BOX, "Linear Models and Squrious Observations," *Applied Statistics*, Vol. 27, No. 2, pp131-138, 1978.

3) D. Pena and I. Guttman, "Comparing probabilitic methods for outlier detection in linear models," *Biometrika*, vol.80, No. 3, pp.603-610, 1993.

4) J. Hoeting, A. E. Raftry and D. Madigan, "A Method for simultaneous Variable Selection and Outlier Identification in Linear Regression," *Computational Statistics and Data Analysis*, Vol. 22, pp.251-270, 1996.

5) J.M.Bernardo, A.F.M.Smith, *Bayesian theory*, John Wiley&Sons, 1994.

6) 松嶋 敏泰,"帰納・演繹推論と予測 -決定理論による学習モデル-," 第 1 回情報論的学習理論ワークショップ予稿集, pp.1-8, 1998.