1K-07

河川における水中大腸菌数予測のための符号制約回帰分析

下山　愛祐美 * 黒岩　祐有美 † 佐野　大輔 ‡ 加藤 毅 * § ¶
Ayumi Shimoyama Yumi Kuroiwa Daisuke Sano Tsuyoshi Kato

1. はじめに

環境水や飲料水の微生物学的安全性を維持するには水中病原体濃度を的確に予測しなければならない．水中病原体濃度の予測モデルを構築するには，水文水質データなどの説明変量と病原体濃度のペアのデータセットが必要になる．しかしながら，通常，水中の病原体濃度は低いことから，多くのサンプリング時において**検出限界**未満になり，このため最小二乗推定など標準的な統計推定の枠組みから逸脱する複雑な推定問題になっている．**非検出値**を含むデータセットは**左側打ち切りデータ**と呼ばれる．左側打ち切りデータに対して，線形回帰分析を行う際，**トビットモデル** (Tobit model) [2] と呼ばれる特殊な確率モデルが有効であることが知られている．トビットモデルを使用しても，訓練用データセットが小さかったり，検出されたデータが少なすぎたりすると，訓練用データに過適合して汎化性能が悪くなる．しかし，病原体濃度の計測は多くのコストがかかるため，訓練用データセットを大きくしたり，検出データを増やしたりすることは容易ではない．そこで，本研究では，水質工学において蓄積された**ドメイン知識**をトビットモデルの推定に積極的に活用して，過適合を抑制し汎化性能を高めることを目標とする．

本論文では，ドメイン知識として符号制約を利用してトビットモデルの汎化性能を向上させる．説明変量と病原体の濃度との間に相関があるとき，データが大量にあれば，一つ一つの説明変量の相関が小さくても，複数の説明変量の線形結合によって病原体濃度の予測精度は高まる．しかし，トビットモデルを推定するためのデータが少ない場合，真の相関とは符号が逆の**標本相関**になってしまうことがあり，この現象が予測精度を下げる原因となる．ところで，水質工学においてはこれまで用いられてきた説明変量のうちの大部分は病原体濃度との**母相関**の符号が自明である．本研究では，この事実に注目して，回帰係数に次のようなシンプルな制約（**符号制約**と呼ぶ）を課すこととした．

- 正の母相関を持つことが既知の説明変量の回帰係数に対して，**非負制約**を課す．
- 負の母相関を持つことが既知の説明変量の回帰係数に対して，**非正制約**を課す．

著者らは，この符号制約下でトビットモデルを最尤推定するための**期待値最大化法 (EM 法)** を開発した．本論文では，その期待値最大化法を提案し，大腸菌数と水文水質データの実データを使った数値実験によって，符号制約により汎化性能が顕著に向上することを示す．

*群馬大学大学院理工学府電子情報・数理教育プログラム
†群馬大学理工学部電子情報理工学科
‡東北大学大学院工学研究科土木工学専攻
§群馬大学次世代モビリティ社会実装研究センター (CRANTS)
¶早稲田大学規範科学総合研究所 (IIRS)

2. 符号制約最小二乗推定

本論文では左側打ち切りデータに対する回帰分析に関して議論するが，本節ではまず打ち切られていないデータに対する回帰分析に符号制約を課す方法について述べる．最小二乗推定では，訓練用データ $(\boldsymbol{x}_1, y_1)$, $\ldots$, $(\boldsymbol{x}_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ に対して，平均二乗誤差

$$P(\boldsymbol{w}) := \frac{1}{n}\sum_{i=1}^{n}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{w}\rangle)^2 \tag{1}$$

が最小になるように回帰係数 $\boldsymbol{w} := [w_1, \ldots, w_d]^\top$ を求める．ここで，$P(\boldsymbol{w})$ の最小化の際に符号制約を課すことを考える．$\mathcal{I}_+ \subseteq [n] := \{1, \ldots, n\}$ を目的変量との相関が正とわかっている説明変量の添え字の集合とし，$\mathcal{I}_- \subseteq [n]$ を目的変量との相関が負とわかっている説明変量の添え字の集合とする．$h \in \mathcal{I}_+$ なる w_h に非負制約，$h' \in \mathcal{I}_-$ なる $w_{h'}$ に非正制約を課すとき，実行可能領域は

$$\mathcal{S} := \left\{\boldsymbol{w} \in \mathbb{R}^d \,|\, \forall h \in \mathcal{I}_+, w_h \geq 0, \forall h' \in \mathcal{I}_-, w_{h'} \leq 0\right\}$$

と表すことができる．実行可能領域 $\mathcal{S}$ 内で $P(\boldsymbol{w})$ を最小化する問題は，スラック変数を導入することで等価な**非負制約最小二乗問題** (**NNLS**) [1] に変換でき，NNLS のソルバーを使うと効率的に最適解を見つけられる．

3. トビットモデルの符号制約最尤推定

トビットモデルは，非検出データが，検出限界以下になる確率で表現することで尤度関数を表現し，回帰変数 $\boldsymbol{w}$ を尤度最大になる値に決定する．いま，n 個の例題のうち，n_{v} 個の目的変量の値が検出限界 u を上回ったとし，その n_{v} 個のデータペア

$$(\boldsymbol{x}^{\mathrm{v}}_1, y^{\mathrm{v}}_1), \ldots, (\boldsymbol{x}^{\mathrm{v}}_{n_{\mathrm{v}}}, y^{\mathrm{v}}_{n_{\mathrm{v}}}) \tag{2}$$

の値はすべて得られているとする．一方，残りの $n_{\mathrm{h}}(:= n - n_{\mathrm{v}})$ 個の例題に関しては，目的変量の値が検出限界 u を下回ったために，n_{h} 個のデータペア

$$(\boldsymbol{x}^{\mathrm{h}}_1, y^{\mathrm{h}}_1), \ldots, (\boldsymbol{x}^{\mathrm{h}}_{n_{\mathrm{h}}}, y^{\mathrm{h}}_{n_{\mathrm{h}}}) \tag{3}$$

に対して，y^{h}_i の値は不明とする．このようなデータセットに対して，トビットモデルでは，回帰係数 $\boldsymbol{w}$ を対数尤度関数

$$L(\boldsymbol{w}, \beta) := \sum_{i=1}^{n_{\mathrm{v}}} \log \mathcal{N}(y^{\mathrm{v}}_i ; \langle \boldsymbol{w}, \boldsymbol{x}^{\mathrm{v}}_i \rangle, \beta^{-1}) + \sum_{i'=1}^{n_{\mathrm{h}}} \log \int_{-\infty}^{u} \mathcal{N}(y^{\mathrm{h}}_{i'} ; \langle \boldsymbol{w}, \boldsymbol{x}^{\mathrm{h}}_{i'} \rangle, \beta^{-1}) dy^{\mathrm{h}}_{i'} \tag{4}$$

を最大化するように定める．従来のトビットモデルでは，制約なしで対数尤度関数 $L(\boldsymbol{w}, \beta)$ を最大化していた．ここで，符号制約を付けて最大化することを考える．すなわち，以下の問題を解くことになる：

$$\max \quad L(\boldsymbol{w}, \beta) \qquad \text{wrt} \quad \boldsymbol{w} \in \mathcal{S}, \quad \beta \in \mathbb{R}. \tag{5}$$

本研究では，この最大化問題を解くための EM 法を新たに開発した．EM 法では，各非検出データに対して，事後分布 $q(\boldsymbol{y}^{\mathrm{h}})$ を導入して，E-step と M-step を収束するまで繰り返す方法である．

反復 t におけるパラメータの値 $(\boldsymbol{w}^{(t)}, \beta^{(t)})$ とおくと，反復 t における E-step では事後分布 $q_i(y_i^{\mathrm{h}})$ を

$$q_t^i(y_i^{\mathrm{h}}) := \begin{cases} \frac{1}{\Phi(\xi_i^{(t)})} \mathcal{N}(y_i^{\mathrm{h}}; \mu_i^{(t)}, 1/\beta^{(t)}) & \text{if} \quad y_i^{\mathrm{h}} \le u, \\ 0 & \text{if} \quad y_i^{\mathrm{h}} > u \end{cases}$$

求める．ただし，$\mu_i^{(t)}$ および $\xi_i^{(t)}$ は次のようにおいた：

$$\mu_i^{(t)} := \left\langle \boldsymbol{w}^{(t)}, \boldsymbol{x}_i \right\rangle, \quad \xi_i^{(t)} := \left(u - \mu_i^{(t)}\right) \sqrt{\beta^{(t)}}. \tag{6}$$

M-step では，次のように Q 関数を最大化させるパラメータに更新する．Q 関数は以下で定義される：

$$Q(\boldsymbol{w}, \beta, q^{(t)}) := \frac{n}{2} \log(\beta) - \frac{\beta}{2} \left(\left\| \boldsymbol{X}^\top \boldsymbol{w} - \bar{\boldsymbol{y}}^{(t)} \right\|^2 + v^{(t)} \right). \tag{7}$$

ただし，

$$\begin{aligned} \boldsymbol{X} &:= \left[\boldsymbol{x}_1^{\mathrm{v}}, \ldots, \boldsymbol{x}_{n_{\mathrm{v}}}^{\mathrm{v}}, \boldsymbol{x}_1^{\mathrm{h}}, \ldots, \boldsymbol{x}_{n_{\mathrm{h}}}^{\mathrm{h}}\right], \\ \boldsymbol{y}^{\mathrm{v}} &:= \left[y_1^{\mathrm{v}}, \ldots, y_{n_{\mathrm{v}}}^{\mathrm{v}}\right]^\top, \quad \boldsymbol{y}^{\mathrm{h}} := \left[y_1^{\mathrm{h}}, \ldots, y_{n_{\mathrm{h}}}^{\mathrm{h}}\right]^\top, \\ \bar{\boldsymbol{y}}^{(t)} &:= \left[\left(\boldsymbol{y}^{\mathrm{v}}\right)^\top, \mathbb{E}_{q^{(t)}} \left[\left(\boldsymbol{y}^{\mathrm{h}}\right)^\top \right] \right]^\top, \\ v^{(t)} &:= \mathbb{E}_{q^{(t)}} \left[\left\| \boldsymbol{y}^{\mathrm{h}} \right\|^2 \right] - \left\| \mathbb{E}_{q^{(t)}} \left[\boldsymbol{y}^{\mathrm{h}} \right] \right\|^2 \end{aligned} \tag{8}$$

であり，$\mathbb{E}_{q^{(t)}}$ は反復 t における事後分布に関して期待値を取る期待値演算子である．この Q 関数を使って，次のように $(\boldsymbol{w}, \beta)$ の値を更新する：

$$\boldsymbol{w}^{(t+1)} := \operatorname*{argmax}_{\boldsymbol{w} \in \mathcal{S}} Q(\boldsymbol{w}, \beta^{(t)}, q^{(t)}), \tag{9}$$

および

$$\begin{aligned} \beta^{(t+1)} &:= \operatorname*{argmax}_{\beta \in \mathbb{R}} Q(\boldsymbol{w}^{(t+1)}, \beta, q^{(t)}) \\ &= \frac{n}{\|\boldsymbol{X}^\top \boldsymbol{w}^{(t+1)} - \bar{\boldsymbol{y}}^{(t)}\|^2 + v^{(t)}}. \end{aligned} \tag{10}$$

(9) における $\boldsymbol{w}$ の更新は，Q 関数の定義から見てわかるように，NNLS に帰着できるので，NNLS のソルバーを使えば容易に $\boldsymbol{w}$ の値を更新できる．

4. 実験

実験条件：提案法の性能を調査するために，水文水質データと大腸菌数の実データを用いた．水文水質データは WT, pH, EC, SS, DO, BOD, TN, TP および flow rate の 9 個の説明変量からなる．このうち pH は，次のように 2 個の説明変量に変換した．

$$\mathrm{pH}_+ := \max(0, \mathrm{pH} - 7), \quad \mathrm{pH}_- := \max(0, 7 - \mathrm{pH}).$$

pH_+ は酸性の強さを表し，pH_- はアルカリ性の強さを表す．本研究では実際に，2011 年 12 月から 2013 年 4 月までの間に 96 回大腸菌数を測定した．回帰分析を行う時は，この 96 個のデータのうち，無作為に選択した n 個を訓練用データとし，残り $(96 - n)$ 個を評価に使った．検出限界を疑似的にある値 u に設定し，n 個の訓練用データのうち，u 以下のデータを非検出データとして扱った．検出限界 u には $u = 2.0$ および $u = 3.0$ を選んだ．大腸菌数の常用対数を目的変数とした．96 個の大腸菌数データのうち，26.0%が 2.0 以下であり，47.9%が 3.0 以下であった．

水質工学の分野では，WT, EC, SS, BOD, TN, TP が増加するほど，大腸菌数が増加し，pH_+, pH_-, DO, flow rate が増加するほど，大腸菌数は減少することが知られている．このことから，符号制約として，説明変量 WT, EC, SS, BOD, TN, TP に対する係数 w_h は非負に制限し，説明変量 pH_+, pH_-, DO, flow rate に対する係数 w_h は非正に制限することにした．

実験結果： 96 個から n 個を無作為選択して，訓練用データと評価用データに分けた．訓練用データから回帰係数 $\boldsymbol{w}$ の値を決定し，評価用データ $(\boldsymbol{x}_i^{\mathrm{tst}}, y_i^{\mathrm{tst}}) \in \mathbb{R}^d \times \mathbb{R}$ $(i = 1, \ldots, (96 - n))$ に対して，Root mean square deviation (RMSD) を計算した．図 1 は，これを 30 回繰り返して，30 回の平均をプロットしたものである．検出限界を $u = 2.0$ に設定したときは，訓練用データ数が $n = 40$ 以下で符号制約によって平均 RMSD が小さくなった．$u = 3.0$ にあげると，訓練用データ数が多くなっても符号制約の効果が確認できた．

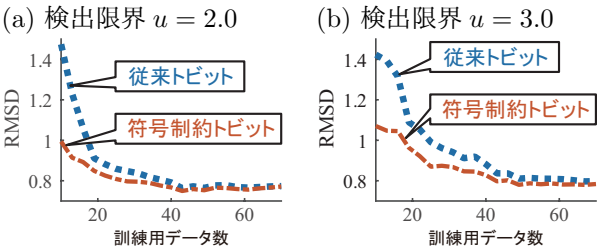

図 1: 従来最尤推定と符号制約最尤推定の比較．

5. 結論

検出限界に伴う回帰分析の方法として用いられてきたトビットモデルに符号制約を導入する方法を開発した．また，符号制約によって，訓練用データが少なくても高い汎化性能が得られることを実験により確認した．

謝辞：本研究は JSPS 科研費 40401236 の助成を受けた．

参考文献

[1] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems*. SIAM, jan 1995.

[2] J. Tobin. Estimation of relationships for limited dependent variables. *Econometrica*, 26(1):24, jan 1958.