Vol.2021-MPS-132 No.6
2021/3/1

サイド情報を活用した水中病原体と指標微生物の相関解析

曹 洪源[1]　佐野 大輔[2]　加藤 毅[1,3]

概要：ヒトの糞便に含まれる病原性微生物による水資源の汚染は公衆衛生に関する世界的課題である．水中微生物濃度の定期的モニタリングによって汚染を監視する場合，費用および技術的な制約から微生物の種類は単一または少数に限られる．病原体から暴露するリスクを適切に評価するためには，水中微生物の濃度間の相関関係を推定するための信頼できるアルゴリズムを確立する必要がある．一般に病原性微生物濃度は薄いため定量限界を下回ることが多い．そのため，観測できたデータだけから標本相関をとる方法では，相関係数の推定精度が低下しやすい．本研究では，この問題に対処するため，サイド情報を利用したトビット解析を援用して相関計算を行う相関計算法を検討した．また，ドメイン知識を活用できるよう，非対称正規分布をトビット解析に新たに導入した．数値実験により，その性能を比較調査したところ，サイド情報およびドメイン知識の活用が効果的であることを確認した．

キーワード：打ち切りデータ，トビット解析，EM法，非負最小二乗推定

HongYuan Cao[1]　Daisuke Sano[2]　Tsuyhoshi Kato[1,3]

1. 序論

ヒトの糞便に含まれる病原微生物による水資源の汚染は公衆衛生に関する世界的課題である．ヒトの疾病を引き起こす水中病原微生物には，細菌やウイルスなどが含まれる．ヒトに有害な腸内細菌として有名なものにはサルモネラ菌，赤痢菌，大腸菌 O157:H7 などがある．病原ウイルスには，腸内ウイルス，ノロウイルス，ロタウイルスなどがある．水系感染症の主要な伝搬経路は経口感染である．処理された下水にも数多くの腸内病原体が残っており，それらによって，海水，河川，湖沼，地下水などの環境水が汚染されている [5], [14]．海水中の病原体はカキなどの貝類に濃縮され，これらを未調理もしくは不十分な調理で摂取することで水系感染が発生している [6]．地表水と比べて，地下水は土層がろ過の機能を果たすため，微生物学的水質が比較的安定している．しかし，米国においては，水系感染流行事由の約半数は汚染された地下水から発生している [9]．河川，湖沼，海水における水浴など未処理のレクレーション用の水利用が引き起こした流行はしばしば糞便汚染により引き起こされている（図 1 参照）．以上の理由より，病原微生物の暴露に由来する公衆衛生リスクを制御するには，それぞれの域内での微生物学的水質の適切な評価と管理が必要になる．

すべての水中病原体を定期モニタリングすることはおよそ困難である．現在の技術では，病原体の多くは測定に高い費用や労力を要してしまう．そのため，定期モニタリングを行う病原体は少数に限定するのが現実的な方法と言える．現在は，病原性のない指標微生物や物理化学的水質データを使って，水質管理がされている．しかし，これらの指標と物理化学的水質データは病原性微生物との相関が必ずしも高くなく，あらゆる水系感染のリスクを制御するには十分ではない．一方で，水中微生物の測定技術は進化を続けている [12], [13]．将来の測定技術の発達を見越して，水中微生物濃度の解析技術を整備しておけば，水質管理ルーチンの適切な設計が可能になる．

ピアソン相関係数は，水中微生物の濃度の関係のみならず，様々な応用分野における 2 変量の関係を評価するための標準的なツールである．ある水中微生物がほかの水中微生物の濃度と高い相関の濃度を持つなら，前者の水中微生物は後者の水中微生物が及ぼす健康リスクに対する予測能力は高いといえる．相関解析の解析対象が水中微生物濃度の場合，ある困難に遭遇する．その困難は，統計解析の標準的な設定からの逸脱に起因している．それは，**定量限**

[1] 群馬大学 Gunma University
[2] 東北大学 Tohoku University
[3] 早稲田大学 Waseda University

界の存在である．多くの病原体は水中に少数しか含まれていないため，左打ち切りデータとなってしまう．左打ち切りデータには，標本中には観測されているデータもあれば，非検出となったデータもある．ただし，非検出となったデータは定量限界未満であったという情報は利用できる．適切な水質管理を行うために，このような打ち切りデータから，なるだけ正確に相関解析を行う方法の確立が望まれる．

本研究では，水中微生物濃度間の相関係数を推定するためのいくつかの方法の性能を調査した．最も簡単な方法は観測できたデータだけから標本相関をとることである．しかし，一般に病原性微生物濃度は薄いため定量限界を下回ることが多い．すると可視データは少なく，小標本からの標本相関となり，推定精度が低下しやすい．代替法として，**サイド情報**を使って非検出濃度を補完してから相関係数を計算する方法も考えられる．それは，**トビットモデル** [1] を打ち切りデータにフィットさせ，期待値で非検出値を補完するものである．すると，すべてのデータを相関係数の計算に用いることができる．このアプローチの推定精度は補完精度に依存する．本研究では，補完精度をよりよくするために，**ドメイン知識**を用いる方法を考案した．水質データにおいて，任意の２変量間の相関の符号は既知である．本研究では，ドメイン知識を活用するために，トビットモデルの回帰係数の事前分布として，非対称正規分布 [8] を用いることにした．

本研究の技術的貢献は，非対称正規事前分布を伴ったトビットモデルのフィッティングに対する高速なアルゴリズムの発見である．従来のトビットモデルのフィッティングには，しばしば，**期待値最大化法**（ＥＭ法）が用いられる．ＥＭ法の各反復はＥステップとＭステップからなる．回帰係数の事前分布が通常の正規分布の場合，Ｍステップは連立線形方程式を解くだけで実行できる．しかし，一般に事前分布を変更するとＭステップは複雑になる．本研究では，非対称正規分布を事前分布に使ってもＭステップを効率的に実行できることを理論的に証明した．

数値実験により，(i) 非対称正規事前分布を使ったモデルの計算時間と従来のモデルの計算時間はほぼ変わらないこと，また，(ii) 非対称正規事前分布によりドメイン知識がモデルの精度を改善し，推定精度を向上できること，を確認した．

2. 準備

2.1 ピアソン相関係数

ピアソン相関係数はペアデータ $(y_{1,\mathrm{a}}, y_{1,\mathrm{b}}), \ldots, (y_{n,\mathrm{a}}, y_{n,\mathrm{b}}) \in \mathbb{R} \times \mathbb{R}$ の統計量である．ピアソン相関係数の定義は

図 1 水資源と水利用．人類の生活は海，河川，湖沼，地下水といった水資源に依存しており，これらは工業，産業，水浴，飲用に利用されている．水資源はヒトの排泄によって汚染される．公衆衛生を確保するためには，利用形態に応じて適切な規制と管理が必要となる．

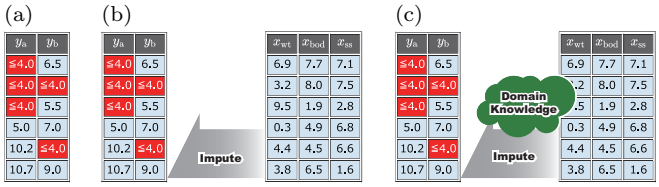

図 2 ３つの相関計算法．解析対象となる微生物濃度データは打ち切られている．(a) ナイーブ相関計算法は両方の微生物濃度が利用可能なペアのみから相関係数を計算する．(b) 従来トビット相関計算法はサイド情報を使って非検出値を補完してから相関係数を計算する．(c) 非対称トビット相関計算法はドメイン知識を使って非検出値の補完精度を向上させる．

$$R(\boldsymbol{y}_{\mathrm{a}}, \boldsymbol{y}_{\mathrm{b}}) := \frac{\sum_{i=1}^{n}(y_{i,\mathrm{a}} - \bar{y}_{\mathrm{a}})(y_{i,\mathrm{b}} - \bar{y}_{\mathrm{b}})}{\sqrt{\sum_{i=1}^{n}(y_{i,\mathrm{a}} - \bar{y}_{\mathrm{a}})^2}\sqrt{\sum_{i=1}^{n}(y_{i,\mathrm{b}} - \bar{y}_{\mathrm{b}})^2}} \tag{1}$$

で与えられる．ただし，$\boldsymbol{y}_{n,\mathrm{a}} := [y_{1,\mathrm{a}}, \ldots, y_{n,\mathrm{a}}]^\top$, $\boldsymbol{y}_{n,\mathrm{b}} := [y_{1,\mathrm{b}}, \ldots, y_{n,\mathrm{b}}]^\top$, $\bar{y}_{\mathrm{a}} := \frac{1}{n}\sum_{i=1}^{n} y_{i,\mathrm{a}}$, $\bar{y}_{\mathrm{b}} := \frac{1}{n}\sum_{i=1}^{n} y_{i,\mathrm{b}}$ とした．

2.2 トビット解析

トビット解析 [1] は打ち切りデータのための回帰分析法である．トビット解析では，**目的変量** $y \in \mathbb{R}$（本研究では，水中微生物の濃度）は次の生成モデルから生成されると仮定している：

$$y = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + \epsilon \tag{2}$$

ただし，ϵ は正規乱数 $\epsilon \sim \mathcal{N}(0, \beta^{-1})$ である；ベクトル $\boldsymbol{x} \in \mathbb{R}^d$ は**説明変量**で構成されている（本研究では，説明変量は物理化学データやほかの水中微生物の濃度となる）；$\boldsymbol{w} \in \mathbb{R}^d$ は**回帰係数ベクトル**である．この生成モデルの設定は通常の最小二乗推定と同じである．最小二乗推定との違いは，トビット解析は標本中の打ち切りの存在を許すことである．濃度 y が検出限界 θ 未満のため検出できなかったとき，その濃度の**期待値**は

$$\mathbb{E}[y | y < \theta, \boldsymbol{x}] = \langle \boldsymbol{w}, \boldsymbol{x} \rangle - \beta^{-1/2}\lambda_{\mathrm{IMR}}((\theta - \langle \boldsymbol{w}, \boldsymbol{x} \rangle)\sqrt{\beta}) \tag{3}$$

で与えられる．ただし，$\lambda_{\mathrm{IMR}}(\xi) := \phi(\xi)/\Phi(\xi)$ は**逆ミルズ比** である；ϕ および Φ は標準正規分布の密度関数と累

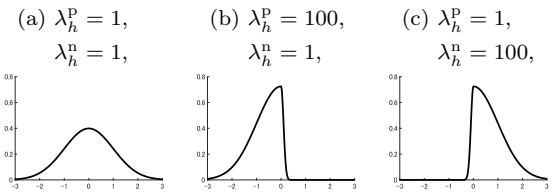

図 3 非対称トビットモデルの回帰係数の事前分布．それぞれのパネルには，横軸を回帰係数 w_h とした時の事前分布 $p(w_h)$ の値をプロットしている．(a) $\lambda_h^{\mathrm{p}} = 1$ および $\lambda_h^{\mathrm{n}} = 1$ の時，対称正規分布に戻る．(b) $\lambda_h^{\mathrm{p}} \gg \lambda_h^{\mathrm{n}}$ の時, 正の回帰係数は抑制される．(c) $\lambda_h^{\mathrm{p}} \ll \lambda_h^{\mathrm{n}}$ の時, 負の回帰係数は抑制される．

積密度関数である．目的変量 y は，条件 $y < \theta$ の下で，次の**切断正規分布**に従う：

$$p(y|y < \theta, \boldsymbol{x}) = f_{\mathrm{tn}}(y \mid \langle \boldsymbol{w}, \boldsymbol{x} \rangle, \beta, \theta) \tag{4}$$

ただし，

$$f_{\mathrm{tn}}(y \mid \mu, \beta, \theta) := \begin{cases} \frac{\sqrt{\beta}\phi(\sqrt{\beta}(y-\mu))}{\Phi(\sqrt{\beta}(\theta-\mu)} & \text{for } y \in (-\infty, \theta), \\ 0 & \text{for } y \in [\theta, +\infty). \end{cases} \tag{5}$$

この性質から，式 (3) に与える期待値は導出できる．**2次モーメント**も閉形式で次のように表すことができる：

$$\mathbb{E}[y^2|y < \theta, \boldsymbol{x}] = \frac{1 - \xi\lambda_{\mathrm{IMR}}((\theta - \langle \boldsymbol{w}, \boldsymbol{x} \rangle)\sqrt{\beta})}{\beta} + \langle \boldsymbol{w}, \boldsymbol{x} \rangle^2 - \frac{2\lambda_{\mathrm{IMR}}((\theta - \langle \boldsymbol{w}, \boldsymbol{x} \rangle)\sqrt{\beta})\langle \boldsymbol{w}, \boldsymbol{x} \rangle}{\sqrt{\beta}}. \tag{6}$$

モデルパラメータ $\boldsymbol{w}$ および β の値は，モデルを打ち切りデータセット $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ $(i = 1, \ldots, n)$ にフィットさせることで決定される．$y_1, \ldots, y_{n_{\mathrm{v}}}$ は観測された微生物濃度，$y_{n_{\mathrm{v}}+1}, \ldots, y_n$ は検出限界 θ により非検出になった濃度とする．フィッティングは次の正則化対数尤度関数の最大化によって行う：

$$L_{\mathrm{sym}}(\boldsymbol{w}, \beta) := \log p_{\mathrm{sym}}(\boldsymbol{w}) + L_0(\boldsymbol{w}, \beta) \tag{7}$$

ただし $p_{\mathrm{sym}}(\boldsymbol{w})$ は回帰係数 $\boldsymbol{w}$ の正規事前分布である：$p_{\mathrm{sym}}(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{0}, \lambda^{-1}\boldsymbol{I})$．式 (7) における第２項は，トビットモデルの対数尤度関数である：

$$L_0(\boldsymbol{w}, \beta) := \frac{n_{\mathrm{v}}}{2}\log\beta + \sum_{i=1}^{n_{\mathrm{v}}} \log\phi\left(\sqrt{\beta}(y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle)\right) + \sum_{i=n_{\mathrm{v}}+1}^{n} \log\Phi\left(\sqrt{\beta}(\theta - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle)\right). \tag{8}$$

L_{sym} の最大化にはＥＭ法がしばしば用いられる．詳細は [1] を参照されたい．

2.3 非負最小二乗問題

非負最小二乗問題は

$$\begin{aligned} &\min \quad \|\boldsymbol{A}^\top \boldsymbol{x} - \boldsymbol{b}\| \quad \text{wrt} \quad \boldsymbol{x} \in \mathbb{R}^m, \\ &\text{where} \quad \boldsymbol{A} \in \mathbb{R}^{m \times n}, \boldsymbol{b} \in \mathbb{R}^n. \end{aligned} \tag{9}$$

のように表される２次計画問題である．この問題を，以後，NNLS($\boldsymbol{A}, \boldsymbol{b}$) と表す．非負最小二乗問題を解くには Lawson & Hanson の算法 [10] を皮切りに，長年にわたって改良が重ねられ，現在では，非負最小二乗問題は高速に解くことができる凸問題の一つとして知られている [2], [4], [11].

3. 相関解析法

本研究では，相関解析法として，**ナイーブ相関計算法**，**従来トビット相関計算法**，**非対称トビット相関計算法**を調査する（図 2 参照）．

ナイーブ相関計算法：標本に n ペアのデータが含まれているとする：$(y_{1,\mathrm{a}}, y_{1,\mathrm{b}}), \ldots, (y_{n,\mathrm{a}}, y_{n,\mathrm{b}})$．これらは，２個の水中微生物の濃度で，左打ち切りである．これらの検出限界を，それぞれ，θ_{a} および θ_{b} とする．$y_{i,\mathrm{a}} < \theta_{\mathrm{a}}$ および $y_{i,\mathrm{b}} < \theta_{\mathrm{b}}$ なるデータの濃度は利用できない．ナイーブ相関計算法は，双方の微生物で利用できる濃度のみで標本相関係数を計算する．この方法の短所は，２個の微生物に共通して観測される要素は少なくなりがちな点である．そうなると，推定誤差が大きくなってしまう．

従来トビット相関計算法：筆者らは，ナイーブ相関計算法に代替する方法として，検出されなかった濃度の要素も利用する相関計算法を考えた．その方法を「従来トビット相関計算法」と呼ぶ．いま，２種類の水中微生物とは異なる，ほかの物理化学データがサイド情報として利用可能と仮定する．実際に，水温，DO，SS，TN，TP などの物理化学データは，水中微生物濃度と比べて容易に観測できる．従来トビット相関計算法では，検出できなかった水中微生物 B の濃度をサイド情報を使って補完し，次に検出できなかった水中微生物 A の濃度をサイド情報と水中微生物 B の濃度を使って補完する．非検出濃度の補完にはトビット解析（2.2 節参照）を用いる．

非対称トビット相関計算法：３番目のアプローチはドメイン知識を活用してトビット解析を高精度化し，これによって相関係数の推定精度を向上させるものである．水質工学においては，典型的な物理化学データと典型的な病原体は正の相関があるのか負の相関があるのかすでに分かっている．たとえば，水が暖かいほうが病原体微生物が増えるので，病原体と水温には正の相関があることが知られている．目的変量と正の相関がある説明変量の回帰係数は正であることが望ましい．しかし，標本によっては可視濃度値のデータ数は少なくなることがある．このような場合，本来微生物濃度と正の相関がある説明変量が負の標本相関を持ってしまうことあり，この現象がトビットモデルの能力を悪化させる．非対称トビット相関計算法はトビットモデルを改良して，非検出値の補完性能を向上させる．

本節の残りでは，非対称トビット相関計算法ではトビットモデルをどのように改良するか述べる．改良したトビットモデルを**非対称トビットモデル**と呼ぶことにする．ドメイン知識と矛盾する回帰係数を抑制するために，従来のトビットモデルで用いていた通常の正規分布 p_{sym} の代わりに，非対称正規分布 [8] (図 3 参照) を用いる．非対称正規分布を使った回帰係数の事前分布は

$$p_{\mathrm{asym}}(\boldsymbol{w}) := \prod_{h=1}^{d} \frac{1}{Z_h} \exp\left(-\frac{\lambda_h^{\mathrm{p}}(w_h)_+^2 + \lambda_h^{\mathrm{n}}(-w_h)_+^2}{2}\right) \tag{10}$$

と与えられる．ただし，$(x)_+ := \max(0, x)$，$Z_h := \sqrt{\frac{\pi}{2\lambda_h^{\mathrm{p}}}} + \sqrt{\frac{\pi}{2\lambda_h^{\mathrm{n}}}}$．目的変量と正の相関があることが既知の説明変量の添え字集合を $\mathcal{I}_{\mathrm{p}} \subseteq [d]$，目的変量と負の相関があることが既知の説明変量の添え字集合を $\mathcal{I}_{\mathrm{n}} \subseteq [d]$，と書くことにする．定数ベクトル $\boldsymbol{\lambda}^{\mathrm{p}}, \boldsymbol{\lambda}^{\mathrm{h}} \in \mathbb{R}^d$ の各要素は $\lambda_h^{\mathrm{p}} = (1 + 99\mathbb{1}[h \in \mathcal{I}_{\mathrm{n}}])\lambda$ 及び $\lambda_h^{\mathrm{n}} = (1 + 99\mathbb{1}[h \in \mathcal{I}_{\mathrm{p}}])\lambda$ $(h \in [d])$ と設定することにする．非対称トビットモデルのための正則化対数尤度関数は

$$L_{\mathrm{asym}}(\boldsymbol{w}, \beta) := \log p_{\mathrm{asym}}(\boldsymbol{w}) + L_0(\boldsymbol{w}, \beta) \tag{11}$$

と表される．非対称トビットモデルを打ち切りデータセットにフィットさせるために，この正則化対数尤度関数 (11) を最大化させる．次節で $L_{\mathrm{asym}}(\boldsymbol{w}, \beta)$ 最大化のためのアルゴリズムを示す．

4. 非対称トビットモデルのフィッティング法

本研究では，非対称トビットモデルをデータにフィットさせるための新しい最適化アルゴリズムを開発した．正則化対数尤度関数 (11) の最大値を見つけるために，EM法を採用する．一般に，事前分布を替えると技術的な難しさが発生する．本節では，たとえ事前分布を通常の正規分布を非対称正規分布に置き換えても，EM法の反復の高速計算を維持できることを示す．

EM法は潜在変数モデルをデータにフィットさせるための汎用的な枠組みである．各反復はEステップとMステップからなっている．トビット解析のためのEM法は次のQ関数を用いる：

$$\begin{aligned} Q(\boldsymbol{w}, \beta, q) &:= \log p(\boldsymbol{w}) + \frac{n}{2}\log\beta \\ &+ \sum_{i=1}^{n_{\mathrm{v}}} \log\phi\left(\sqrt{\beta}(y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i\rangle)\right) \\ &+ \sum_{i=n_{\mathrm{v}}+1}^{n} \mathbb{E}_{q_i(y_i)}\left[\log\phi\left(\sqrt{\beta}(y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i\rangle)\right)\right] \end{aligned} \tag{12}$$

ただし，q は $(n - n_{\mathrm{v}})$ 個の密度関数 $q_{n_{\mathrm{v}}+1}(y_{n_{\mathrm{v}}+1}), \ldots, q_n(y_n)$ の集合である．$p(\boldsymbol{w})$ は回帰係数 $\boldsymbol{w}$ の事前分布である．従来のトビットモデルでは $p = p_{\mathrm{sym}}$，非対称トビットモデルでは $p = p_{\mathrm{asym}}$ である．第 $(t-1)$ 反復で得られたモデルパラメータの値を $(\boldsymbol{w}^{(t-1)}, \beta^{(t-1)})$ と書くことにする．第 t 反復における分布集合 q を $q^{(t)} := \left(q_i^{(t)}\right)_{i=n_{\mathrm{v}}+1}^{n}$ と書くことにする．第 t 反復は次の手続きからなる：

(1) 密度関数 $q_i^{(t)}$ を $(\boldsymbol{w}^{(t-1)}, \beta^{(t-1)})$ に基づく y_i の事後分布に設定し，Q 関数中の期待値を含む項を更新する．

(2)
$$\boldsymbol{w}^{(t)} := \operatorname*{argmax}_{\boldsymbol{w}\in\mathbb{R}^d} Q(\boldsymbol{w}, \beta^{(t-1)}, q^{(t)}); \tag{13}$$

(3)
$$\beta^{(t)} := \operatorname*{argmax}_{\beta\in\mathbb{R}} Q(\boldsymbol{w}^{(t)}, \beta, q^{(t)}); \tag{14}$$

1行目はEステップと呼ばれ，2～3行目はMステップと呼ばれる．Eステップと β の更新則は $\boldsymbol{w}$ の事前分布を変更しても変わらない．これに対して，$\boldsymbol{w}$ の更新則は事前分布の変更によって複雑化しうる．本研究では，次の定理を見つけた：

Theorem 1. トビットモデルのフィッティングのためのEM法を考える．$p = p_{\mathrm{asym}}$ としたとき，$\boldsymbol{w}$ の更新則 (13) は非負最小二乗問題に帰着される．

この定理は，回帰係数 $\boldsymbol{w}$ の非対称正規分布に置き換えても，EM法の各反復は高速に行えることを示唆している．

回帰係数 $\boldsymbol{w}$ の更新則を議論する前に，Eステップと逆分散パラメータ β の更新則を述べる．4個の変数

$$\begin{aligned} \boldsymbol{y}^{\mathrm{v}} &:= [y_1, \ldots, y_{n_{\mathrm{v}}}]^\top, & \boldsymbol{y}^{\mathrm{h}} &:= [y_{n_{\mathrm{v}}+1}, \ldots, y_n]^\top, \\ \boldsymbol{X}^{\mathrm{v}} &:= [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_{\mathrm{v}}}], & \boldsymbol{X}^{\mathrm{h}} &:= [\boldsymbol{x}_{n_{\mathrm{v}}+1}, \ldots, \boldsymbol{x}_n] \end{aligned}$$

を導入し，$\boldsymbol{X} := \left[\boldsymbol{X}^{\mathrm{v}}, \boldsymbol{X}^{\mathrm{h}}\right]$ とおく．第 t 反復でのEステップで計算される事後分布は次のように更新される：

$$q_i^{(t)}(y_i) = f_{\mathrm{tn}}\left(y_i \,\middle|\, \left\langle \boldsymbol{w}^{(t-1)}, \boldsymbol{x}_i \right\rangle, \beta^{(t-1)}, \theta\right). \tag{15}$$

これにより，Q関数に含まれる次の統計量を計算できるようになる：

$$\begin{aligned} \bar{\boldsymbol{y}}^{(t)} &:= \left[(\boldsymbol{y}^{\mathrm{v}})^\top, \mathbb{E}_{q^{(t)}}\left[(\boldsymbol{y}^{\mathrm{h}})^\top\right]\right]^\top, \\ v^{(t)} &:= \mathbb{E}_{q^{(t)}}\left[\left\|\boldsymbol{y}^{\mathrm{h}}\right\|^2\right] - \left\|\mathbb{E}_{q^{(t)}}\left[\boldsymbol{y}^{\mathrm{h}}\right]\right\|^2. \end{aligned} \tag{16}$$

$\bar{\boldsymbol{y}}^{(t)}$ および $v^{(t)}$ に含まれる期待値は (3) および (6) を使うと，閉じた形で計算することができる．

逆分散パラメータ β の更新則 (14) はQ関数の偏導関数を 0 とおくことですぐに得られる：

$$\beta^{(t)} = \frac{n}{\|\boldsymbol{X}^\top\boldsymbol{w} - \bar{\boldsymbol{y}}^{(t)}\|^2 + v^{(t)}}. \tag{17}$$

このようにEステップと逆分散パラメータ β の更新は大きな計算コストをかけずに実行できる．これらの更新則は回帰係数の事前分布の定義に対して不変である．

定理 1 は次のように証明される．

定理 1 の証明：$2d \times (n+2d)$ 行列 $\boldsymbol{A}^{(t)}$ および $(n+2d)$ 次元ベクトル $\boldsymbol{b}^{(t)}$ を

$$\boldsymbol{A}^{(t)} := \begin{bmatrix} \boldsymbol{X} & \mathrm{diag}\left(\frac{\boldsymbol{\lambda}^{\mathrm{p}}}{\beta^{(t-1)}}\right)^{1/2} & \boldsymbol{O} \\ -\boldsymbol{X} & \boldsymbol{O} & \mathrm{diag}\left(\frac{\boldsymbol{\lambda}^{\mathrm{n}}}{\beta^{(t-1)}}\right)^{1/2} \end{bmatrix},$$
$$\boldsymbol{b}^{(t)} := \begin{bmatrix} \bar{\boldsymbol{y}}^{(t)} \\ \boldsymbol{0}_{2d} \end{bmatrix} \tag{18}$$

とおく．回帰係数ベクトル $\boldsymbol{w} \in \mathbb{R}^d$ を２つの非負ベクトル $\boldsymbol{w}_+, \boldsymbol{w}_- \in \mathbb{R}^d_+$ を使って，$\boldsymbol{w} = \boldsymbol{w}_+ - \boldsymbol{w}_-$ のように分解する．この２つの非負ベクトルを使って，Q関数は次のように書き直すことができる：

$$Q(\boldsymbol{w}_+ - \boldsymbol{w}_-, \beta^{(t-1)}, q^{(t)}) = -\frac{\beta}{2}\left\| (\boldsymbol{A}^{(t)})^\top \begin{bmatrix} \boldsymbol{w}_+ \\ \boldsymbol{w}_- \end{bmatrix} - \boldsymbol{b}^{(t)} \right\|^2 + \mathrm{const} \tag{19}$$

ただし，const は回帰係数に依存しない項をあらわす．式 (19) より，$Q(\cdot, \beta^{(t-1)}, q^{(t)})$ を最大化するための部分問題は非負最小二乗問題 $\mathrm{NNLS}(\boldsymbol{A}^{(t)}, \boldsymbol{b}^{(t)})$ に帰着されることが分かる．□

$\mathrm{NNLS}(\boldsymbol{A}^{(t)}, \boldsymbol{b}^{(t)})$ の最適解を $\begin{bmatrix} \boldsymbol{w}_+^{(t)} \\ \boldsymbol{w}_-^{(t)} \end{bmatrix}$ と書くとすると，回帰係数ベクトルは $\boldsymbol{w}^{(t)} := \boldsymbol{w}_+^{(t)} - \boldsymbol{w}_-^{(t)}$ によって復元できる．これらをまとめたＥＭ法の疑似コードは文献 [3] の Algorithm 1 に記述した．

5. 実験

3 節で述べた相関計算法の性能を調査するために数値実験を実施した．３個の水質データセット Indian, NY Harbor, Sapporo を用いた．Indian データセットは 1,580 個のレコードからなり，１レコードは 6 変量 FC, TC, pH, Cond, N, BOD からなる．NY Harbor データセットおよび Sapporo データセットの概要と実験条件の詳細は，文献 [7] を参照されたい．推定された相関係数 $\hat{R}$ は打ち切られていないデータにおける相関係数との差の絶対値で評価することにした．すなわち，その誤差は $|\hat{R} - R(\boldsymbol{y}_{\mathrm{a}}, \boldsymbol{y}_{\mathrm{b}})|$ と表される．ただし，$\boldsymbol{y}_{\mathrm{a}} \in \mathbb{R}^n$ および $\boldsymbol{y}_{\mathrm{b}} \in \mathbb{R}^n$ は検出限界導入前の微生物濃度である．この手続きを 50 回試行して，50 個の誤差を得た．

本稿の表 1，および文献 [3] の表 2，表 3 に，それぞれ Indian, NY Harbor, Sapporo におけるすべての微生物の組み合わせに対する推定誤差の平均を示す．括弧内の数値は標準偏差である．３手法から得られる３個の誤差のうち最小の誤差を太字にした．Indian では，非対称トビット相関計算法と従来トビット相関計算法は，それぞれ 28 ペア，2 ペアで最小誤差となった．NY Harbor では，非対称トビット相関計算法と従来トビット相関計算法は，それぞれ 12 ペア，2 ペアで最小誤差となった．Sapporo では 39 ペア，19 ペアで最小となった．ナイーブ相関計算法はどのデータセット，どのペアでも最小誤差にならなかった．これらの結果は，非検出濃度の補完によって相関計算の精度を向上できることを示している．では補完が推定精度を向上させるのか考察しよう．２つの微生物どうしの相関が弱

表 1 Indian データセットにおける推定誤差．3,4,5 列目には非対称トビット相関計算法，従来トビット相関計算法，ナイーブ相関計算法が算出した推定誤差の平均を記載．括弧内の数値は標準偏差．

A	B	非対称トビット	従来トビット	ナイーブ
FC	TC	**0.025** (0.020)	0.075 (0.064)	0.083 (0.100)
FC	pH	**0.134** (0.104)	0.171 (0.115)	0.623 (0.267)
FC	Cond	**0.112** (0.091)	0.119 (0.093)	0.522 (0.335)
FC	N	**0.116** (0.083)	0.131 (0.092)	0.419 (0.272)
FC	BOD	0.156 (0.123)	**0.151** (0.132)	0.453 (0.250)
TC	FC	**0.028** (0.022)	0.060 (0.057)	0.083 (0.100)
TC	pH	**0.116** (0.084)	0.163 (0.115)	0.635 (0.308)
TC	Cond	**0.142** (0.082)	0.178 (0.104)	0.732 (0.343)
TC	N	**0.101** (0.081)	0.114 (0.087)	0.376 (0.313)
TC	BOD	**0.091** (0.069)	0.096 (0.068)	0.441 (0.347)
pH	FC	**0.141** (0.091)	0.191 (0.107)	0.623 (0.267)
pH	TC	**0.124** (0.091)	0.165 (0.116)	0.635 (0.308)
pH	Cond	**0.144** (0.068)	0.167 (0.081)	0.684 (0.364)
pH	N	**0.114** (0.094)	0.127 (0.107)	0.978 (0.036)
pH	BOD	**0.131** (0.087)	0.156 (0.123)	0.600 (0.318)
Cond	FC	**0.098** (0.081)	0.111 (0.086)	0.522 (0.335)
Cond	TC	**0.135** (0.078)	0.167 (0.093)	0.729 (0.341)
Cond	pH	**0.128** (0.068)	0.161 (0.081)	0.684 (0.364)
Cond	N	**0.066** (0.060)	0.090 (0.074)	0.558 (0.302)
Cond	BOD	**0.066** (0.049)	0.070 (0.051)	0.518 (0.318)
N	FC	**0.133** (0.087)	0.145 (0.093)	0.419 (0.272)
N	TC	**0.115** (0.088)	0.127 (0.098)	0.376 (0.313)
N	pH	**0.070** (0.052)	0.098 (0.080)	0.978 (0.036)
N	Cond	**0.071** (0.064)	0.098 (0.080)	0.558 (0.302)
N	BOD	**0.143** (0.072)	**0.143** (0.072)	0.342 (0.215)
BOD	FC	0.165 (0.118)	**0.161** (0.140)	0.453 (0.250)
BOD	TC	**0.103** (0.075)	0.110 (0.078)	0.441 (0.347)
BOD	pH	**0.111** (0.088)	0.154 (0.130)	0.600 (0.318)
BOD	Cond	**0.053** (0.041)	0.070 (0.057)	0.518 (0.318)
BOD	N	**0.139** (0.085)	0.143 (0.088)	0.342 (0.215)

表 2 Indian データセットにおける計算時間の実測値

n	非対称トビット	従来トビット
10	0.330 (0.001)	**0.321** (0.001)
17	0.536 (0.001)	**0.529** (0.001)
31	**0.957** (0.006)	0.964 (0.012)
56	**1.706** (0.002)	1.715 (0.003)
100	3.026 (0.010)	**3.025** (0.002)
177	**5.315** (0.011)	5.322 (0.001)
316	**9.429** (0.034)	9.434 (0.032)
562	16.665 (0.046)	**16.633** (0.021)
1000	**29.610** (0.094)	29.660 (0.080)

いと共通可視濃度 $\mathcal{I}_{vv}$ は少なくなりやすい．このような場合，ナイーブ相関計算法は，小さなペアデータから相関を計算することになるため，相関係数の精度は悪くなりやすい．非対称トビット相関計算法は従来トビット相関計算法より推定精度が高くなることが多かった．これは，非対称トビット相関計算法が効果的にドメイン知識を活用して非検出値の補完の精度を向上させ，ゆえに相関係数の誤差を小さくしたから，といえる．

最後に，計算時間に関して報告する．本研究の理論的発見の一つは非対称正規事前分布を使っても，トビットモデルのためのＥM法のMステップは高速に計算できることであった．通常の正規分布を事前分布に使っていた従来のトビットモデルではMステップは制約なし最小二乗推定問題に帰着できる．これに対して，定理 1 で述べたように，本研究では，非対称正規事前分布を使用した場合，非負最小二乗問題に帰着できることを見つけた．では，従来トビットモデルと比べて，非対称トビットモデルはどれほどフィッティングの計算時間が増加するか確認しよう．従来モデルと非対称モデルに対するＥM法 30 反復の計算時間を計測した．本稿の表 2，および文献 [3] の表 4(b)(c) それぞれに Indian, NY Harbor，Sapporo での 10 試行の平均時間を示す．単位は秒，括弧内の数値は標準偏差である．驚くことに，いずれの場合も２つのモデルの計算時間の差はほとんどなかった．これは，水質データ解析の場合，回帰係数の個数があまり多くないことに起因する．非負最小二乗問題は d が小さければ高速に解くことができる．Ｅステップでは $(n - n_{\mathrm{v}})$ 個の非検出データに対して累積密度関数の値を計算する必要があり，このＥステップの計算コストが比較的重いので，Mステップの変更による全体の計算時間の差はほぼ無視できるほど小さかった．

6. 結論

本稿では，水中の病原微生物の濃度の相関解析において，サイド情報を用いて非検出値を補完してから相関解析を行うと精度が向上することを示した．数値実験の結果，非検出値補完のためにドメイン知識を活用するとサイド情報の効果が増幅されることが分かった．非検出値補完で重要となる技術はトビットモデルである．本研究では，非対称正規分布を回帰係数の事前分布として導入することで，ドメイン知識を活用した．トビットモデルのフィッティングにはＥM法が使われる．本研究では，非対称正規分布を事前分布にした場合，ＥM法のMステップは非負最小二乗問題に帰着されることを理論的に示した．また，数値実験により，非負最小二乗問題を解くための計算時間はＥステップにかかる計算時間よりかなり軽いため，非対称正規分布に入れ替えたことにより全体の計算時間はほとんど増加しないことを示した．今後の課題としては，本研究で開発した相関解析法を実際の病原体濃度の解析に応用して，進化を続ける病原微生物濃度の測定法にあわせて適切なモニタリングのルーチンを設計していくことがあげられる．

謝辞 本研究の一部は、（独）環境再生保全機構の環境研究総合推進費（JPMEERF20205006）により実施された．また，JSPS 科研費 19K04661 の助成を受けた．

参考文献

[1] Takeshi Amemiya. Tobit models: A survey. *Journal of Econometrics*, 24(1-2):3–61, January 1984.

[2] Stefania Bellavia, Maria Macconi, and Benedetta Morini. An interior point newton-like method for non-negative least-squares problems with degenerate solution. *Numerical Linear Algebra with Applications*, 13(10):825–846, 2006. doi: 10.1002/nla.502.

[3] HongYuan Cao and Tsuyoshi Kato. Asymmetric tobit analysis for correlation estimation from censored data. *arXiv*, -(-):http://arxiv.org/abs/2101.10853, Jan 2021.

[4] Donghui Chen and Robert J. Plemmons. Nonnegativity constraints in numerical analysis. In Adhemar Bultheel and Ronald Cools, editors, *The Birth of Numerical Analysis*, pages 109–139. World Scientific, Nov 2009. doi: 10.1142/9789812836267_0008.

[5] James Dobrowoski, Michael O'Neill, Lisa Duriancik, and Joanne Throwe. Opportunities and challenges in agricultural water reuse: Final report. *USDA-CSREES*, 89:–, -2008.

[6] J. Gentry, J. Vinje, and E. K. Lipp. A rapid and efficient method for quantitation of genogroups i and ii norovirus from oysters and application in other complex environmental samples. *J Virol Methods*, 156(1–2):59–65, Mar 2009.

[7] T. Kato, A. Kobayashi, W. Oishi, S. S. Kadoya, S. Okabe, N. Ohta, M. Amarasiri, and D. Sano. Sign-constrained linear regression for prediction of microbe concentration based on water quality datasets. *J Water Health*, 17(3):404–415, Jun 2019.

[8] Tsuyoshi Kato, Shinichiro Omachi, and Hirotomo Aso. Asymmetric gaussian and its application to pattern recognition. In *S+SSPR2002*, pages 405–413, 2002.

[9] M. H. Kramer, B. L. Herwaldt, G. F. Craun, R. L. Calderon, and D. D. Juranek. Surveillance for waterborne-disease outbreaks–united states, 1993-1994. *MMWR CDC Surveill Summ*, 45(1):1–33, Apr 1996.

[10] Charles L. Lawson and Richard J. Hanson. *Solving Least Squares Problems*. Society for Industrial and Applied Mathematics, jan 1995. doi:10.1137/1.9781611971217.

[11] Nicolai Meinshausen. Sign-constrained least squares estimation for high-dimensional regression. *Electronic Journal of Statistics*, 7:1607–1631, 2013. doi: 10.1214/13-ejs818.

[12] Rachel T. Noble and Stephen B. Weisberg. A review of technologies for rapid detection of bacteria in recreational waters. *Journal of Water and Health*, 3(4):381–392, December 2005.

[13] Committee on Indicators for Waterbone Pathogens. *Indicators for Waterborne Pathogens*. National Academies Press, 2004.

[14] Francisco Pedrero, Ioannis Kalavrouziotis, Juan Jose Alarcon, Prodromos Koukoulakis, and Takashi Asano. Use of treated municipal wastewater in irrigated agriculture–review of some practices in spain and greece. *Agricultural Water Management*, 97(9):1233–1241, Sept 2010.