A-031

# 打ち切りデータ補完のための複数ターゲットトビットモデル
# Multi-target Tobit model for completion of censored data

高田 裕也 * 加藤 毅 †
Yuya Takada Tsuyoshi Kato

## 1. はじめに

飲用水や環境水には感染症を引き起こす病原体が含まれており，水利用における微生物学的安全性が脅かされている [2]．病原体には，サルモネラ菌，赤痢菌，ノロウイルス，ロタウイルスなど多岐にわたっており，その多くはヒトの糞便による汚染に由来している．水利用の使途と水中微生物の生態に応じて適切な水質衛生基準値を設定するためには，複数の水中微生物濃度およびほかの水質データから，それらの関係性を解析する．水中微生物濃度や水質データの関係性は，採水ごとに同時に測定された観測値を使って解析する．水中微生物濃度の定量化に際して最も大きな障害は陽性率の低さであり，少なからぬ割合で濃度が定量限界を下回っている（図 1(a),(c)）．相関解析などの統計解析を可能にするためには，非検出値の真の値を高精度に推定することが求められている．

打ち切りデータの解析には，**トビットモデル** [1] という統計学的手法がある．トビットモデルは回帰分析の一種で，目標変量のみに打ち切りを許している．病原体濃度の解析には，各病原体に対して目標変量にしてトビットモデルをあてはめることで非検出データを補完することができる．しかし，このアプローチを使うには，説明変量に使われる病原体濃度は打ち切られていても何らかの値で埋めてから適用する必要がある．そのため，トビットモデルは複数の病原体濃度を同時に推定することができない．

本研究では，トビットモデルの目標変量を複数化するように拡張を施すことで，打ち切られた複数の目的変量を同時に補完する新しい方法論，**複数ターゲットトビットモデル** (*MTTM*) を考案した．本研究の主要な理論的結果は 2 個の発見からなっている．

> **第 1 の発見**：トビットモデルで用いられている尤度関数の一部の項がカルバックライブラ距離（KL 距離）の最小値で表される（定理 1）．

目標変量の複数化は，その表現を使った尤度関数を複数の目標変量について重ね合わせることで実現した．この拡張によって得られた目的関数最大化のための最適化アルゴリズムに関して，次のことを発見をした．

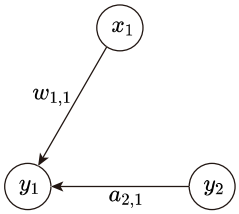
> **第 2 の発見**：目的関数の最大化にブロック座標上昇法を用いたとき，各ステップが閉形式で表される（定理 3）．

この発見により，ステップサイズのような調整を要するパラメータが不要となり，数値的に安定した最適化が可能となる．水中微生物濃度の実データを用いた数値実験によって検証したところ，提案する複数ターゲットトビットモデルは従来のトビットモデルを顕著に上回る精度で非検出値を補完できることを確認した．

(a)

| $y_1$ | $y_2$ | $x_1$ |
|---|---|---|
| ≦4.0 | 6.5 | 6.9 |
| ≦4.0 | 3.0 | 3.2 |
| ≦4.0 | 5.5 | 9.5 |
| 5.0 | 7.0 | 0.3 |
| 10.2 | 3.8 | 4.4 |
| 10.7 | 9.0 | 3.8 |

(b) 従来のトビットモデル




(c)

| $y_1$ | $y_2$ | $x_1$ |
|---|---|---|
| ≦4.0 | 6.5 | 6.9 |
| ≦4.0 | ≦4.0 | 3.2 |
| ≦4.0 | 5.5 | 9.5 |
| 5.0 | 7.0 | 0.3 |
| 10.2 | ≦4.0 | 4.4 |
| 10.7 | 9.0 | 3.8 |

(d) 提案モデル (MTTM)

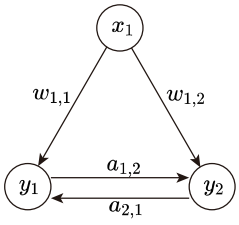



図 1: **従来トビットモデルと提案モデル．**従来のトビットモデルは，打ち切りがある列を含む表 ((a) 参照) から，その列を目標変数とした 回帰モデル ((b) 参照) をデータにフィットさせる枠組みである．提案モデルである MTTM は，複数の打ち切りを含む表 ((c) 参照) から，複数の回帰モデルを重ね合わせる ((d) 参照) ことで，同時にフィットさせることができる．

## 2. 従来のトビットモデル（STTM）

トビットモデル [1] はデータセット $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ $(i \in [n] := \{1, \ldots, n\})$ に対し，回帰モデル

$$y_i = \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + \epsilon_i$$

を仮定して，偏回帰係数 $\boldsymbol{w} \in \mathbb{R}^d$ の値を推定するモデルである．$\epsilon_i$ は雑音であり，$\epsilon_i \sim \mathcal{N}(0, \beta^{-1})$ が仮定される．定量限界を $u \in \mathbb{R}$ とおき，$n_{\mathrm{v}}$ 個が定量限界を越えているとする．$i \in [n_{\mathrm{v}}]$ に対し，$y_i > u$，$i \in [n] \setminus [n_{\mathrm{v}}]$ に対し，$y_i \leq u$ となるよう，例題番号を入れ替える．トビットモデルが標準的な線形回帰分析と異なる点は，$y_i$ が定量限界 $u$ を下回ったとき，$y_i \leq u$ という情報は利用できるが，$y_i$ の値そのものは利用できないことである．トビットモデルのフィッティングには次の対数

*群馬大学理工学府
†群馬大学情報学部

尤度関数が用いられている：

$$\mathcal{L}_{\mathrm{s}}(\boldsymbol{\theta}_{\mathrm{s}}) := \sum_{i=1}^{n_{\mathrm{v}}} \log \mathcal{N}(y_i\,;\,\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle, \beta^{-1}) + \sum_{i=n_{\mathrm{v}}+1}^{n} \log \int_{-\infty}^{u} \mathcal{N}(t_i\,;\,\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle, \beta^{-1}) dt_i, \tag{1}$$

ただし，$\boldsymbol{\theta}_{\mathrm{s}} := (\boldsymbol{w}, \beta)$ はモデルパラメータを表す．$\boldsymbol{\theta}_{\mathrm{s}}$ の最適解はニュートン法などで見つけることができる [1]．

筆者らは (1) の第 2 項の符号を逆転すると，KL 距離の最小値に等しいことを見つけた．次節では，この事実を利用するとトビットモデルの自然な拡張が得られることを示す．$\mathcal{P}$ を $\mathbb{R} \to \mathbb{R}$ なる確率密度関数の集合とし，次のような $n$ 個の確率密度関数 $q_1, \ldots, q_n \in \mathcal{P}$ を導入する．最初の $n_{\mathrm{v}}$ 個は Dirac デルタ関数 $\delta(\cdot)$ で定義する：$\forall i \in [n_{\mathrm{v}}]$,

$$q_i(t) = \delta(t - y_i).$$

$i \in [n_{\mathrm{v}}]$ に対して，$Q_i := \{\delta(\cdot - y_i)\}$ とおく．残りの $n_{\mathrm{h}} (:= n - n_{\mathrm{v}})$ 個の確率密度関数 $q_{n_{\mathrm{v}}+1}, \ldots, q_n \in \mathcal{P}$ は以下を満たすとする：$\forall i \in [n] \setminus [n_{\mathrm{v}}]$,

$$\forall t > u, \quad q_i(t) = 0. \tag{2}$$

$i \in [n] \setminus [n_{\mathrm{v}}]$ に対して，$\mathcal{Q}_i := \{q_i \in \mathcal{P} \,|\, \forall t > u,\, q_i(t) = 0\}$ とおき，また，$\mathcal{Q}_{\mathrm{s}} := \mathcal{Q}_1 \times \cdots \times \mathcal{Q}_n$ とおく．(1) の第 2 項に関して，

$$-\log \int_{-\infty}^{u} \mathcal{N}(t_i\,;\,\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle, \beta^{-1}) dt_i = \min_{q_i \in \mathcal{Q}_i} \mathrm{KL}\left[q_i || \mathcal{N}(\cdot\,;\,\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle, \beta^{-1})\right] \tag{3}$$

が成り立つことを示すことが出来る．ただし，$\mathrm{KL}(q||p)$ は $p \in \mathcal{P}$ から $q \in \mathcal{P}$ への KL 距離を表す‡．等式 (3) の導出は 5 節参照．この事実は直ちに次の定理を導く．

**定理 1.** $\beta > 0$ なる任意の $\boldsymbol{\theta}_s = (\boldsymbol{w}, \beta) \in \mathbb{R}^d \times \mathbb{R}$ に対して，

$$\mathcal{L}_s(\boldsymbol{\theta}_s) = \max_{q \in \mathcal{Q}_s} \mathcal{F}_s(\boldsymbol{\theta}_s, q)$$

が成り立つ．ただし，$\mathcal{F}_s(\boldsymbol{\theta}_s, q)$ は，$q := (q_1, \ldots, q_n) \in \mathcal{Q}_s$ に対して，

$$\mathcal{F}_s(\boldsymbol{\theta}_s, q) := \sum_{i=n_v+1}^{n} \mathcal{H}[q_i] + \sum_{i=1}^{n} \int q_i(t_i) \cdot \log \mathcal{N}(t_i\,;\,\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle, \beta^{-1}) dt_i$$

とおいた．$\mathcal{H}[\cdot] : \mathcal{P} \to \mathbb{R}$ はエントロピーを表す‡．

‡本稿では，KL 距離 $KL[q||p]$ およびエントロピー $\mathcal{H}[q]$ の積分は $q(x) > 0$ なる区間 $I$ のみとする．すなわち，$\mathcal{H}[q] := -\int_{x \in I} q(x) \log q(x) dx$ および $\mathrm{KL}[q||p] := \int_{x \in I} q(x) \log(q(x)/p(x)) dx$ という定義を用いる．

## 3. 複数ターゲットトビットモデル (MTTM)

本節では，トビットモデルを複数のターゲット変数を扱えるようにするために，定理 1 から自然に導かれる拡張した MTTM を提案する．

目標変量の個数を $m$ とおき，MTTM をフィットさせるデータセットを

$$(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n) \in \mathbb{R}^d \times \mathbb{R}^m$$

とする．各例題は説明変量 $\boldsymbol{x}_i \in \mathbb{R}^d$ と目標変量

$$\boldsymbol{y}_i := [y_{1,i}, \ldots, y_{m,i}]^\top$$

からなる．ベクトル $\boldsymbol{y}_i \in \mathbb{R}^m$ のうち，第 $k$ 要素を除いたベクトルを $\boldsymbol{y}_{\setminus k,i} := [y_{1,i}, \ldots, y_{k-1,i}, y_{k+1,i}, \ldots, y_{m,i}]^\top$ と書くことにする．MTTM の回帰モデルは $m$ 個の回帰モデル

$$k \in [m], \quad y_{k,i} = \left\langle \boldsymbol{a}_k, \boldsymbol{y}_{\setminus k,i} \right\rangle + \langle \boldsymbol{w}_k, \boldsymbol{x}_i \rangle + \epsilon_{k,i}$$

で構成する．ただし，$\boldsymbol{a}_k \in \mathbb{R}^{m-1}$ および $\boldsymbol{w}_k \in \mathbb{R}^d$ は第 $k$ 回帰モデルの偏回帰係数である；$\epsilon_{k,i} \sim \mathcal{N}(0, \beta^{-1})$ は雑音である．各目標変量はほかの目標変量に対する説明変量になっていることに注意されたい．

MTTM では，目標変量の行列 $\boldsymbol{Y} := [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n] \in \mathbb{R}^{m \times n}$ が不完全になっている状況を仮定する．目標変量 $y_{k,i} \in \mathbb{R}$ が定量限界 $u_k \in \mathbb{R}$ より大きければ，$y_{k,i}$ の値を利用できる．しかし，$y_{k,i} \le u_k$ ならば，$y_{k,i}$ の値を利用できない．欠損要素の添え字集合を $\mathcal{E}_{\mathrm{h}}$，観測要素の添え字集合を $\mathcal{E}_{\mathrm{v}}$ とおく; i.e.

$$\mathcal{E}_{\mathrm{h}} := \{(k,i) \in [m] \times [n] \,|\, y_{k,i} \le u_k\},$$
$$\mathcal{E}_{\mathrm{v}} := \{(k,i) \in [m] \times [n] \,|\, y_{k,i} > u_k\}.$$

偏回帰係数の決定に用いる目的関数を定義するために，次のような $mn$ 個の確率密度関数 $q_{k,i} \in \mathcal{P}$ を導入する：

$$\forall (k,i) \in \mathcal{E}_{\mathrm{v}}, \qquad q_{k,i}(t) = \delta(t - y_{k,i}) \tag{4}$$

$$\forall (k,i) \in \mathcal{E}_{\mathrm{h}}, \qquad \begin{array}{ll} q_{k,i}(t) = 0 & \text{if } t > u_k, \\ q_{k,i}(t) > 0 & \text{if } t \le u_k. \end{array} \tag{5}$$

また，$(k,i) \in \mathcal{E}_{\mathrm{v}}$ に対して，$\mathcal{Q}_{k,i} := \{\delta(\cdot - y_{k,i})\}$ とおく．$(k,i) \in \mathcal{E}_{\mathrm{h}}$ に対して，$\mathcal{Q}_{k,i}$ を式 (5) を満たす確率密度関数の集合とおく．これらの直積集合を $\mathcal{Q}_{\mathrm{m}} := \prod_{(k,i) \in [m] \times [n]} \mathcal{Q}_{k,i}$ とおく．$\mathcal{Q}_{\mathrm{m}}$ を使って，$\mathcal{F}_{\mathrm{s}}$ を複数目標変量に対応できるよう次のように拡張する：

$$\mathcal{F}_{\mathrm{m}}(\boldsymbol{\theta}_{\mathrm{m}}, q) := \sum_{(k,i) \in \mathcal{E}_{\mathrm{h}}} \mathcal{H}[q_{k,i}] + \sum_{i=1}^{n} \sum_{k=1}^{m} \int q_{1,i}(t_{1,i}) \cdot \cdots \cdot q_{m,i}(t_{m,i}) \log \mathcal{N}(t_{k,i}\,;\,\langle \boldsymbol{a}_k, \boldsymbol{t}_i \rangle + \langle \boldsymbol{w}_k, \boldsymbol{x}_i \rangle, \beta^{-1}) d\boldsymbol{t}_i$$

ただし，$\boldsymbol{t}_i := [t_{1,i}, \ldots, t_{m,i}]^\top$, $q := (q_{k,i})_{k,i} \in \mathcal{Q}_{\mathrm{m}}$ であり，$\boldsymbol{\theta}_{\mathrm{m}} := (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_m, \beta)$ と再定義した．MTTM のパラメータ $\boldsymbol{\theta}_{\mathrm{m}}$ の値を決定するために，次の最適化問題を解く：

**MTTM のフィッティング問題**

$$\begin{aligned} &\max \quad \mathcal{L}_{\mathrm{m}}(\boldsymbol{\theta}_{\mathrm{m}}) \qquad \text{wrt } \boldsymbol{\theta}_{\mathrm{m}}, \\ &\text{where} \quad \mathcal{L}_{\mathrm{m}}(\boldsymbol{\theta}_{\mathrm{m}}) := \max_{q \in \mathcal{Q}_{\mathrm{m}}} \mathcal{F}_{\mathrm{m}}(\boldsymbol{\theta}_{\mathrm{m}}, q). \end{aligned} \tag{6}$$

すなわち，$\boldsymbol{\theta}_{\mathrm{m}}$ と $q \in \mathcal{Q}_{\mathrm{m}}$ の同時最適化によって偏回帰係数の値を決定する．

**欠損値の補完．** 最適解 $(\boldsymbol{\theta}^{\star}_{\mathrm{m}}, q^{\star})$ を得ることが出来れば，定量限界以下の値を期待値 $\mathbb{E}_{q_{k,i}(t_{k,i})}[t_{k,i}]$ を使って推定できる．次の定理によって，期待値計算が容易であることが分かる．(証明は 5 節参照.)

**定理 2.** 最大化問題 (6) の最適解 $(\boldsymbol{\theta}^{\star}_{m}, q^{\star})$ において，$q^{\star} \in \mathcal{Q}_m$ の各要素 $q^{\star}_{k,i} \in \mathcal{Q}_{k,i}$ は切断正規分布で表すことができる．すなわち，

$$\exists \mu^{\star}_{k,i} \in \mathbb{R}, \exists \sigma^{\star}_{k,i} \in \mathbb{R}, \quad q^{\star}_{k,i} = f_{TN}(\cdot\,;\, \mu^{\star}_{k,i}, \sigma^{\star}_{k,i}, u_k).$$

ただし，$f_{TN}(\cdot\,;\, \mu, \sigma, u)$ は切断正規分布の確率密度関数を表す：

$$f_{TN}(x\,;\, \mu, \sigma, u) := \begin{cases} \frac{\phi((x-\mu)/\sigma)}{\sigma \Phi((x-\mu)/\sigma)} & for\, x \leq u, \\ 0 & for\, x > u. \end{cases}$$

ただし，$\phi$ および $\Phi$ はそれぞれ標準正規分布の確率密度関数および累積分布関数を表す．

定理 2 より，$y_{k,i}$ の推定値は，$\widehat{u}^{\star}_{k,i} := (u_k - \mu^{\star}_{k,i})/\sigma^{\star}_{k,i}$ を使って

$$\mathbb{E}_{q_{k,i}(t_{k,i})}[t_{k,i}] = \mu^{\star}_{k,i} - \sigma^{\star}_{k,i} \frac{\phi(\widehat{u}^{\star}_{k,i})}{\Phi(\widehat{u}^{\star}_{k,i})}$$

と求められる．

## 4.MTTM のフィッティング

前節で，MTTM のデータへのフィッティングを $\mathcal{F}_m(\boldsymbol{\theta}_{\mathrm{m}}, q)$ の最大化によって行うことを述べた．筆者らは，$\mathcal{F}_m(\boldsymbol{\theta}_{\mathrm{m}}, q)$ の最大化のためにブロック座標上昇法を採用した．Algorithm 1 に示すように，ブロック座標上昇法では，ある $(k,i) \in \mathcal{E}_{\mathrm{h}}$ に対し，$q_{k,i}$ 以外の $q$ の構成要素および $\boldsymbol{\theta}_{\mathrm{m}}$ を固定して，$q_{k,i}$ に関して $\mathcal{F}_m(\boldsymbol{\theta}_{\mathrm{m}}, q)$ を最大化するステップ (Line 5) と，$q \in \mathcal{Q}_{\times}$ を固定して $\boldsymbol{\theta}_{\mathrm{m}}$ に関して $\mathcal{F}_m(\boldsymbol{\theta}_{\mathrm{m}}, q)$ を最大化するステップ (Line 7) からなる．このアルゴリズムに関して，筆者らは次の定理を見つけた．(証明は 5 節参照.)

**定理 3.** *Algorithm 1* の各ステップ *(Line 5* および *Line 7)* の解は閉形式で表すことができる．

**Algorithm 1:** Block coordinate ascent algorithm for maximizing $\mathcal{F}_{\mathrm{m}}(\boldsymbol{\theta}_{\mathrm{m}}, q)$.

```
1 begin
2     Initialize θ_m and q ∈ Q_×;
3     for t := 1, 2, ... do
4         for (k, i) ∈ E_h do
5             q_{k,i} := argmax_{q_{k,i} ∈ Q_{k,i}} F_m(θ_m, q);
6         end
7         θ_m := argmax_{θ_m} F_m(θ_m, q);
8     end
9 end
```

各ステップの具体的な更新則を以下に述べる．

### 4.1. 密度関数 $q$ の更新

Algorithm 1 の Line 5 では，$q_{k,i}$ を $\mathcal{F}_m(\boldsymbol{\theta}_{\mathrm{m}}, q)$ を最大化する密度関数に更新する．そのような $q_{k,i}$ は次のような切断正規分布で表される：

$$q^{\mathrm{new}}_{k,i}(t_{k,i}) = f_{\mathrm{TN}}(t_{k,i}; \mu_{k,i}, \sigma_{k,i}, u_k). \tag{7}$$

$k = 1$ のとき，$\mu_{k,i}$ および $\sigma_{k,i}$ はベクトル $\overline{\boldsymbol{y}}_{\mathrm{r}} = \left[\overline{y}_{2,i}, \ldots, \overline{y}_{m,i}\right]^{\top}$ および行列 $\boldsymbol{A}^{m \times m}$ を使って算出する．ただし，$\overline{y}_{k',i} := \mathbb{E}_{q_{k',i}(t_{k',i})}[t_{k',i}]$ であり，行列 $\boldsymbol{A}$ の第 $k'$ 列には，$k' \in [m]$ に対し，$\boldsymbol{a}_{k'} \in \mathbb{R}^{m-1}$ の第 $k'$ 要素に 0 を挿入した $m$ 次元列ベクトルを入れる．これらを使って，次のように $\mu_{k,i}$, $\sigma_{k,i}$ を計算する：

$$\mu_{k,i} := \frac{\left\langle \boldsymbol{b}_1, \boldsymbol{W}^{\top}\boldsymbol{x} - \boldsymbol{B}_{\mathrm{r}}^{\top}\overline{\boldsymbol{y}_{\mathrm{r}}} \right\rangle}{\|\boldsymbol{b}_1\|^2}, \; \sigma_{k,i} := \frac{1}{\sqrt{\beta}\|\boldsymbol{b}_1\|}.$$

ただし，$\boldsymbol{b}_1 \in \mathbb{R}^m$ および $\boldsymbol{B}_{\mathrm{r}} \in \mathbb{R}^{(m-1)\times m}$ は $\left[\boldsymbol{b}_1, \boldsymbol{B}_{\mathrm{r}}^{\top}\right] = \boldsymbol{I} - \boldsymbol{A}^{\top}$ を満たすように算出する．$k \neq 1$ のときは，添え字 $k$ と 1 を入れ替えてこの手続きを実行する．

### 4.2. モデルパラメータ $\boldsymbol{\theta}_{\mathrm{m}}$ の更新

Algorithm 1 の Line 7 における $\boldsymbol{\theta}_{\mathrm{m}}$ の更新式も閉じた形で表される．分布 $q_{k,i}$ に関する期待値

$$\overline{y}_{k,i} := \mathbb{E}_{q_{k,i}(t_{k,i})}[t_{k,i}], \quad v_{k,i} := \mathbb{E}_{q_{k,i}(t_{k,i})}[t^2_{k,i}]$$

から構成されるベクトル

$$\overline{\boldsymbol{y}}_i := \left[\overline{y}_{1,i}, \ldots, \overline{y}_{m,i}\right]^{\top}, \quad \boldsymbol{v}_i := \left[v_{1,i}, \ldots, v_{m,i}\right]^{\top}$$

を考え，それぞれから第 $k$ 要素を除いたベクトルを $\overline{\boldsymbol{y}}_{\setminus k,i}$, $\boldsymbol{v}_{\setminus k,i}$ とおく．また，$\forall k \in [m], \forall i \in [n]$ について，

$$\begin{aligned} \widetilde{\boldsymbol{w}}_k &:= \begin{bmatrix} \boldsymbol{a}_k \\ \boldsymbol{w}_k \end{bmatrix}, \qquad \widetilde{\boldsymbol{x}}_{k,i} := \begin{bmatrix} \overline{\boldsymbol{y}}_{\setminus k,i} \\ \boldsymbol{x}_i \end{bmatrix}, \\ \boldsymbol{G}_k &:= \mathrm{diag}\left( \begin{bmatrix} \sum_{i=1}^{n} \boldsymbol{v}_{\setminus k,i} \\ \boldsymbol{0}_d \end{bmatrix} \right) \end{aligned} \tag{8}$$

とおく．すると，密度関数 $q$ を固定して，目的関数 $\mathcal{F}_{\mathrm{m}}(\boldsymbol{\theta}_{\mathrm{m}}, q)$ を最大化する $(\widetilde{\boldsymbol{w}}_1, \ldots, \widetilde{\boldsymbol{w}}_m, \beta)$ は以下のよ

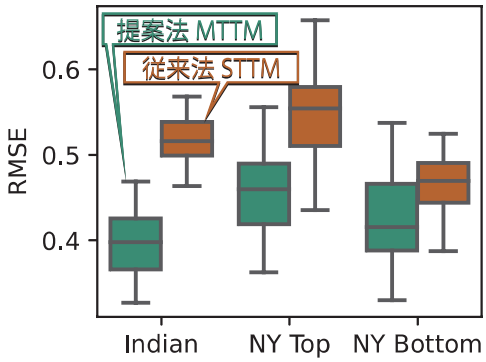



図 2: **性能評価.**

うに表される：

$$\widetilde{\boldsymbol{w}}_k^{\mathrm{new}} = \left(\boldsymbol{G}_k + \sum_{i=1}^{n} \widetilde{\boldsymbol{x}}_{k,i}\widetilde{\boldsymbol{x}}_{k,i}^{\top}\right)^{-1} \sum_{i=1}^{n} \bar{y}_{k,i}\widetilde{\boldsymbol{x}}_{k,i},$$

$$\frac{1}{\beta^{\mathrm{new}}} = \frac{1}{mn}\sum_{k=1}^{m} \langle \widetilde{\boldsymbol{w}}_k^{\mathrm{new}}, \boldsymbol{G}_k \widetilde{\boldsymbol{w}}_k^{\mathrm{new}} \rangle + \frac{1}{mn}\sum_{k,i} (\bar{y}_{k,i} - \langle \widetilde{\boldsymbol{w}}_k^{\mathrm{new}}, \widetilde{\boldsymbol{x}}_{k,i} \rangle)^2 + \frac{1}{mn}\sum_{k,i} v_{k,i}.$$

## 5. 導出と証明

定理 1 を導く等式 (3) の導出，定理 2，および定理 3 の証明において，次の補助定理がカギとなる：

**補助定理 4.** 関数 $f : \mathbb{R} \to \mathbb{R}$ に対して，次の最適化問題を考える：

$$\begin{aligned} max \quad & F[q] \quad wrt\ q \in \mathcal{P} \\ subject\ to \quad & \forall x > u, \quad q(x) = 0, \\ where \quad & F[q] := \mathbb{E}_{q(x)}\left[\log f(x)\right] + \mathcal{H}[q], \\ & \forall x \le u, \quad f(x) > 0. \end{aligned} \tag{9}$$

この問題の最適解は

$$q(x) := \begin{cases} \frac{f(x)}{\int_{-\infty}^{u} f(t)dt} & for\ x \le u, \\ 0 & for\ x > u \end{cases} \tag{10}$$

で与えられる．

補助定理 4 において，一部の区間で 0 という縛りが確率密度関数 $q \in \mathcal{P}$ になければ，平均場近似で使われるトリックに帰着する．補助定理 4 の証明は汎関数 $F[q]$ に対してラグランジュ乗数法を適用することで得ることができる．

**式 (3) の導出 (スケッチ)：**補助定理 4 における $F[q]$ は $f := \mathcal{N}(\cdot\,; \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle, \beta^{-1})$ のとき，$\mathcal{F}[q] = -\mathrm{KL}[q||f]$ が成り立つ．$Z := \int_{-\infty}^{u} f(t)dt$ とし，$\widehat{q}(x) := \mathbb{1}[x \le u] \cdot f(x)/Z$ とすると，補助定理 4 より，

$$\begin{aligned} \text{RHS of (3)} &= -\min_{q \in \mathcal{Q}} \mathrm{KL}[q||f] = \mathcal{F}[\widehat{q}] \\ &= -\mathrm{KL}[\widehat{q}||f] = \int_{-\infty}^{u} \frac{f(t)}{Z} \log \frac{p(t)}{f(t)/Z} dt \\ &= \frac{\log Z}{Z} \int_{-\infty}^{u} f(t)dt = \log Z = \text{LHS of (3)}. \end{aligned}$$

**定理 3 の証明 (スケッチ)：** ここでは，本文中にすでに与えた閉形式解に対するそれぞれの導出のスケッチを示す．$f_{k,i}(t_{k,i}) := \mathcal{N}\left(t_{k,i}\,; \mu_{k,i}, \sigma_{k,i}^2\right)$ とおくと，

$$\mathcal{F}_{\mathrm{m}}(\boldsymbol{\theta}, q) = \mathcal{H}[q_{k,i}] + \mathbb{E}_{q_{k,i}(t_{k,i})}\left[\log f_{k,i}(t_{k,i})\right] + \mathrm{const} \tag{11}$$

のように整理できる．ただし，const は $q_{k,i}$ に依存しない項を表す．補助定理 4 より，Algorithm 1 の Line 5 の解が導かれる．

Line 7 の解は $\frac{\partial \mathcal{F}_{\mathrm{m}}(\boldsymbol{\theta}, q)}{\partial \boldsymbol{\theta}} = \boldsymbol{0}$ とおくことにより得られる． □

**定理 2 の証明 (スケッチ)：** 補助定理 4 を式 (11) に適用することにより，$\forall (k,i) \in \mathcal{E}_{\mathrm{h}}$ に対して $q_{k,i}^{\star}$ は切断正規分布の密度関数となることが導かれる． □

## 6. 数値実験

MTTM の性能を評価するため，3 個の水質データセット Indian, NY Top, NY Bottom を用いた．総大腸菌群数 (TC) と糞便性大腸菌群数 (FC) を検出限界のある 2 種類の病原体濃度と見なし，陰性率 10% として疑似的に打ち切りデータを用意した．データセット全体から無作為に選んだ $n = 100$ 個のサブセットを使った．これを 50 回繰り返して，それぞれで推定値の RMSE を算出した．各データセットにおける RMSE の分布を図 2 にプロットした．図中の STTM は，説明変量に検出限界値を代入して STTM を適用したときの結果である．3 個のデータセットいずれにおいても MTTM の RMSE が STTM の RMSE より顕著に低かった．よって，MTTM は打ち切られた複数の変量を補完するための有用な枠組みといえる．

**謝辞** 本研究の一部は，環境研究総合推進費により実施され，JSPS 科学研究費 22K04372 の助成を受けたものである．

## 参考文献

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

[2] 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.