

グループテストによる検査の誤り訂正と 検査数削減の理論とアルゴリズム

坂田 綾香[†]

統計数理研究所 数理・推論研究系

キーワード：グループテスト，確率伝搬法

1. はじめに

グループテストとは、患者から採取した検体を混ぜ合わせて検査する方法である¹⁾。混ぜ合わせた検体 (プール) 数は患者の数より少ないとするため、検査数は患者数よりも少ない数に抑制される。ただし検査結果から患者の状態を推定するための手続きが必要となる (図 1 (a))。有病率が十分小さい場合には、適切な推定法を用いれば、高い精度で患者の状態を推定できることが期待される。検査数の削減は検査コストの抑制につながるため、さまざまな疾患への適用が議論されてきた。たとえば HIV 検査²⁾ や肝炎ウイルス検査^{3, 4)} に対してグループテストの適用が議論されている。

グループテストの精度は、プールの作り方と推定方法に大きく依存する。プールの作り方については、ランダムに混ぜ合わせる方法⁵⁾ や、また任意の方法で作ったプールに対して検査を行い、陽性となったプールを分割していく方法も提案されている^{6, 7)}。実験の観点からは、ハイスループットスクリーニングにおけるアッセイのデザイン方法として shifted transversal design と呼ばれる方法が提案されている⁸⁾。また最近では能動学習に基づく方法も議論されている^{9, 10)}。

本稿ではグループテストにおける患者の状態推定について、数理的観点から考察する。グループテストにおける検査数は患者数よりも小さいため劣決定問題であるが、感染者数が十分小さいという仮定のもとではグループテストにより感染者が特定できる場合がある。特に、検査にエラーが含まれる場合の患者の状態推定について議論する。検査において全くエラーが起これないという状況は非現実的であり、また人為的ミスによるエラーも起こりうる。たとえば PCR 検査では、検査対象としない DNA が増えてしまったり、DNA の配列が異なるものに置き換わることによりエラーが生じる場合がある。本稿では、ベイズ推定の枠組みでエラーの統計性を考慮し、患者の状態を推定する。そのためのアルゴリズムとして確率伝搬法 (belief propagation, BP) を導入する^{5, 11)}。ただし患者の状態推定にベイズ推定を適用する場合、推論の拠り所となるのは事後分布であるが、これは連続的な確率分布である。その連続的な分布から、陽性/陰性という離散の状態を決める必要がある。このような連続値を出力する検査からの離散状態の推定のための、カットオフ値 (閾値) の決定法を紹介する。

2. グループテストの数学的定式化

ここでは、プールに含まれる患者数 K が全プールで一定、各患者が属すプール数 C も一定

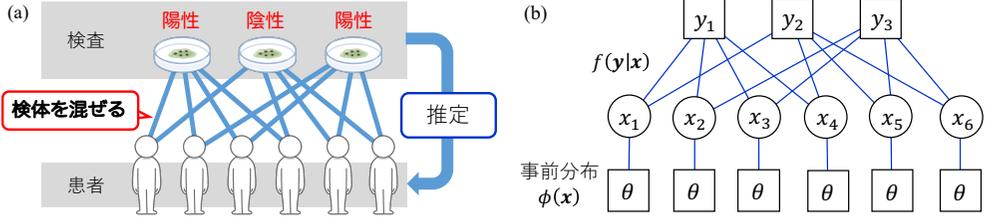


図 1. (a) グループテストの模式図. 実線は患者がどのプールに属するかを示している. (b) 推定の際に導入するモデルのグラフ表現. 患者の状態 \mathbf{x} , 検査の出力を \mathbf{y} として, \mathbf{y} の生成過程を関数 f でモデル化する. また事前分布 ϕ を導入する.

という条件を満たすランダムなプールを作る. ここで, K, C とも患者数 N に比べて十分小さいとする. N 人の患者集団に対して, $M (< N)$ 回のグループテストを行う. 図 1 は, $K = 4, M = 3, N = 6$ の例である. N 人の真の状態はベクトル $\mathbf{x}^{(0)} \in \{0, 1\}^N$ で表し, 患者 i が感染していれば $x_i^{(0)} = 1$, していなければ $x_i^{(0)} = 0$ とする. 検体の混ぜ方はプール行列 $F \in \{0, 1\}^{M \times N}$ により指定され, $F_{\mu i} = 1$ であれば患者 i がプール μ に含まれ, $F_{\mu i} = 0$ であれば含まれないことを意味する. プール μ の真の状態 $y_\mu^{(0)} \in \{0, 1\}$ は, プール内に一人でも感染者がいれば陽性 (1), 一人もいなければ陰性 (0) とする. N 成分の論理和を $\bigvee_{i=1}^N f_i = f_1 \vee \dots \vee f_N$ とすると $y_\mu^{(0)}(\mathbf{x}^{(0)}) = \bigvee_{i=1}^N F_{\mu i} x_i^{(0)}$ となる.

以下では, 全ての検査において誤りが同じ確率で独立に起こるとし, 真陽性確率 p_{TP} と偽陽性確率 p_{FP} を用いて検査をモデル化する¹¹⁾. プール μ に対する検査結果 y_μ の, 患者の状態に関する条件付き分布は次のように与えられる.

$$(2.1) \quad f^{(0)}(y_\mu | \mathbf{x}) = \left((1 - p_{\text{TP}})^{y_\mu^{(0)}(\mathbf{x})} (1 - p_{\text{FP}})^{(1 - y_\mu^{(0)}(\mathbf{x}))} \right)^{1 - y_\mu} \left(p_{\text{TP}}^{y_\mu^{(0)}(\mathbf{x})} p_{\text{FP}}^{(1 - y_\mu^{(0)}(\mathbf{x}))} \right)^{y_\mu}$$

すなわち, 検査データ \mathbf{y} は (2.1) に従って生成されていると考える.

グループテストにおける推定の目標は, 検査結果 $\mathbf{y} = [y_1, \dots, y_M]^T$ から患者の真の状態 $\mathbf{x}^{(0)}$ を特定することである. ここでは, 患者の状態をモデルパラメータとして検査結果 \mathbf{y} を記述するという考え方から, 尤度 $f(\mathbf{y} | \mathbf{x})$ を導入する. 事前分布については, 全患者の検査前事前確率を有病率 θ として, 次のように与える.

$$(2.2) \quad \phi(\mathbf{x}) = \prod_{i=1}^N \{(1 - \theta)(1 - x_i) + \theta x_i\}$$

ベイズの定理より, 事後分布は次のように与えられる.

$$(2.3) \quad P(\mathbf{x} | \mathbf{y}) = \frac{1}{Z(\mathbf{y})} f(\mathbf{y} | \mathbf{x}) \phi(\mathbf{x}), \quad Z(\mathbf{y}) = \sum_{\mathbf{x} \in \{0, 1\}^N} f(\mathbf{y} | \mathbf{x}) \phi(\mathbf{x})$$

本稿では, 事後分布から評価した周辺事後分布を診断変数として患者の状態を決定する. その周辺事後分布の評価において, 確率伝搬法と呼ばれるアルゴリズムを導入する.

3. 確率伝搬法による推定

確率伝搬法は, 近似的に周辺事後分布を求めるアルゴリズムであり, グループテストにおけるベイズ推定にも応用されている^{5, 11-13)}. 確率伝搬法は, 確率変数の依存関係により確率分布を表現する, グラフィカルモデルの上で定義される. 図 1(a) のグループテストに対応するグラ

フィカルモデルを図 1(b) に示す. 変数ノード (○) と因子ノード (□) はそれぞれ患者の状態とプールに対する検査結果を示す. グラフのエッジはプールの作り方を意味し, 1 番目のプールに 1 番, 2 番, 3 番, 4 番の患者が属するので $F_{1,1} = F_{1,2} = F_{1,3} = F_{1,4} = 1$, それ以外の成分はゼロであり, 1 番目のプールの真の状態は $y_1^{(0)} = x_1^{(0)} \vee x_2^{(0)} \vee x_3^{(0)} \vee x_4^{(0)}$ である. 因子ノードと変数ノードの次数はそれぞれ K, C に対応する. 確率伝搬法では, 図 1(b) のグラフが局所的にツリーであるとの仮定のもと, エッジ上で二種類のメッセージ $p_{i \rightarrow \nu}(x_i)$ と $\tilde{p}_{\nu \rightarrow i}(x_i)$ を定義する. ここで i と ν はそれぞれ変数ノードと因子ノードのラベルである. 直観的に, これらのメッセージは ν 回目のテストを行う前と行った後の周辺事後確率を表す. 確率伝搬法は次のメッセージの更新から定義される.

$$(3.1) \quad \tilde{p}_{\nu \rightarrow i}(x_i) \propto \sum_{\mathbf{x} \setminus x_i} f(y_\nu | \mathbf{x}) \prod_{j \in \mathcal{L}(\nu) \setminus i} p_{j \rightarrow \nu}(x_j)$$

$$(3.2) \quad p_{i \rightarrow \nu}(x_i) \propto \phi(x_i) \prod_{\eta \in \mathcal{G}(i) \setminus \nu} \tilde{p}_{\eta \rightarrow i}(x_i),$$

ここで $\mathcal{L}(\nu)$ と $\mathcal{G}(i)$ は, それぞれ ν 番目のプールに属す患者の番号と i 番目の患者が属すプールの番号の集合を意味する. これらのメッセージを用いると, 周辺事後確率は次のように与えられる.

$$(3.3) \quad p_i(x_i) \propto \phi(x_i) \prod_{\eta \in \mathcal{G}(i)} \tilde{p}_{\eta \rightarrow i}(x_i).$$

グラフが真にツリーの場合, ツリーの端から端へとメッセージを一度伝搬させることで厳密な周辺分布 (3.3) を得るが, 一般のグラフについては, メッセージを再帰的に更新する必要がある. そこで, 以下では t 回目の更新のメッセージを $\tilde{p}_{\nu \rightarrow i}^{(t)}, p_{i \rightarrow \nu}^{(t)}$ と表現する. グループテストの場合, ランダムプールにおいて $N, M \rightarrow \infty$ また $\alpha = M/N \sim O(1)$ の極限で確率伝搬法は妥当な推定結果を与えることが示されている¹³⁾.

グループテストの場合, 推定される変数 x_i は二値をとる. したがって, メッセージは次のように $[0, 1]$ のベルヌーイ変数 $\{m_{j \rightarrow \nu}^{(t)}\}$ と $\{\tilde{m}_{\nu \rightarrow i}^{(t)}\}$ を用いて表現される.

$$(3.4) \quad p_{j \rightarrow \nu}^{(t)}(x_i) = (1 - m_{j \rightarrow \nu}^{(t)})(1 - x_i) + m_{j \rightarrow \nu}^{(t)} x_i$$

$$(3.5) \quad \tilde{p}_{\nu \rightarrow i}^{(t)}(x_i) = (1 - \tilde{m}_{\nu \rightarrow i}^{(t)})(1 - x_i) + \tilde{m}_{\nu \rightarrow i}^{(t)} x_i.$$

(3.1) - (3.2) から, ベルヌーイ変数に対する更新式は次のように得られる.

$$(3.6) \quad m_{j \rightarrow \nu}^{(t)} = \frac{\theta \prod_{\eta \in \mathcal{G}(i) \setminus \nu} \tilde{m}_{\eta \rightarrow i}^{(t-1)}}{(1 - \theta) \prod_{\eta \in \mathcal{G}(i) \setminus \nu} (1 - \tilde{m}_{\eta \rightarrow i}^{(t-1)}) + \theta \prod_{\eta \in \mathcal{G}(i) \setminus \nu} \tilde{m}_{\eta \rightarrow i}^{(t-1)}}$$

$$(3.7) \quad \tilde{m}_{\nu \rightarrow i}^{(t)} = \frac{U_\nu}{U_\nu \left\{ 2 - \prod_{j \in \mathcal{L}(\nu) \setminus i} (1 - m_{j \rightarrow \nu}^{(t-1)}) \right\} + W_\nu \prod_{j \in \mathcal{L}(\nu) \setminus i} (1 - m_{j \rightarrow \nu}^{(t-1)})}$$

ここで次のように定義した.

$$(3.8) \quad U_\nu = p_{TP} y_\nu + (1 - p_{TP})(1 - y_\nu), \quad W_\nu = p_{FP} y_\nu + (1 - p_{FP})(1 - y_\nu)$$

周辺事後確率 (3.3) もまた, ベルヌーイ変数により表される. これを m_i とすると, メッセージに対するベルヌーイ変数を用いて次のように与えられる.

$$(3.9) \quad m_i = \frac{\theta \prod_{\eta \in \mathcal{G}(i)} \tilde{m}_{\eta \rightarrow i}}{(1 - \theta) \prod_{\eta \in \mathcal{G}(i)} (1 - \tilde{m}_{\eta \rightarrow i}) + \theta \prod_{\eta \in \mathcal{G}(i)} \tilde{m}_{\eta \rightarrow i}}$$

4. カットオフ値の決め方

ベイズ推定および確率伝搬法により与えられる周辺事後確率 m_i は連続値である。一方で、グループテストの目標は患者の状態（陽性/陰性）という離散的な値を決めることである。連続値から離散値を決める際に、カットオフ値を適切に決める必要がある。ここでは、真のモデル $f^{(0)}(\mathbf{y}|\mathbf{x})$ と仮定したモデル $f(\mathbf{y}|\mathbf{x})$ が一致するという設定のもと、ベイズ決定理論に基づきカットオフを決める方法を紹介する。

カットオフ値を決める際に基本となる指標は、真陽性率（感度）と真陰性率（特異度）である。カットオフ値に応じて偽陽性と偽陰性が生じ、カットオフ値が大きくなれば偽陽性率（ $= 1 - \text{真陰性率}$ ）は下がり偽陰性率（ $= 1 - \text{真陽性率}$ ）は上がり、またカットオフ値が小さくなれば偽陰性率は下がり偽陽性率が上がるというトレードオフがある。ベイズ推定を用いたグループテストにおいては、最大周辺事後確率（Maximum Posterior Marginal, MPM）推定量により患者の状態を決定する先行研究がある¹¹⁾。MPM 推定量はカットオフ値を 0.5 とした場合に対応する。グループテストに限らず、MPM 推定量は特定の条件下では真値と推定値の平均二乗誤差を最小化する推定量として知られるため、広く用いられている¹⁴⁾。

ここでは、ユーティリティ関数 U を次のように定義し、ユーティリティ関数の最大化法によりカットオフ値を決定する¹⁵⁻¹⁷⁾。

$$(4.1) \quad \begin{aligned} U(\mathbf{x}^{(0)}, \hat{\mathbf{x}}(\mathbf{y}); \mathbf{u}) &= \theta(u_{\text{TP}}\text{TP} + u_{\text{FN}}\text{FN}) + (1 - \theta)\text{FP}u_{\text{FP}} + (1 - \theta)\text{TN}u_{\text{TN}} \\ &= \theta u_{\text{TP}} + (1 - \theta)u_{\text{FP}} + \theta(u_{\text{FN}} - u_{\text{TP}})\text{FN} + (1 - \theta)(u_{\text{FP}} - u_{\text{TN}})\text{FP} \end{aligned}$$

TP, FN, FP, TN はそれぞれ真陽性率、偽陰性率、偽陽性率、真陰性率であり、また $\text{TP} = 1 - \text{FN}$, $\text{TN} = 1 - \text{FP}$ である。(4.1) の係数 $\mathbf{u} = \{u_{\text{TP}}, u_{\text{FN}}, u_{\text{FP}}, u_{\text{TN}}\}$ はユーティリティと呼ばれ、それぞれ真陽性、偽陰性、偽陽性、真陰性がもたらす影響を表す。一般に、起こってほしくない事象のユーティリティを小さく、起こってほしい事象のユーティリティを大きく設定して、ユーティリティ関数を最大化するカットオフ値を求める。したがって $u_{\text{TP}} > u_{\text{FN}}$, $u_{\text{TN}} > u_{\text{FP}}$ とする。カットオフ値はユーティリティに依存するため、ユーティリティを適切に設定することが重要である¹⁷⁻¹⁹⁾。

(4.1) の第一項、第二項はモデルの設定から決まる定数であるので、便利のため (4.1) の第 3 項と第 4 項からリスク関数 $R(\mathbf{x}^{(0)}, \hat{\mathbf{x}}(\mathbf{y}); \boldsymbol{\lambda})$ を次のように定義する。

$$(4.2) \quad R(\mathbf{x}^{(0)}, \hat{\mathbf{x}}(\mathbf{y}); \boldsymbol{\lambda}) = \lambda_{\text{FN}}\text{FN}(\mathbf{x}^{(0)}, \hat{\mathbf{x}}(\mathbf{y})) + \lambda_{\text{FP}}\text{FP}(\mathbf{x}^{(0)}, \hat{\mathbf{x}}(\mathbf{y}))$$

$\boldsymbol{\lambda} = \{\lambda_{\text{FN}}, \lambda_{\text{FP}}\}$ はパラメータであり、 $\lambda_{\text{FN}} = -\theta(u_{\text{FN}} - u_{\text{TP}}) > 0$, $\lambda_{\text{FP}} = -(1 - \theta)(u_{\text{FP}} - u_{\text{TN}}) > 0$ とした。また $\text{FN}(\mathbf{x}^{(0)}, \hat{\mathbf{x}}(\mathbf{y}))$ は推定値 $\hat{\mathbf{x}}(\mathbf{y})$ と真の状態 $\mathbf{x}^{(0)}$ から決まる真陰性率の意である。 $\text{FN}(\mathbf{x}^{(0)}, \hat{\mathbf{x}}(\mathbf{y}))$ も同様である。ユーティリティ関数の最大化はリスク関数の最小化と等価であるため、以下ではリスク関数を最小化する推定値 $\hat{\mathbf{x}}(\mathbf{y})$ を考える。あらゆる \mathbf{y} についてリスク関数 $\hat{R}(\mathbf{x}^{(0)}, \hat{\mathbf{x}}(\mathbf{y}); \boldsymbol{\lambda})$ を最小化する推定量を $\hat{\mathbf{x}}^*(\mathbf{y})$ と表記すると、仮定したモデルが真のモデルに一致する際、各成分は次のように与えられる²⁰⁾。

$$(4.3) \quad \hat{x}_i^*(\mathbf{y}) = \mathbb{I} \left(\rho_i(\mathbf{y}) > \frac{\theta \lambda_{\text{FP}}}{\lambda_{\text{FN}}(1 - \theta) + \theta \lambda_{\text{FP}}} \right)$$

(4.3) を用いて、カットオフ値 0.5 に対応する MPM 推定量について考察してみよう。カットオフ値が 0.5 となるのは $\lambda_{\text{FN}} = \theta$, $\lambda_{\text{FP}} = 1 - \theta$ のときである。つまり、有病率 θ が 0.5 以下の時は偽陽性を減らすことを優先し、0.5 以上の時は偽陰性を減らすことを優先する。言い換えると陽性/陰性集団について、大きい集団の方の属性を誤らないことを重要視していることが

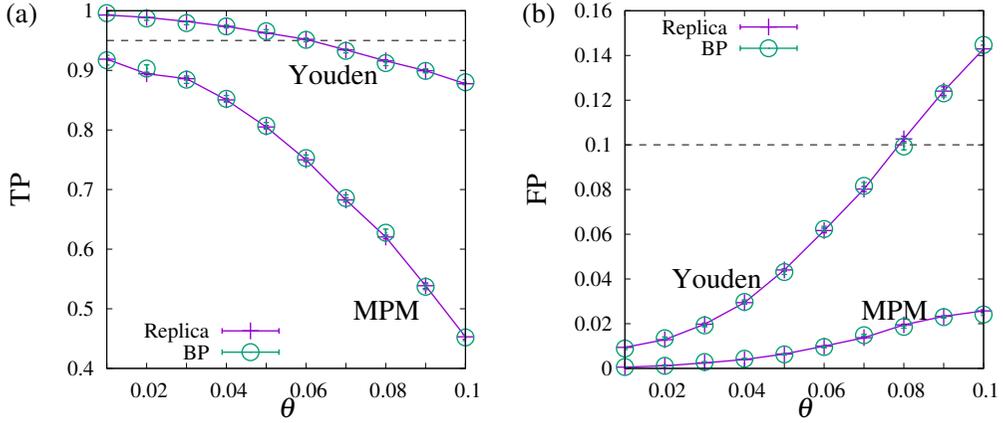


図 2. $N = 1000$, $\alpha = 0.5$, $K = 10$, $p_{TP} = 0.95$, $p_{FP} = 0.1$ における最適カットオフ値を用いた場合の TP と FP. ‘BP’ は確率伝搬法による結果を示す. ‘Replica’ は理論的に導出される解を示す²⁰⁾.

わかる. 一般にグループテストは有病率が小さいときに有効であるため, グループテストにおいて MPM 推定値を用いるということは偽陽性を優先的に減らすことを意味する. 一方で, もともと偽陽性率が低い検査へグループテストを適用する場合など, 有病率に関わらず偽陰性を優先的に減らしたいという需要もあるだろう. このような場合に MPM 推定値を用いるのは不適切であると言える.

別のカットオフ値を与える指標として, 次のように与えられる Youden index の最大化を考察してみよう.

$$(4.4) \quad J_Y(\mathbf{x}^{(0)}, \mathbf{y}) = TP(\mathbf{x}^{(0)}, \hat{\mathbf{x}}(\mathbf{y})) - FP(\mathbf{x}^{(0)}, \hat{\mathbf{x}}(\mathbf{y}))$$

リスクの定義より $J_Y(\mathbf{x}^{(0)}, \mathbf{y}) = 1 - R(\mathbf{x}^{(0)}, \mathbf{y}; \lambda_{FN} = 1/2, \lambda_{FP} = 1/2)$ が成立することから, Youden index 最大化によるカットオフ値決定は, 偽陽性, 偽陰性を同程度に減らしたいという要請に基づくことがわかる. (4.3) に $\lambda_{FN} = \lambda_{FP} = 1/2$ を代入するとカットオフ値は θ となり, 事後 Youden index $\hat{J}_Y(\hat{\mathbf{x}}(\mathbf{y})) = \hat{R}(\hat{\mathbf{x}}(\mathbf{y}); \lambda_{FN} = 1/2, \lambda_{FP} = 1/2)$ を最大化するには, カットオフ値を有病率 θ に設定すれば良いことがわかる.

図 2 は, Youden index 最大化という指標のもとでの最適カットオフ値を用いた場合の TP, FP と MPM 推定値を用いた場合の TP, FP の比較である. 点線は元々の検査特性であり, $TP > p_{TP}$ であれば偽陰性が修正されており, $FP < p_{FP}$ であれば偽陽性が修正されていることを意味する. 4 章で見たように, MPM 推定値は FP を下げる傾向にあり, そのトレードオフとして TP は Youden index 最大化を行なった場合よりも低い. Youden index 最大化により TP を上げることと FP を下げることと同じ重みで重視したことにより, $\theta < 0.06$ において偽陰性が修正可能であることが示されている.

5. まとめと今後の展望

本稿では, 検査エラーが確率的に起こる場合のグループテストに対するベイズ推定および確率伝搬法を紹介した. またベイズリスクを最小化するカットオフ値を λ_{FN} , λ_{FP} の関数として示した. このカットオフ値の表現から, グループテストが有効な問題設定においては MPM 推定値が偽陽性を優先的に抑えることが示された. また Youden index 最大化という目的のもと

ではカットオフ値を有病率 θ に設定することが適切であることがわかった。本稿では触れなかったが、仮定したモデルと真のモデルが一致する状況下では、解析的に ROC 解析や AUC の評価を行うことも可能である²⁰⁾。

今後の展望としては、検体の混ぜ合わせにより生じる希釈効果や、検査の非独立性など現実的な設定を取り込んだモデルを考慮することが必要であると考えている。

参考文献

- 1) Dorfman R 1943 *Annals of Mathematical Statistics* **14** 436–440
- 2) Krajden M, Cook D, Mak A, Chu K, Chahil N, Steinberga M, Rekartb M and Gilberta M 2014 *Journal of Clinical Virology* **61** 132–137
- 3) Kleinman S H, Strong D M, Tegtmeier G G E, Holland P V, Gorlin J B, Cousins C R, Chiacchierini R P and Pietrelli L A 2005 *Transfusion* **45** 1247–1257
- 4) Sarov B, Novack L, Beer N, Safi J, Soliman H, Pliskin J S, Litvak E, Yaari A and Shinar E 2007 *Transfusion Medicine* **17** 479–487
- 5) Mézard M, Tarzia M and Toninelli C 2008 *Journal of Physics: Conference Series* **95** 012019
- 6) Sobel M and Groll P A 1959 *Bell Labs Technical Journal* **38** 1179–1252
- 7) Hwang K F 1972 *Journal of the American Statistical Association* **67** 605–608
- 8) Thierry-Mieg N 2006 *BMC Bioinformatics* **7** 1–13
- 9) Cuturi M, Teboul O, Berthet Q, Doucet A and Vert J P 2020 Noisy adaptive group testing using bayesian sequential experimental design <https://arxiv.org/abs/2004.12508>
- 10) Sakata A 2021 *Physical Review E* **103** 022110
- 11) Sejdinovic D and Johnson O 2010 Note on noisy group testing: Asymptotic bounds and belief propagation reconstruction *48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)* (IEEE) pp 998 – 1003
- 12) Kanamori T, Uehara H and Jimbo M 2012 *Journal of Statistical Theory and Practice* **6** 220–238
- 13) Sakata A 2020 *Journal of Physical Society Japan* **89** 084001
- 14) Lehmann E L and Casella G 1998 *Theory of Point Estimation* (Springer (New York))
- 15) McNeil B J and Adelstein S J 1976 *Journal of Nuclear Medicine* **17** 439–48
- 16) Somoza E and Mossman D 1991 *Biological Psychiatry* **29** 811–826
- 17) McFall R M and Treat T A 1999 *Annual Review of Psychology* **50** 215–41
- 18) Metz C E, Goodenough D J and Rossman K 1973 *Radiology* **109** 297–304
- 19) Mossman D and Somoza E 1989 *Archives of general psychiatry* **46** 653–660
- 20) Sakata A and Kabashima Y Decision theoretic cutoff and roc analysis for bayesian optimal group testing arXiv:2110.10877

† 坂田 綾香

〒190-8562 東京都立川市緑町 10-3 統計数理研究所 数理推論研究系
ayaka@ism.ac.jp

本研究に関して利益相反に該当する事項はない。