データ駆動型科学のための選択的推論（Selective Inference）

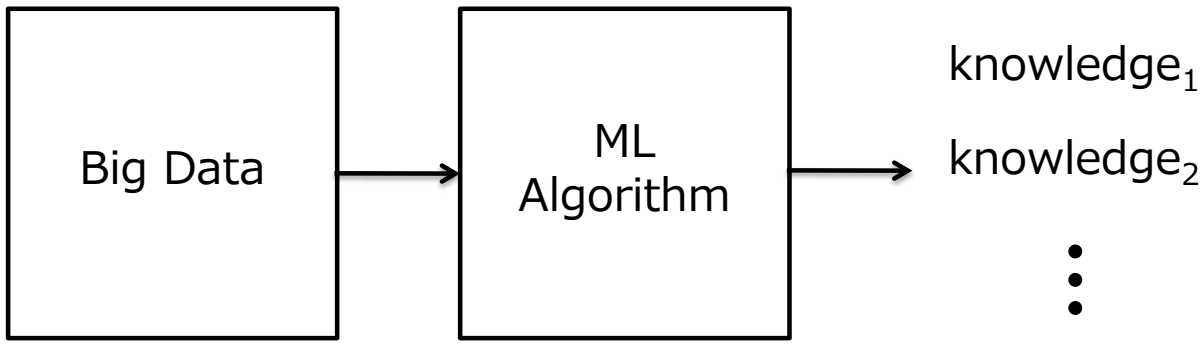
竹内一郎

名古屋工業大学／理化学研究所

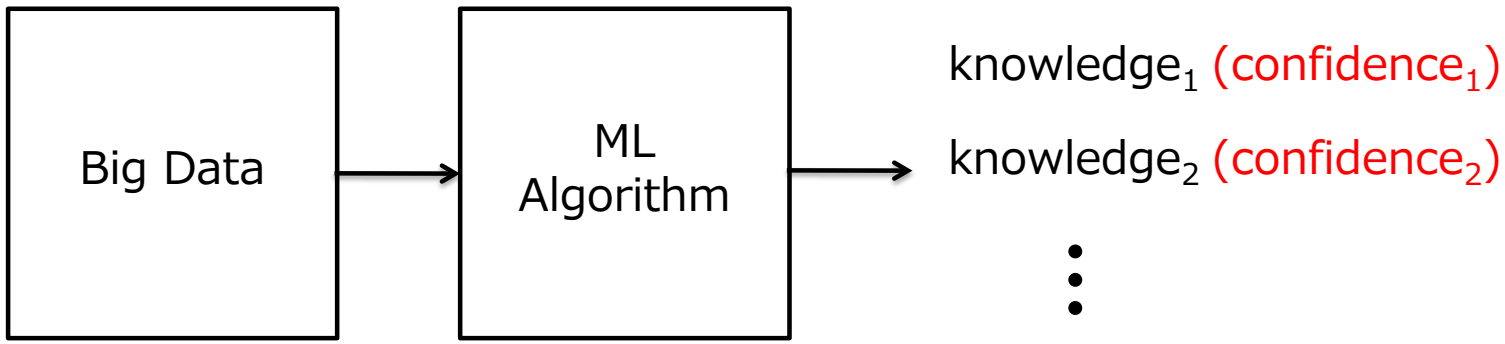
自己紹介

- 所属：名古屋工業大学・情報工学専攻
- 兼任：理化学研究所・革新知能統合研究センター
- 専門：機械学習
- プロジェクト：データ駆動型の生物科学，医療科学，材料科学
- 研究：データ駆動型科学の理論と実践

データ駆動型 *AI*（機械学習）の品質保証



データ駆動型 AI（機械学習）の品質保証



生命科学分野の例題

- 1 万個の遺伝子 $x_1, x_2, \ldots, x_{10000}$ と薬剤効果 y の関係を調べたい
- データ（患者数 n，特徴数 d）

$$\underset{n\times 1}{\boldsymbol{y}} = \begin{bmatrix} 19.5 \\ 18.6 \\ \vdots \\ \vdots \\ \vdots \\ 22.6 \end{bmatrix} \quad \underset{n\times d}{X} = \begin{bmatrix} 0.9 & 0.5 & \cdots & \cdots & \cdots & 0.8 \\ 0.5 & 0.6 & \cdots & \cdots & \cdots & 1.0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0.8 & 0.4 & \cdots & \cdots & \cdots & 0.3 \end{bmatrix}$$

DR

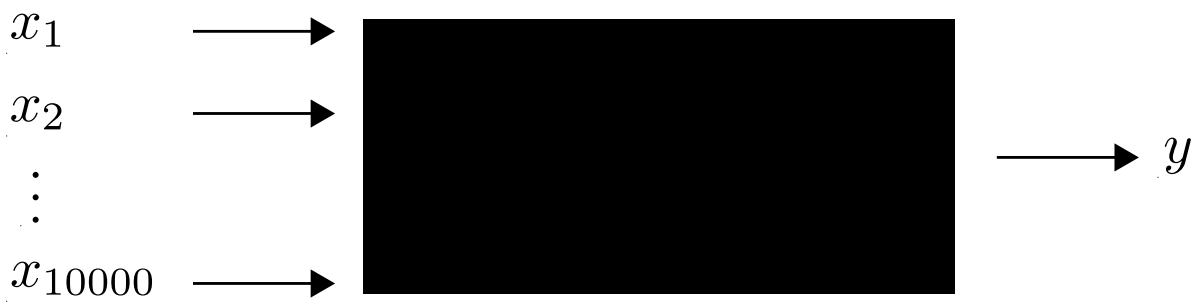
GF 1　GF 2　GF 10000

- 高次元線形モデル

$$y = w_1 x_1 + w_2 x_2 + \ldots + w_{10000} x_{10000}$$

２つの目的：予測と理解

- 予測：新しい患者の薬剤効果を予測したい

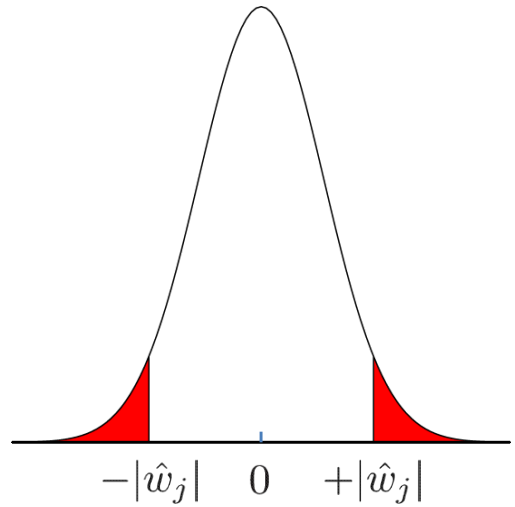


- 理解：どの遺伝子が薬剤効果に影響を与えるかを理解したい

$$\hat{w}_2 = +3.2 \Rightarrow \text{Gene 2 has positive effect}$$
$$\hat{w}_4 = -0.1 \Rightarrow \text{Gene 4 has no effect}$$
$$\hat{w}_5 = -2.5 \Rightarrow \text{Gene 5 has negative effect}$$

統計的仮説検定による品質保証

- 真値が $w_j = 0$（帰無仮説）のとき推定値が $\hat{w}_j$ より極端になる確率を計算



- p 値（p-value）：帰無仮説のもとでは起こりえない度合い

$$p_j = \mathbb{P}(|\hat{w}_j^{\text{observed}}| < |\hat{w}_j|)$$

- 例（有意水準 $\alpha = 0.05$）

$$\hat{w}_2 = +2.5 \;\Rightarrow p_2 = 0.012 < \alpha \;\Rightarrow\; \text{Gene 2 has positive effect}$$
$$\hat{w}_4 = -0.1 \;\Rightarrow p_4 = 0.920 \geq \alpha \;\Rightarrow\; \text{Gene 4 has no effect}$$
$$\hat{w}_5 = -2.0 \;\Rightarrow p_5 = 0.045 < \alpha \;\Rightarrow\; \text{Gene 5 has negative effect}$$

特徴選択

- 特徴選択

- 選択された特徴の線形モデル

$$y = w_2 x_2 + w_{15} x_{15} + w_{321} x_{321}$$

- 統計的仮説検定による品質保証

$$\begin{aligned}
\hat{w}_2 &= +3.8 \Rightarrow p_2 < \alpha \Rightarrow \text{Gene 2 has positive effect} \\
\hat{w}_{15} &= -4.5 \Rightarrow p_{15} < \alpha \Rightarrow \text{Gene 15 has negative effect} \\
\hat{w}_{321} &= -6.2 \Rightarrow p_{321} < \alpha \Rightarrow \text{Gene 321 has negative effect}
\end{aligned}$$

特徴選択

- 特徴選択

- 選択された特徴の線形モデル

$$y = w_2 x_2 + w_{15} x_{15} + w_{321} x_{321}$$

- 統計的仮説検定による品質保証 $\Rightarrow$ 選択バイアスの補正が必要！

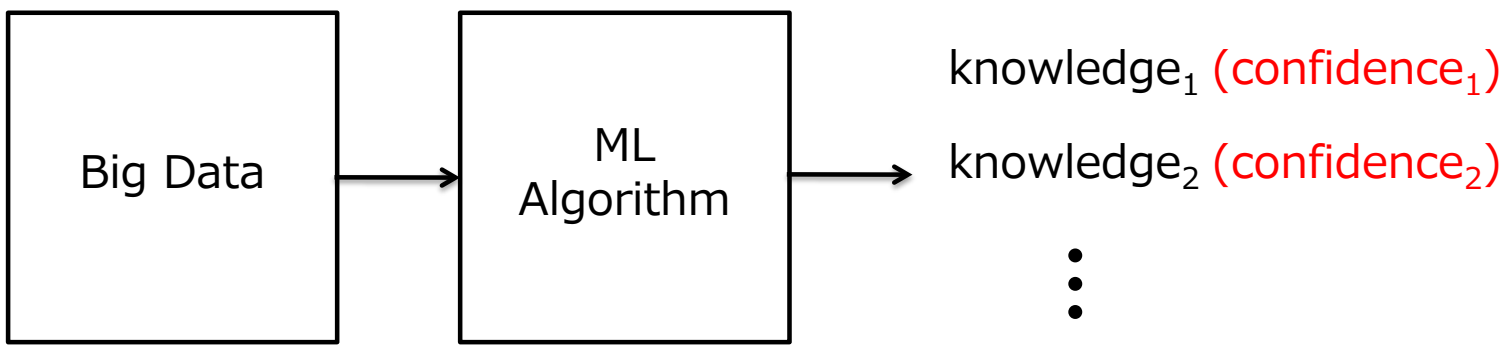
$$\hat{w}_2 = +3.8 \Rightarrow p_2 < \alpha \Rightarrow \text{Gene 2 has positive effect}$$
$$\hat{w}_{15} = -4.5 \Rightarrow p_{15} < \alpha \Rightarrow \text{Gene 15 has negative effect}$$
$$\hat{w}_{321} = -6.2 \Rightarrow p_{321} < \alpha \Rightarrow \text{Gene 321 has negative effect}$$

本講演の概要

- 本講演のメイントピック

  選択的推論：**Selective Inference / Post-Selection Inference**

- 本講演の構成
  - Part1: 統計的仮説検定と選択バイアス
  - Part2: 線形モデルにおける選択的推論
  - Part3: 教師なし学習における選択的推論



Part1

統計的仮説検定と選択バイアス

出版バイアス（*Publication Bias*）

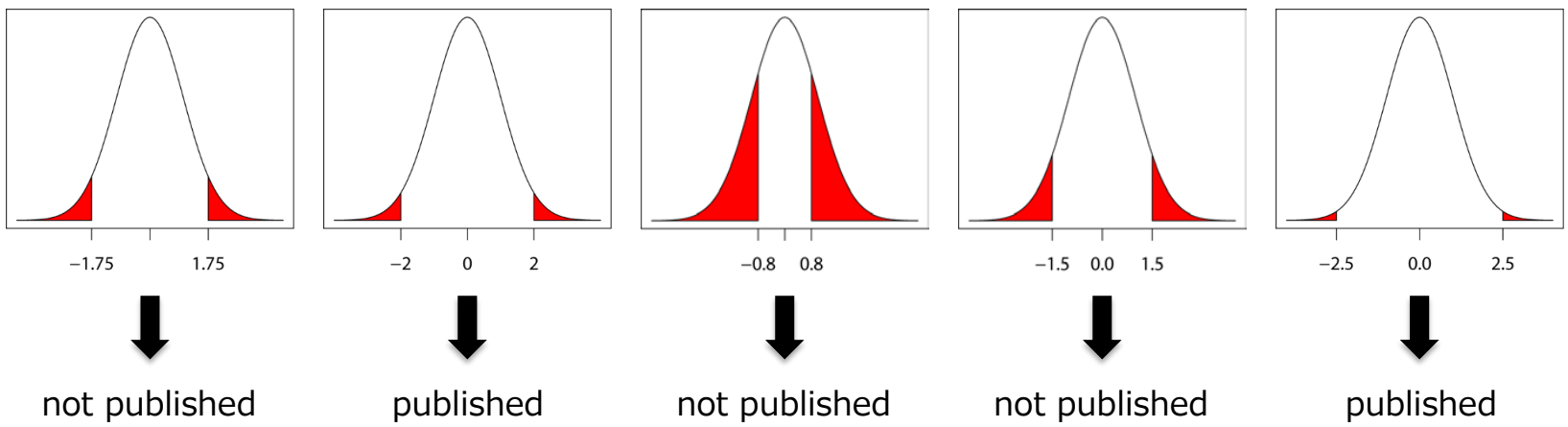
- 研究者１：遺伝子 A の研究
- 研究者２：遺伝子 B の研究
- 研究者３：遺伝子 C の研究
- ...

出版バイアス

出典: フリー百科事典『ウィキペディア（Wikipedia）』

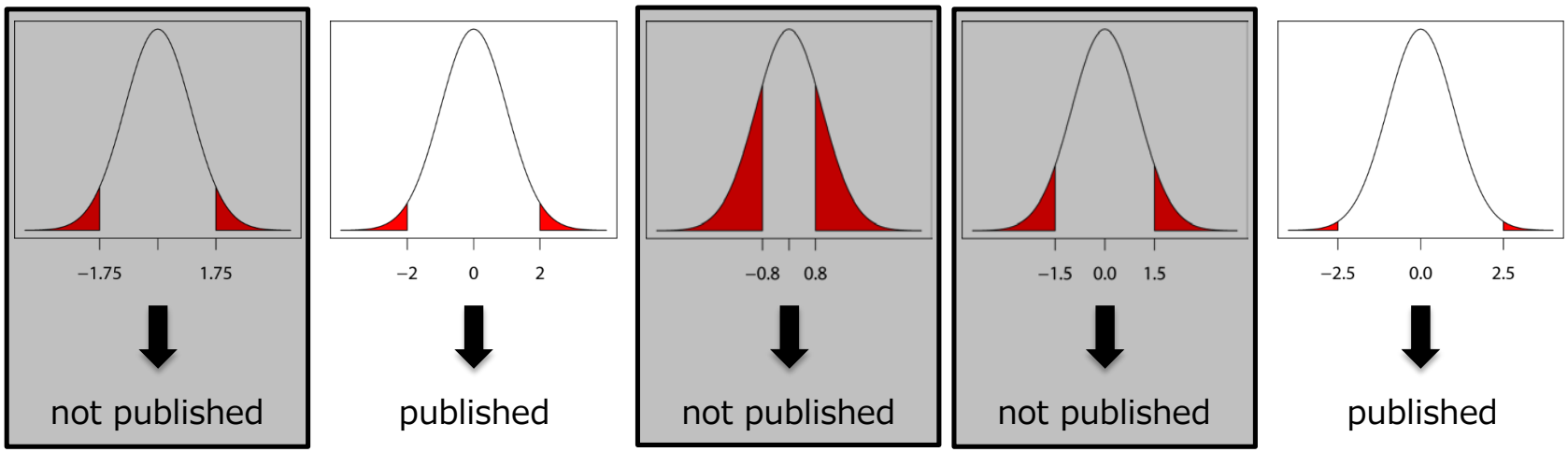
出版バイアス（publication bias）とは、否定的な結果が出た研究は、肯定的な結果が出た研究に比べて公表されにくいというバイアス（偏り）である[1]。**公表バイアス**とも言う。

Wikipediaより転載



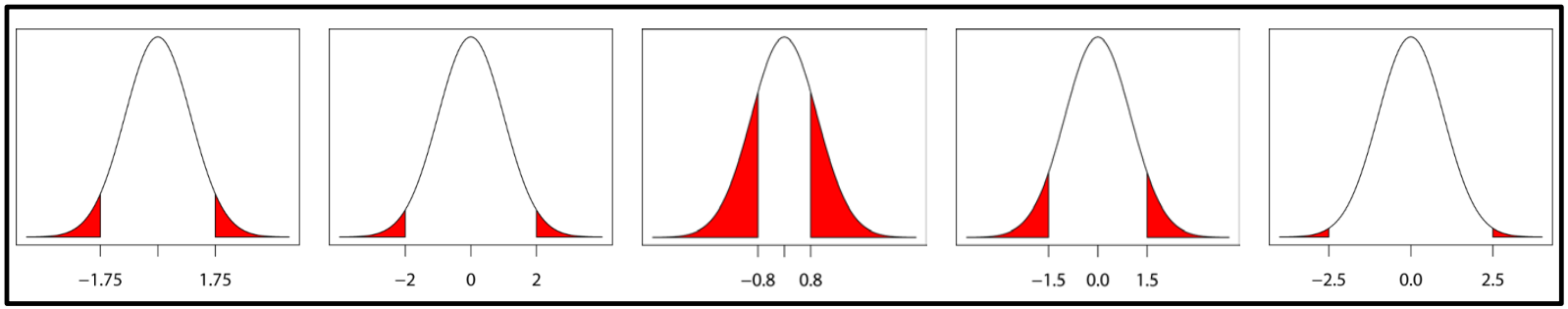
出版バイアスに対する読者の視点

- 読者は出版された論文しか読まない!



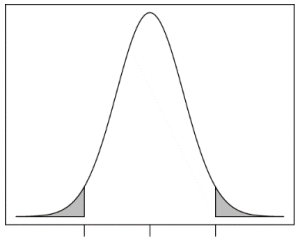
- 論文の読者はどのように結果を信頼すればよいのか？
- 出版された論文のうち，誤発見の確率を $\alpha = 0.05$ 未満にしたい！

仮説選択アルゴリズム



Hypothesis Selection Algorithm $\mathcal{A} : \mathcal{D} \to \mathcal{S}$

$(\mathcal{D} : \text{Data},\ \mathcal{S} : \text{Selected Hypotheses})$



$$\mathcal{A} : \{\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_k\} \mapsto \mathcal{S} := \{k \in [1, \ldots, d] \mid |\hat{w}_k| > 1.96\}$$

出版バイアスの（読者目線からの）補正

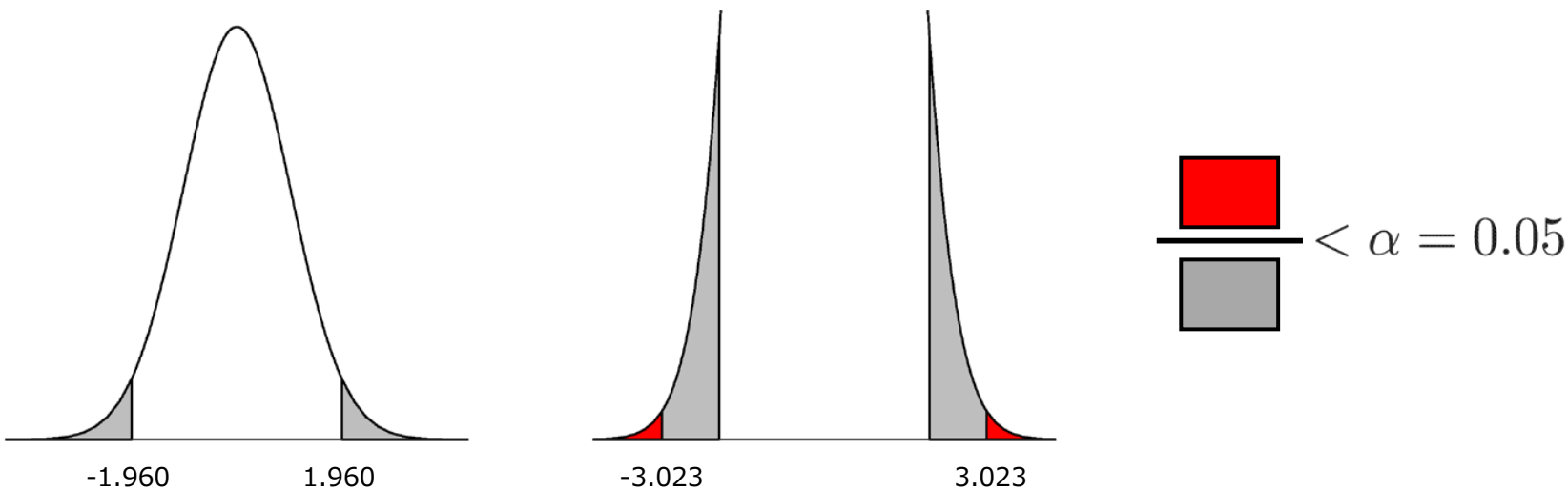
- 仮説選択アルゴリズム $\mathcal{A}$

$$\mathcal{A} : \hat{\boldsymbol{w}} = [\hat{w}_1, \ldots, \hat{w}_d]^\top \mapsto \mathcal{S}$$

- 出版された結果における誤検出率を $\alpha = 0.05$ 未満にするには

$$\frac{\mathbb{P}(|\hat{w}_k| > \theta, k \in \mathcal{A}(\hat{\boldsymbol{w}}))}{\mathbb{P}(k \in \mathcal{A}(\hat{\boldsymbol{w}}))} < \alpha$$

$$\Leftrightarrow \ \mathbb{P}(|\hat{w}_k| > \theta \mid k \in \mathcal{A}(\hat{\boldsymbol{w}})) < \alpha \ \Leftrightarrow \ \theta = 3.023$$



条件付確率に基づく統計的推測

- アルゴリズムによる仮説選択バイアスを補正するため条件付確率に基づく統計的推測を行えばよい

  Find $[\ell_\alpha^h, u_\alpha^h]$ such that $\mathbb{P}(s_h(\text{data}) \notin [\ell_\alpha^h, u_\alpha^h] \mid h \leftarrow \mathcal{A}(\text{data})) < \alpha$

  - $s_h(\text{data})$：選択された仮説の検定統計量
  - $[\ell_\alpha^h, u_\alpha^h]$：採択域
  - $\mathcal{A}$：仮説を選択，生成するアルゴリズム（機械学習）
  - data：データ
  - α：有意水準（e.g., 0.05）

線形モデル：データ例（再掲）

$$\underset{n\times 1}{\boldsymbol{y}} = \begin{bmatrix} 19.5 \\ 18.6 \\ \vdots \\ \vdots \\ \vdots \\ 22.6 \end{bmatrix} \quad \underset{n\times d}{X} = \begin{bmatrix} 0.9 & 0.5 & \cdots & \cdots & \cdots & 0.8 \\ 0.5 & 0.6 & \cdots & \cdots & \cdots & 1.0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0.8 & 0.4 & \cdots & \cdots & \cdots & 0.3 \end{bmatrix}$$

DR

GF 1　GF 2　GF 10000

線形モデルの統計的推測（特徴選択なし）

- 正規線形モデル

$$\boldsymbol{y} = X\boldsymbol{w} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 I)$$

- 最小二乗解

$$\hat{\boldsymbol{w}} := (X^\top X)^{-1} X^\top \boldsymbol{y} \sim N(\boldsymbol{w}, \sigma^2 (X^\top X)^{-1})$$

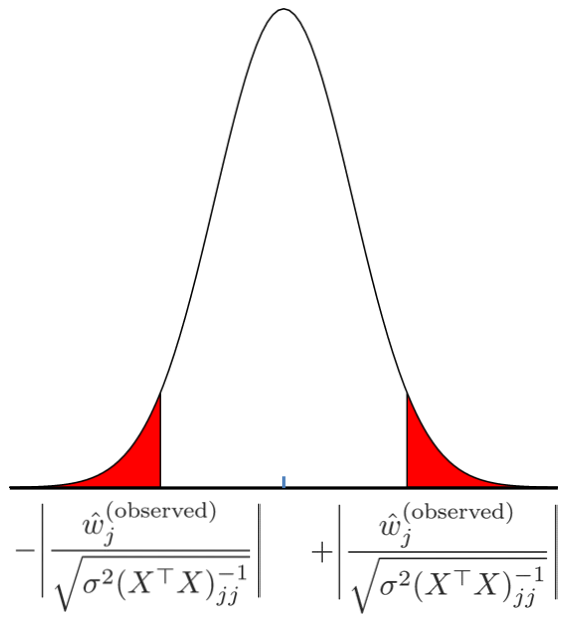
- 統計的仮説検定

$$H_0^j : w_j = 0 \quad \text{v.s.} \quad H_1^j : w_j \neq 0$$

- 誤検出率（p-value）

$$p_j = 2\Phi\left(-\left| \frac{\hat{w}_j^{(\text{observed})}}{\sqrt{\sigma^2 (X^\top X)^{-1}_{jj}}} \right| \right),$$

ただし，Φ は $N(0,1)$ の累積分布関数



線形モデルの統計的推測（特徴選択あり）

- 特徴選択アルゴリズム FS

$$\mathcal{S} \leftarrow \mathrm{FS}(X, \boldsymbol{y})$$

ただし，$\mathcal{S} \subseteq \{1, \ldots, d\}$ は選択された特徴の集合

- 選択された特徴を用いた正規線形モデル

$$\boldsymbol{y} = X\boldsymbol{w}_{\mathcal{S}} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 I)$$

- 最小二乗法

$$\hat{\boldsymbol{w}}_{\mathcal{S}} = (X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1} X_{\mathcal{S}}^\top \boldsymbol{y}$$

- 選択された特徴集合 $\mathcal{S}$ の j 番目の特徴の統計的仮説検定

$$\mathrm{H}_0^{\mathcal{S},j} : w_{\mathcal{S},j} = 0 \quad \text{vs.} \quad \mathrm{H}_1^{\mathcal{S},j} : w_{\mathcal{S},j} \neq 0.$$

線形モデルの特徴選択バイアスの補正

- 特徴選択バイアスを補正するには，

$$\mathbb{P}(\hat{w}_{\mathcal{S},j} \notin [\ell_\alpha^{\mathcal{S},j}, u_\alpha^{\mathcal{S},j}] \mid j \in \mathcal{S}, \mathcal{S} \leftarrow \mathrm{FS}(X, \boldsymbol{y})) < \alpha$$

  を満たすようなパラメータの範囲 $[\ell_\alpha^{\mathcal{S},j}, u_\alpha^{\mathcal{S},j}]$ を求めればよい

- このような条件付分布に基づく統計的推論の枠組は**選択的推論（Selective Inference / Post-Selection Inference）**と呼ばれる
- **Q.** それぞれの特徴選択アルゴリズム $\mathrm{FS} : (X, \boldsymbol{y}) \mapsto \mathcal{S}$ に対し，どのように $[\ell_\alpha^{\mathcal{S},j}, u_\alpha^{\mathcal{S},j}]$ を求めればよいか?
- **A.** LASSO を特徴選択アルゴリズムとして用いた場合に $[\ell_\alpha^{\mathcal{S},j}, u_\alpha^{\mathcal{S},j}]$ を計算する方法が提案され，選択的推論の研究の契機となった

J. Lee, D. Sun, Y. Sun, J. Taylor. Exact post-selection inference, with application to the lasso. The Annals of Statistics, 2016.

Part1のまとめ

- データから生成・選択された仮説の推測では選択バイアスの補正が必要
- 選択バイアスを補正するためには,

$$\mathbb{P}(s_h(\mathsf{data}) \notin [\ell_\alpha^h, u_\alpha^h] \mid h \leftarrow \mathcal{A}(\mathsf{data})) < \alpha$$

  となるように採択域 $[\ell_\alpha^h, u_\alpha^h]$ を決めれればよい.
- 特徴選択された後の線形モデルパラメータの統計的仮説検定は選択的推論の枠組で議論できる

Part2

線形モデルにおける選択的推論

選択的推論

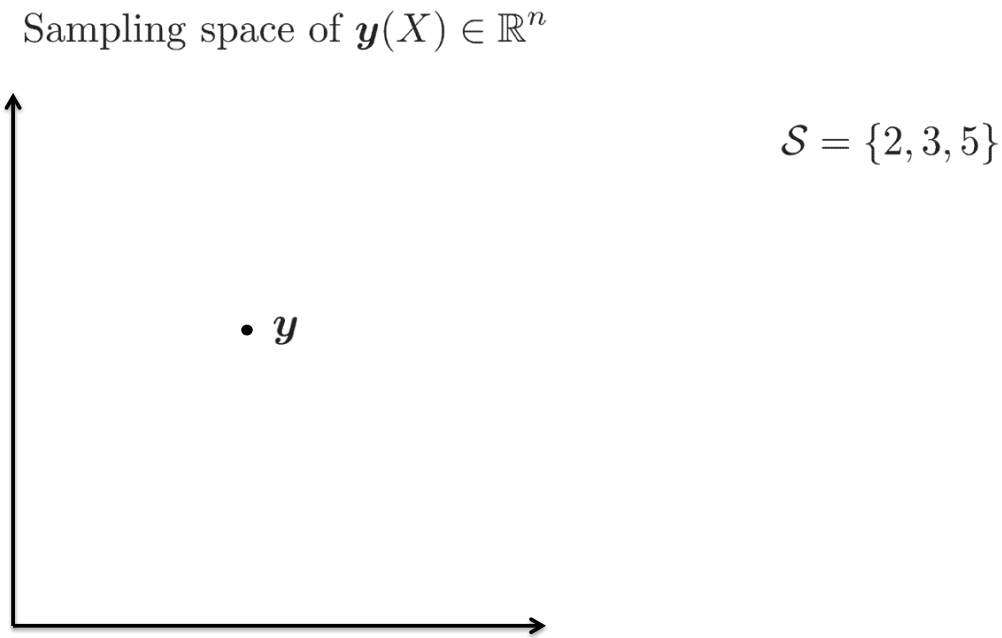
- 線形モデルの選択的推論

$$\mathbb{P}(\hat{w}_{\mathcal{S},j} \notin [\ell_\alpha^{\mathcal{S},j}, u_\alpha^{\mathcal{S},j}] \mid j \in \mathcal{S}, \mathcal{S} \leftarrow \mathrm{FS}(X, \boldsymbol{y})) < \alpha$$

選択的推論の解釈

- 特徴選択の逆像：特徴選択結果が $\mathcal{S}$ となる $\mathbb{R}^n$ の領域

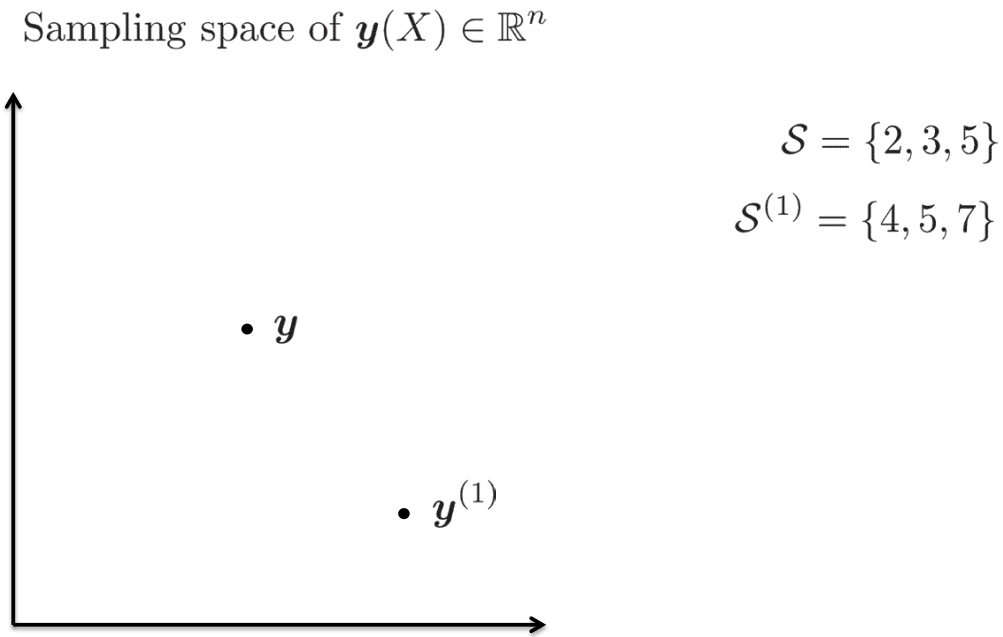
$$\mathcal{Y} := \{\boldsymbol{y} \in \mathbb{R}^n \mid \mathcal{S} \leftarrow \mathcal{A}(X, \boldsymbol{y})\}$$



選択的推論の解釈

- 特徴選択の逆像：特徴選択結果が $\mathcal{S}$ となる $\mathbb{R}^n$ の領域

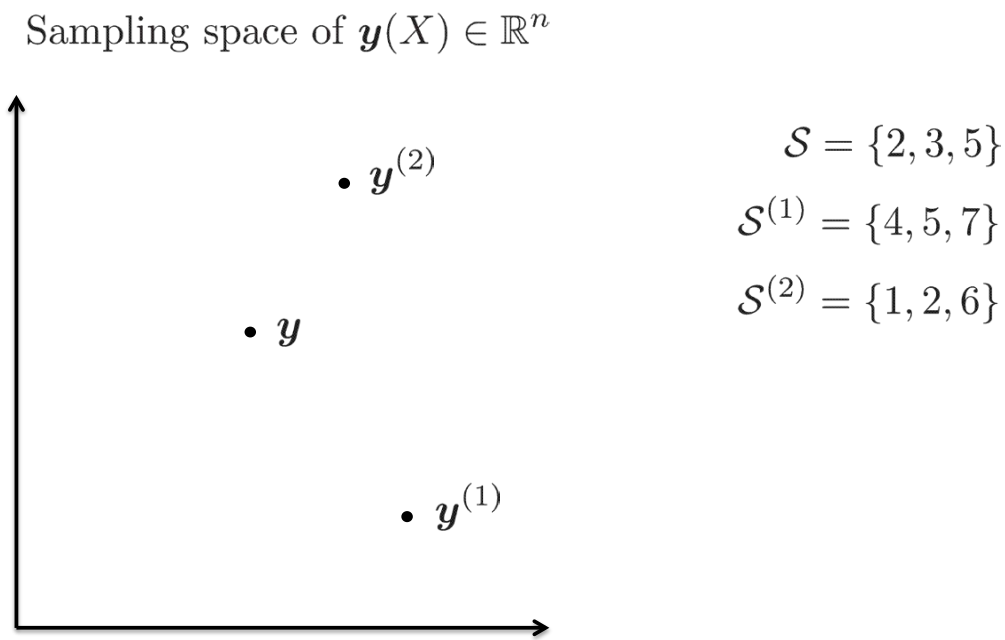
$$\mathcal{Y} := \{\boldsymbol{y} \in \mathbb{R}^n \mid \mathcal{S} \leftarrow \mathcal{A}(X, \boldsymbol{y})\}$$



選択的推論の解釈

- 特徴選択の逆像：特徴選択結果が $\mathcal{S}$ となる $\mathbb{R}^n$ の領域

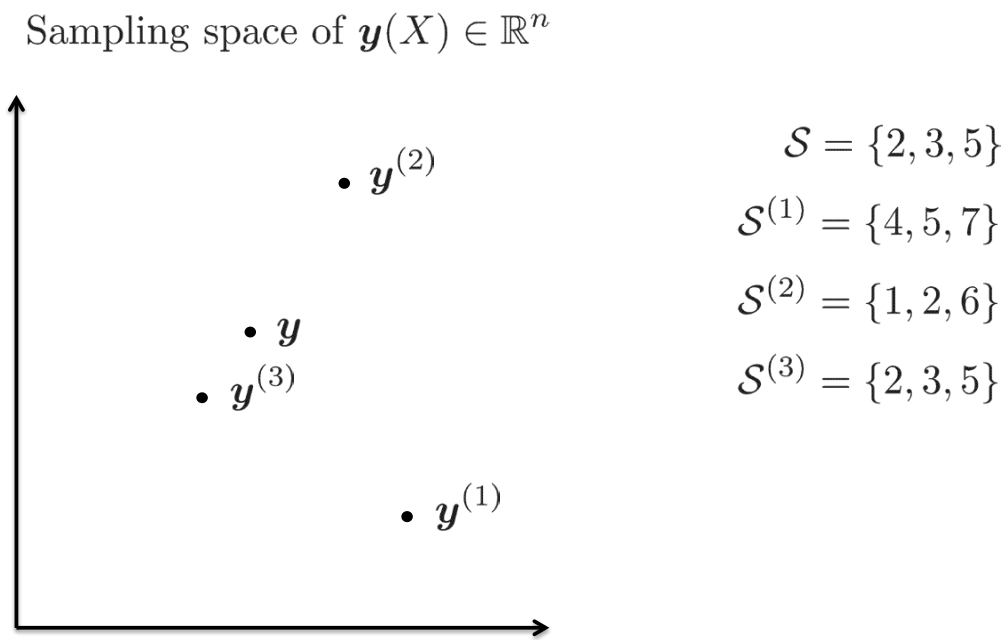
$$\mathcal{Y} := \{\boldsymbol{y} \in \mathbb{R}^n \mid \mathcal{S} \leftarrow \mathcal{A}(X, \boldsymbol{y})\}$$



選択的推論の解釈

- 特徴選択の逆像：特徴選択結果が $\mathcal{S}$ となる $\mathbb{R}^n$ の領域

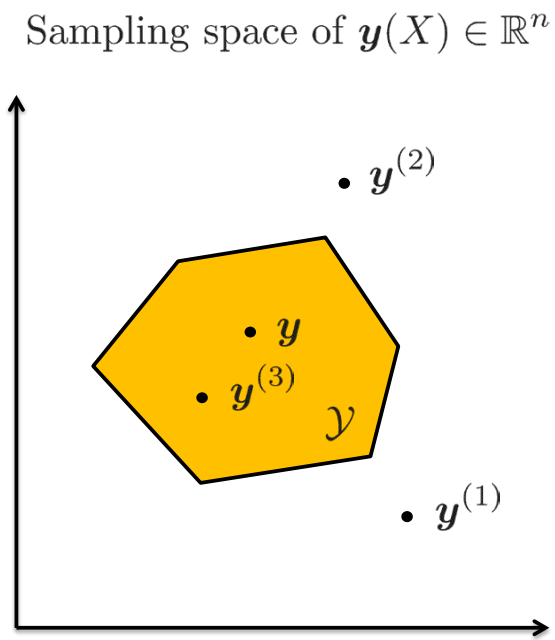
$$\mathcal{Y} := \{\boldsymbol{y} \in \mathbb{R}^n \mid \mathcal{S} \leftarrow \mathcal{A}(X, \boldsymbol{y})\}$$



選択的推論の解釈

- 特徴選択の逆像：特徴選択結果が $\mathcal{S}$ となる $\mathbb{R}^n$ の領域

$$\mathcal{Y} := \{\boldsymbol{y} \in \mathbb{R}^n \mid \mathcal{S} \leftarrow \mathcal{A}(X, \boldsymbol{y})\}$$



$\mathcal{S} = \{2, 3, 5\}$

~~$\mathcal{S}^{(1)} = \{4, 5, 7\}$~~

~~$\mathcal{S}^{(2)} = \{1, 2, 6\}$~~

$\mathcal{S}^{(3)} = \{2, 3, 5\}$

~~$\mathcal{S}^{(4)} = \{2, 6, 8\}$~~

$\mathcal{S}^{(5)} = \{2, 3, 5\}$

~~$\mathcal{S}^{(6)} = \{1, 8, 9\}$~~

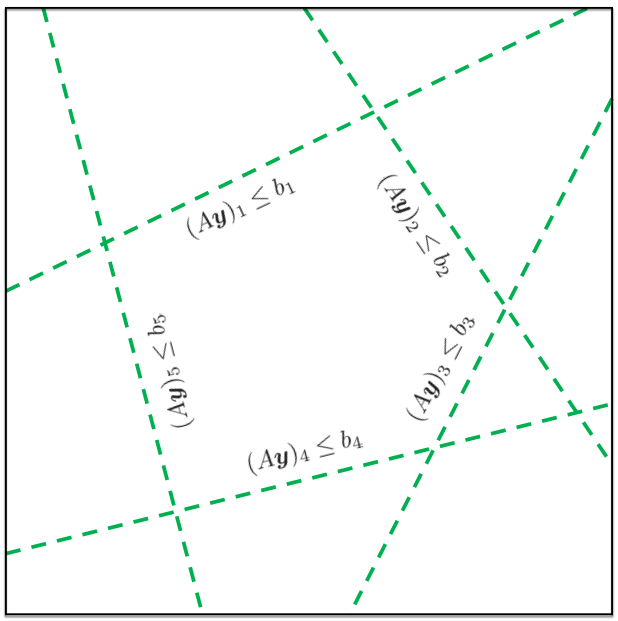
線形選択イベントに基づく選択的推論

- Lee et al.(2016) は，特徴選択イベントが

$$A\boldsymbol{y} \leq \boldsymbol{b} \quad （A，\boldsymbol{b} \text{ は適当な行列とベクトル}）$$

と書けるとき，切断正規分布を用いて，棄却点 $\ell_\alpha^{\mathcal{S},j}$, $u_\alpha^{\mathcal{S},j}$ を計算する手順を提案した

$$\mathbb{P}(\hat{w}_{\mathcal{S},j} \notin [\ell_{\alpha/2}^{\mathcal{S},j}, u_{\alpha/2}^{\mathcal{S},j}] \mid A\boldsymbol{y} \leq \boldsymbol{b}) < \alpha$$



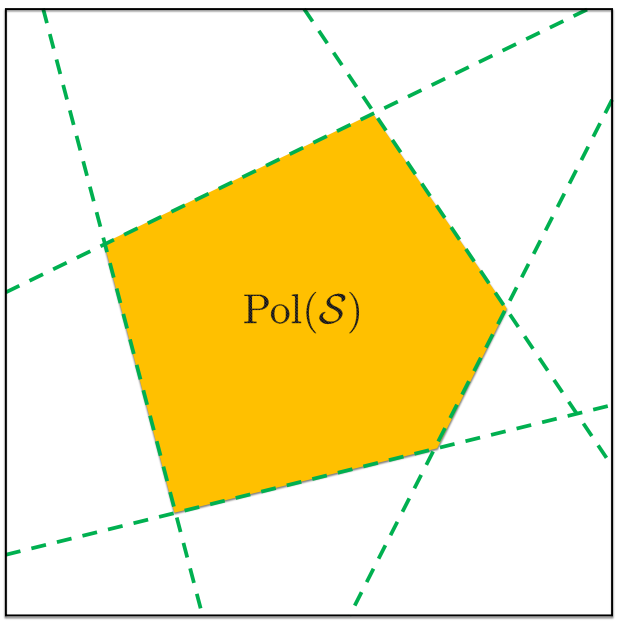
線形選択イベントに基づく選択的推論

- Lee et al.(2016) は，特徴選択イベントが

$$A\boldsymbol{y} \leq \boldsymbol{b} \quad （A，\boldsymbol{b} \text{ は適当な行列とベクトル}）$$

と書けるとき，切断正規分布を用いて，棄却点 $\ell_{\alpha}^{\mathcal{S},j}$, $u_{\alpha}^{\mathcal{S},j}$ を計算する手順を提案した

$$\mathbb{P}(\hat{w}_{\mathcal{S},j} \notin [\ell_{\alpha/2}^{\mathcal{S},j}, u_{\alpha/2}^{\mathcal{S},j}] \mid \boldsymbol{y} \in \mathrm{Pol}(\mathcal{S})) < \alpha$$



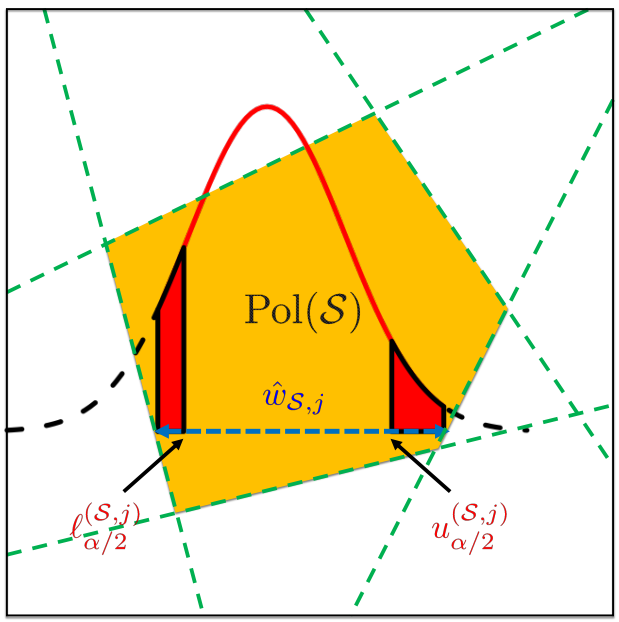
線形選択イベントに基づく選択的推論

- Lee et al.(2016) は，特徴選択イベントが

$$A\boldsymbol{y} \leq \boldsymbol{b} \quad （A，\boldsymbol{b} \text{ は適当な行列とベクトル}）$$

と書けるとき，切断正規分布を用いて，棄却点 $\ell_{\alpha}^{\mathcal{S},j}$, $u_{\alpha}^{\mathcal{S},j}$ を計算する手順を提案した

$$\mathbb{P}(\hat{w}_{\mathcal{S},j} \notin [\ell_{\alpha/2}^{\mathcal{S},j}, u_{\alpha/2}^{\mathcal{S},j}] \mid \boldsymbol{y} \in \mathrm{Pol}(\mathcal{S})) < \alpha$$



線形選択イベントの例：*Marginal Screening*

- Marginal screening (MS) による特徴選択

$$\underbrace{\boldsymbol{x}_{(1)}^\top \boldsymbol{y} \geq \boldsymbol{x}_{(2)}^\top \boldsymbol{y} \geq \boldsymbol{x}_{(3)}^\top \boldsymbol{y}}_{\text{selected (when } k=3)} \geq \underbrace{\boldsymbol{x}_{(4)}^\top \boldsymbol{y} \geq \boldsymbol{x}_{(5)}^\top \boldsymbol{y} \geq \cdots \geq \boldsymbol{x}_{(D)}^\top \boldsymbol{y}}_{\text{not selected (when } k=3)},$$

  MS：内積（正規化後の相関）の高い k 個の特徴を選択する

- 特徴選択イベントは以下のように表される：

$$\begin{array}{ccc} \boldsymbol{x}_{(1)}{}^\top \boldsymbol{y} \geq \boldsymbol{x}_{(4)}{}^\top \boldsymbol{y} & \boldsymbol{x}_{(2)}{}^\top \boldsymbol{y} \geq \boldsymbol{x}_{(4)}{}^\top \boldsymbol{y} & \boldsymbol{x}_{(3)}{}^\top \boldsymbol{y} \geq \boldsymbol{x}_{(4)}{}^\top \boldsymbol{y} \\ \vdots & \vdots & \vdots \\ \boldsymbol{x}_{(1)}{}^\top \boldsymbol{y} \geq \boldsymbol{x}_{(D)}{}^\top \boldsymbol{y} & \boldsymbol{x}_{(2)}{}^\top \boldsymbol{y} \geq \boldsymbol{x}_{(D)}{}^\top \boldsymbol{y} & \boldsymbol{x}_{(3)}{}^\top \boldsymbol{y} \geq \boldsymbol{x}_{(D)}{}^\top \boldsymbol{y}, \end{array}$$

- 適当な行列 A とベクトル $\boldsymbol{b}$ を用いると線形イベントは

$$\{\mathcal{S} \leftarrow \mathrm{FS}(X, \boldsymbol{y})\} \equiv \{A\boldsymbol{y} \leq \boldsymbol{b}\} \equiv \{\boldsymbol{y} \in \mathrm{Pol}(\mathcal{S})\}$$

  と書ける

The Polyhedral Lemma *(Lee et al., AS2016)*

If the selection is represented as a linear selection event

$$\{\boldsymbol{y} \in \mathrm{Pol}(\mathcal{S})\} \quad \Leftrightarrow \quad A\boldsymbol{y} \leq \boldsymbol{b},$$

then, the critical values are computed as

$$\ell_\alpha^{\mathcal{S},j} := (F_{0,\sigma^2_{\mathcal{S},j}}^{[L(\mathcal{S},j),U(\mathcal{S},j)]})^{-1}(\alpha/2),$$
$$u_\alpha^{\mathcal{S},j} := (F_{0,\sigma^2_{\mathcal{S},j}}^{[L(\mathcal{S},j),U(\mathcal{S},j)]})^{-1}(1-\alpha/2),$$

where $F_{\mu,\sigma^2}^{[L,U]}$ is the cumulative distribution function of a truncated Normal distribution $\mathrm{TN}(\mu,\sigma^2,L,U)$, and the truncation points are defined as

$$L(\mathcal{S},j) = \hat{w}_{\mathcal{S},j} + \kappa_L (X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}_{jj},$$
$$\text{where} \quad \kappa_L := \min_{\kappa\in\mathbb{R}} \kappa \text{ s.t. } \boldsymbol{y} + \kappa (X_{\mathcal{S}}^+)^\top \boldsymbol{e}_j \in \mathrm{Pol}(\mathcal{S}),$$
$$U(\mathcal{S},j) = \hat{w}_{\mathcal{S},j} + \kappa_U (X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}_{jj},$$
$$\text{where} \quad \kappa_U := \max_{\kappa\in\mathbb{R}} \kappa \text{ s.t. } \boldsymbol{y} + \kappa (X_{\mathcal{S}}^+)^\top \boldsymbol{e}_j \in \mathrm{Pol}(\mathcal{S}).$$

線形選択イベントとなる特徴選択アルゴリズム

- 多くの特徴選択アルゴリズムが線形選択イベントとして記述可能
  - LASSO

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \ (\boldsymbol{y} - X\boldsymbol{w})^\top (\boldsymbol{y} - X\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_1$$

$$\Leftrightarrow \ X^\top (X\hat{\boldsymbol{w}} - \boldsymbol{y}) + \lambda \hat{\boldsymbol{s}} = \boldsymbol{0}, \quad \begin{array}{ll} \hat{s}_j = \mathrm{sgn}(\hat{w}_j) & \text{if } \hat{w}_j \neq 0, \\ \hat{s}_j \in [-1, 1] & \text{if } \hat{w}_j = 0. \end{array}$$

  - Orthogonal Matching Pursuit
- パターンマイニングへの拡張

  Suzumura S., Nakagawa K., Umezu Y., Tsuda K. and Takeuchi I. Selective inference for sparse high-order interaction models (ICML2017).
- カーネル法への拡張

  Yamada, M., Umezu, Y., Fukumizu, K., and Takeuchi, I. Post Selection Inference with Kernels (AISTATS2018).

Part2のまとめ

- 選択的推論では特徴選択の逆像を計算する

$$\mathcal{Y} := \{\boldsymbol{y} \in \mathbb{R}^n \mid \mathcal{S} \leftarrow \mathcal{A}(X, \boldsymbol{y})\}$$

- LASSO を含む多くの特徴選択アルゴリズムでは逆像が $\boldsymbol{y}$ に関する線形制約の集合で記述できる

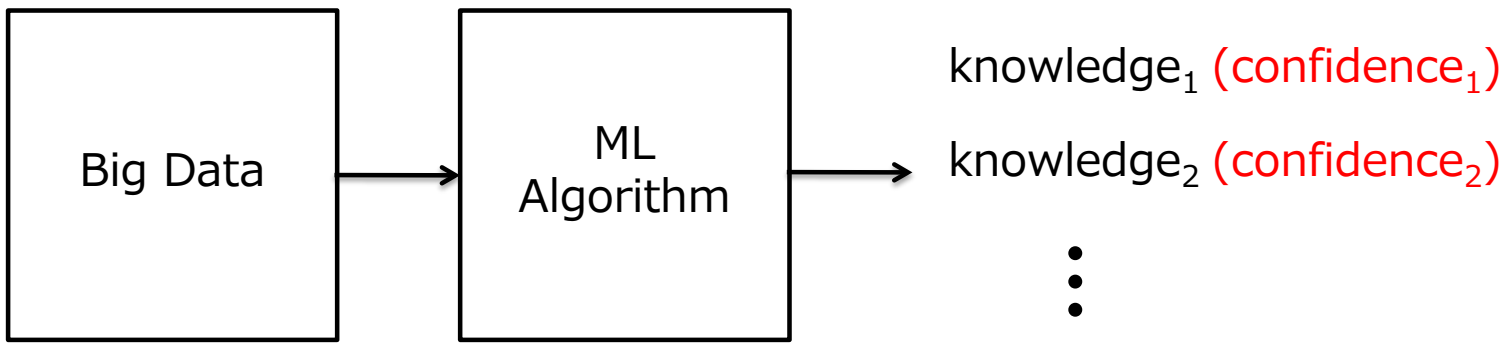
$$\{\mathcal{S} \leftarrow \mathcal{A}(X, \boldsymbol{y})\} \Leftrightarrow \{A\boldsymbol{y} \leq \boldsymbol{b}\} \text{ for certain } A \text{ and } \boldsymbol{b}$$

- $\mathbb{R}^n$ の正規分布が多面体で制約されるため，棄却点 $\ell_\alpha^{\mathcal{S},j}$, $u_\alpha^{\mathcal{S},j}$ は切断正規分布により特徴づけられ，計算可能

Part3

教師なし学習における選択的推論

データ駆動型アプローチとポスト機械学習推論



$$\mathbb{P}(s_h(\mathsf{data}) \in [\ell_\alpha^h, u_\alpha^h] \mid h \leftarrow \mathcal{A}(\mathsf{data})) < \alpha$$

Post Unsupervised Learning Inference

- Post-clustering inference

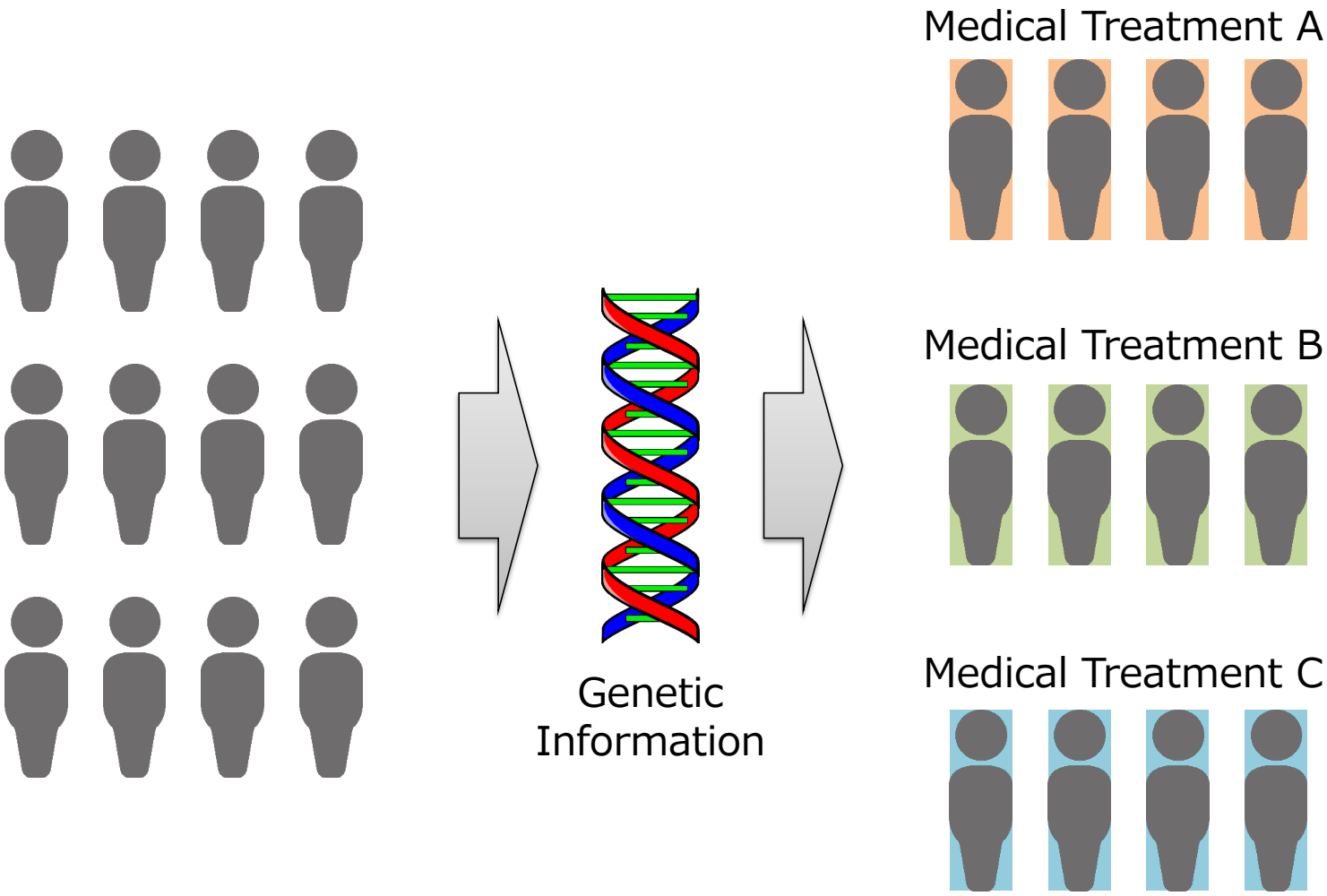
  Joint work with Inoue S, Suzuki K, Shoma T., Umezu Y., DuVerle D., Kadomatsu K., and Tsuda K.

- Post-segmentation inference

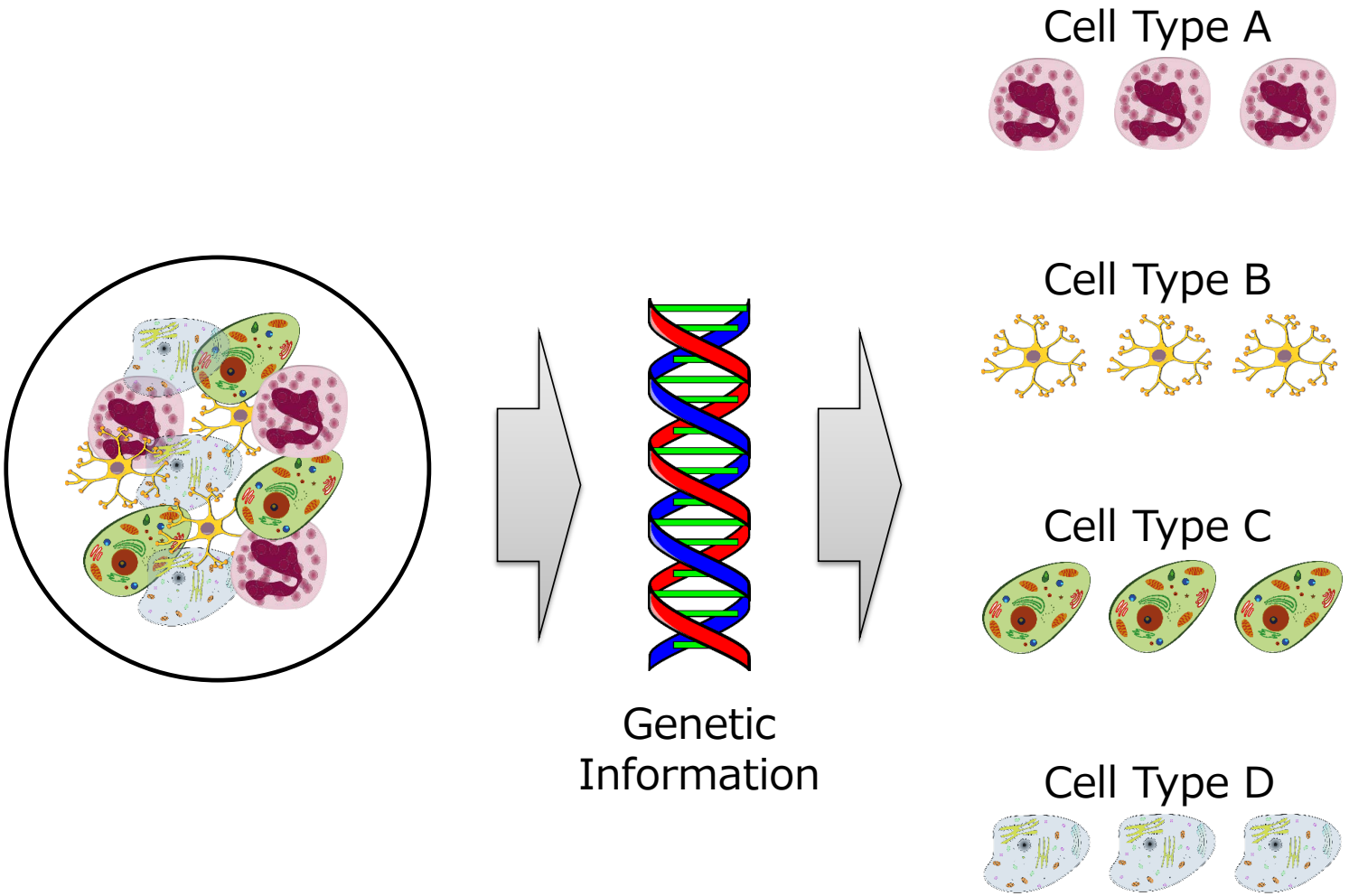
  Joint work with Tanizaki K., Toda H., Sakuma T., Hashimoto, N., and Hontani H.

ポストクラスタリング推論

例題 1 ：精密医療（個別化医療）



例題２：シングル細胞解析



問題設定

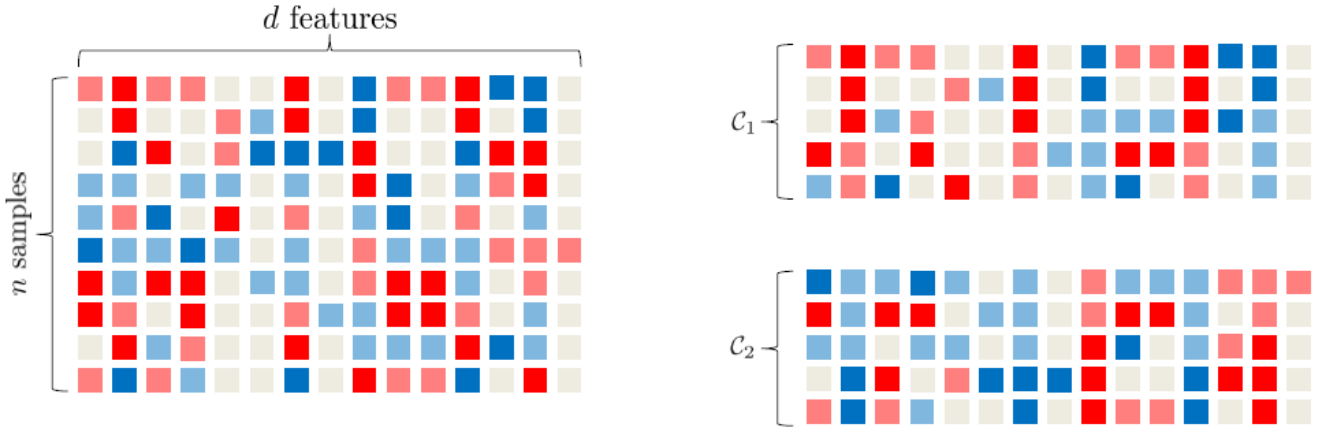
- データ（事例数 n，特徴数 d）

$$
\underset{n\times d}{X} = \begin{array}{c}
\begin{array}{cccccc} \text{GF 1} & \text{GF 2} & & & & \text{GF 1000} \end{array} \\
\begin{bmatrix}
0.9 & 0.5 & \cdots & \cdots & \cdots & 0.8 \\
0.5 & 0.6 & \cdots & \cdots & \cdots & 1.0 \\
\vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\
\vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\
\vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\
0.8 & 0.4 & \cdots & \cdots & \cdots & 0.3
\end{bmatrix}
\end{array}
\begin{array}{l}
\text{patient 1} \\ \text{patient 2} \\ \\ \\ \\ \text{patient 200}
\end{array}
$$

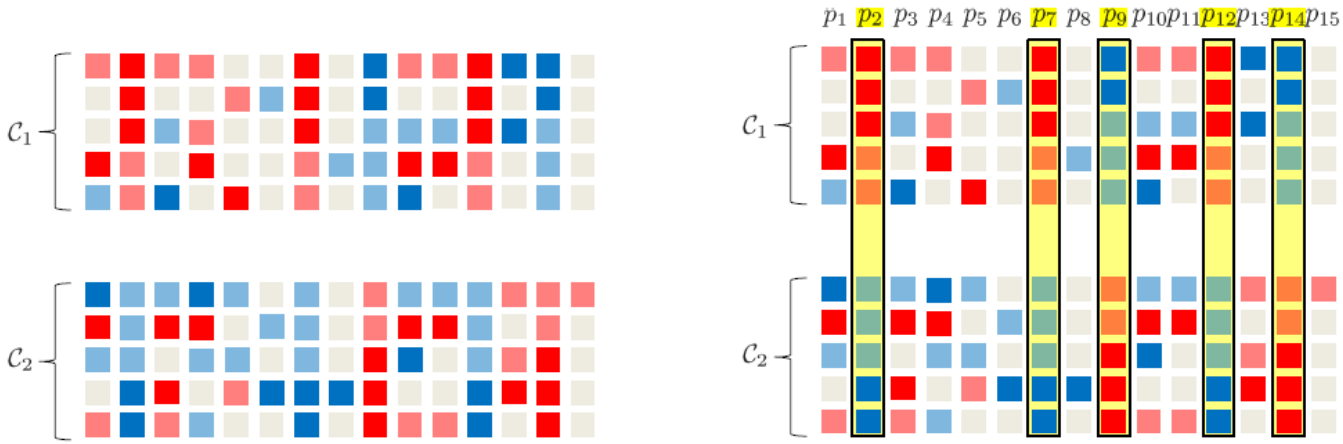
- 問題設定：サブタイプ（クラスタ）を同定し，それぞれのサブタイプに特異的な遺伝的特徴を発見する

不均一データ分析

- Step 1. クラスタリングによりサブタイプ（クラスタ）を同定する

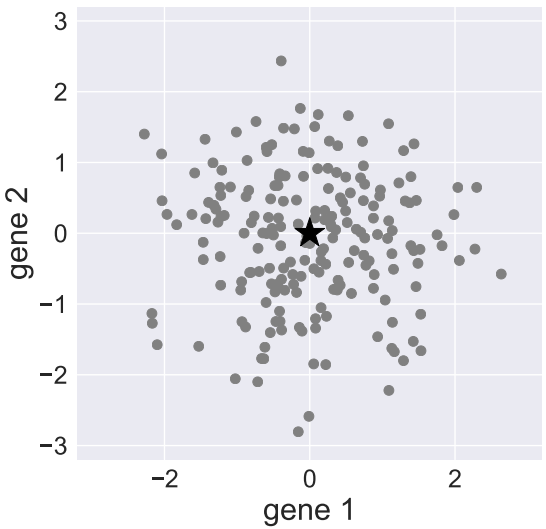


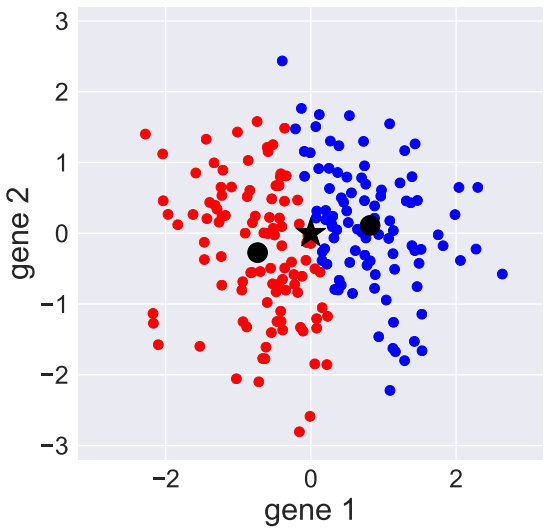
- Step 2. 各クラスタ特異的な特徴を発見し，統計的信頼性を与える

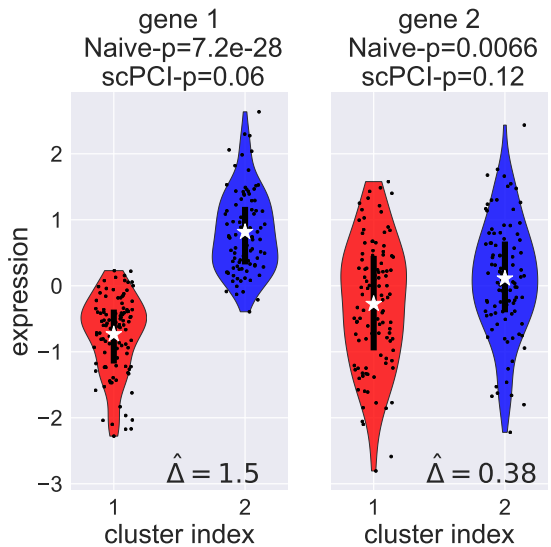


クラスタリングにおけるバイアス

帰無仮説（クラスタなし）におけるシミュレーション



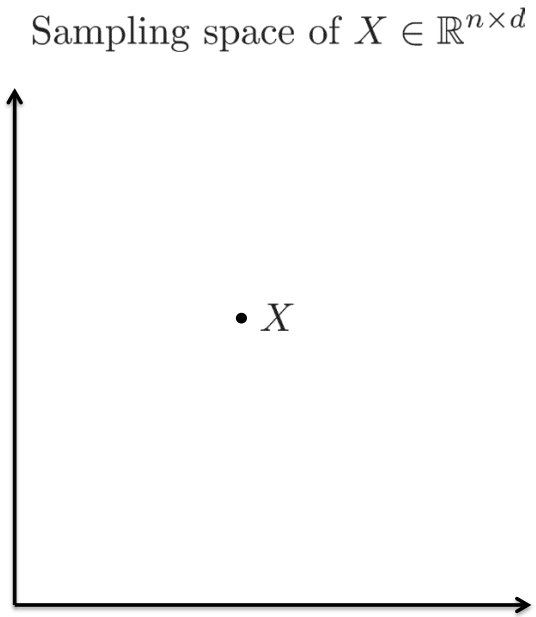




K 平均クラスタリング結果の逆像

- クラスタリング結果が $\mathcal{C}$ となる $\mathbb{R}^{n \times d}$ の領域：

$$\mathcal{X} := \{X \in \mathbb{R}^{n \times d} \mid \mathcal{C} \leftarrow \mathcal{A}(X)\}$$

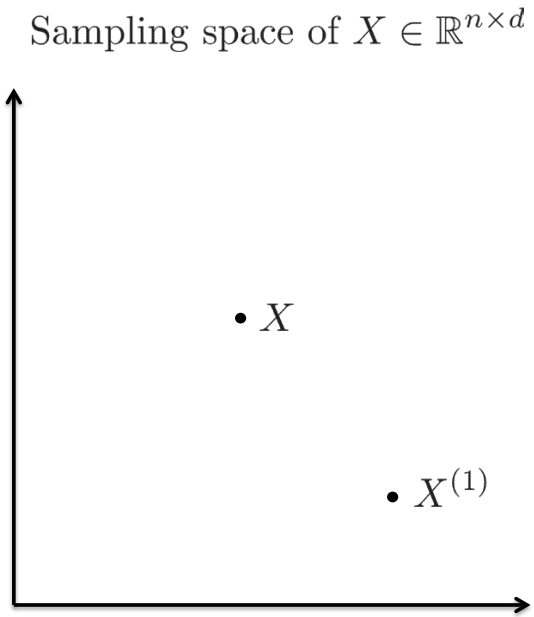


$$\mathcal{S} = \{\{1,3,8\}, \{2,4,5\}, \{6,7,9\}\}$$
$$\mathcal{C}_1 \quad \mathcal{C}_2 \quad \mathcal{C}_3$$

K 平均クラスタリング結果の逆像

- クラスタリング結果が $\mathcal{C}$ となる $\mathbb{R}^{n \times d}$ の領域：

$$\mathcal{X} := \{X \in \mathbb{R}^{n \times d} \mid \mathcal{C} \leftarrow \mathcal{A}(X)\}$$

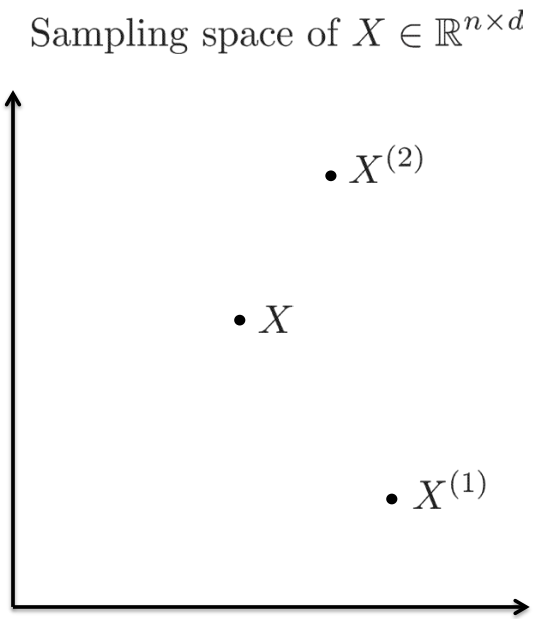


$$\begin{array}{rccc} & \mathcal{C}_1 & \mathcal{C}_2 & \mathcal{C}_3 \\ \mathcal{S} = & \{1,3,8\}, & \{2,4,5\}, & \{6,7,9\}\} \\ \mathcal{S}^{(1)} = & \{1,2,8\}, & \{3,4,7\}, & \{4,5,6\}\} \end{array}$$

K 平均クラスタリング結果の逆像

- クラスタリング結果が $\mathcal{C}$ となる $\mathbb{R}^{n \times d}$ の領域：

$$\mathcal{X} := \{X \in \mathbb{R}^{n \times d} \mid \mathcal{C} \leftarrow \mathcal{A}(X)\}$$

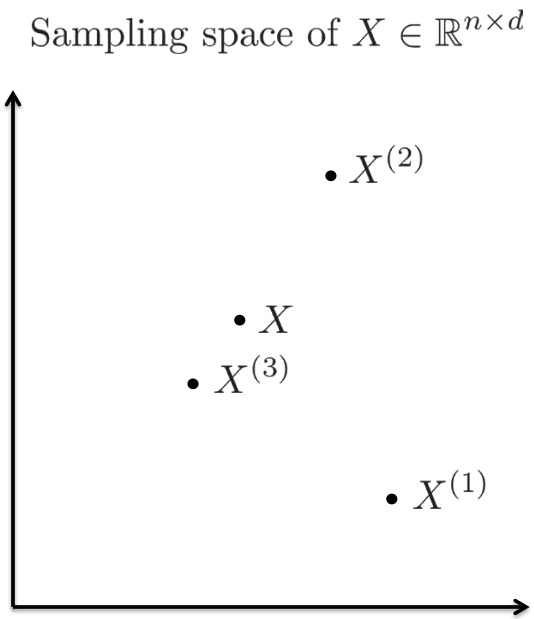


$$\begin{array}{rccc} & \mathcal{C}_1 & \mathcal{C}_2 & \mathcal{C}_3 \\ \mathcal{S} = & \{1,3,8\}, & \{2,4,5\}, & \{6,7,9\}\} \\ \mathcal{S}^{(1)} = & \{1,2,8\}, & \{3,4,7\}, & \{4,5,6\}\} \\ \mathcal{S}^{(2)} = & \{1,2,6\}, & \{4,5,7\}, & \{3,7,9\}\} \end{array}$$

K 平均クラスタリング結果の逆像

- クラスタリング結果が $\mathcal{C}$ となる $\mathbb{R}^{n \times d}$ の領域：

$$\mathcal{X} := \{X \in \mathbb{R}^{n \times d} \mid \mathcal{C} \leftarrow \mathcal{A}(X)\}$$

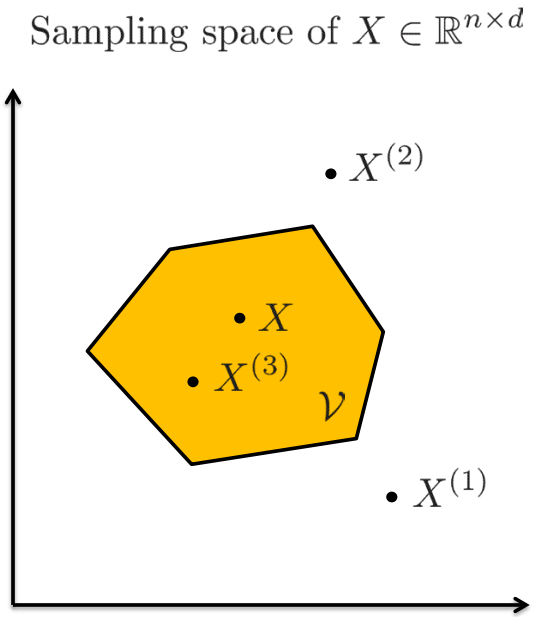


$$\begin{array}{rccc} & \mathcal{C}_1 & \mathcal{C}_2 & \mathcal{C}_3 \\ \mathcal{S} = & \{1,3,8\}, & \{2,4,5\}, & \{6,7,9\}\} \\ \mathcal{S}^{(1)} = & \{1,2,8\}, & \{3,4,7\}, & \{4,5,6\}\} \\ \mathcal{S}^{(2)} = & \{1,2,6\}, & \{4,5,7\}, & \{3,7,9\}\} \\ \mathcal{S}^{(3)} = & \{1,3,8\}, & \{2,4,5\}, & \{6,7,9\}\} \end{array}$$

K 平均クラスタリング結果の逆像

- クラスタリング結果が $\mathcal{C}$ となる $\mathbb{R}^{n \times d}$ の領域：

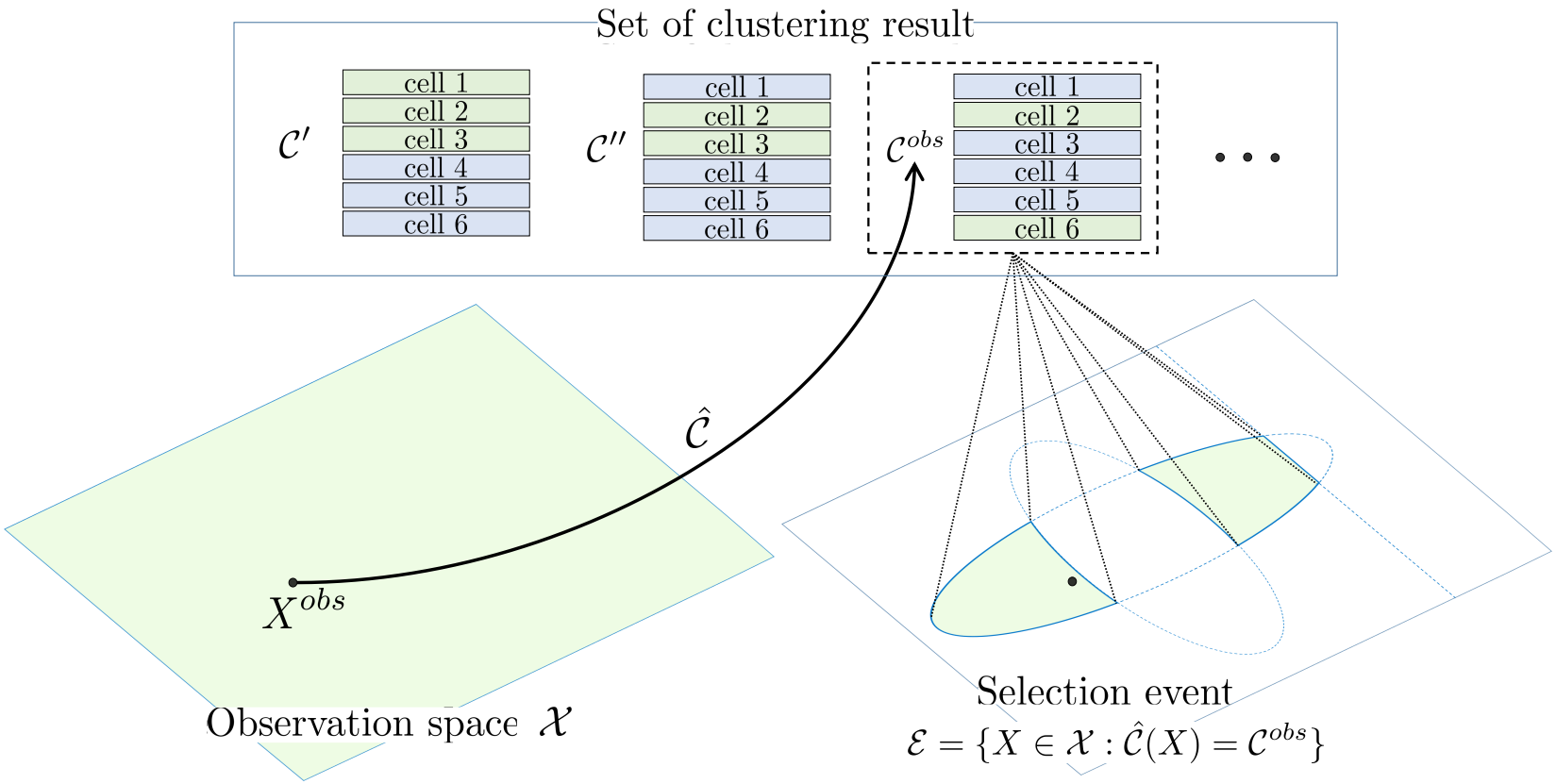
$$\mathcal{X} := \{X \in \mathbb{R}^{n \times d} \mid \mathcal{C} \leftarrow \mathcal{A}(X)\}$$



$$\begin{array}{rccc} & \mathcal{C}_1 & \mathcal{C}_2 & \mathcal{C}_3 \\ \mathcal{S} = & \{1,3,8\}, & \{2,4,5\}, & \{6,7,9\}\} \\ \cancel{\mathcal{S}^{(1)} = \{\{1,2,8\}, \{3,4,7\}, \{4,5,6\}\}} & & & \\ \cancel{\mathcal{S}^{(2)} = \{\{1,2,6\}, \{4,5,7\}, \{3,7,9\}\}} & & & \\ \mathcal{S}^{(3)} = & \{1,3,8\}, & \{2,4,5\}, & \{6,7,9\}\} \\ \cancel{\mathcal{S}^{(4)} = \{\{1,4,6\}, \{2,3,9\}, \{5,7,8\}\}} & & & \\ \mathcal{S}^{(5)} = & \{1,3,8\}, & \{2,4,5\}, & \{6,7,9\}\} \\ \cancel{\mathcal{S}^{(6)} = \{\{1,7,8\}, \{2,6,9\}, \{3,4,5\}\}} & & & \end{array}$$

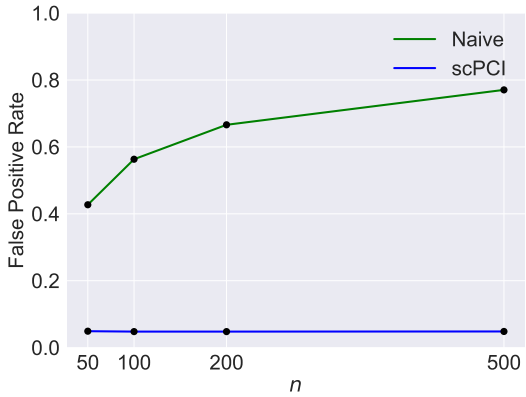
K 平均クラスタリングの選択イベント

- K 平均クラスタリングの選択イベントはデータ $X \in \mathbb{R}^{n \times d}$ に関する線形制約と二次制約で記述可能

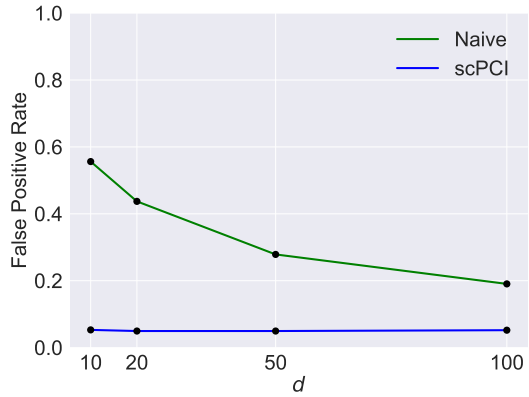


帰無仮説（クラスタなし）に対する結果

- 人工データ

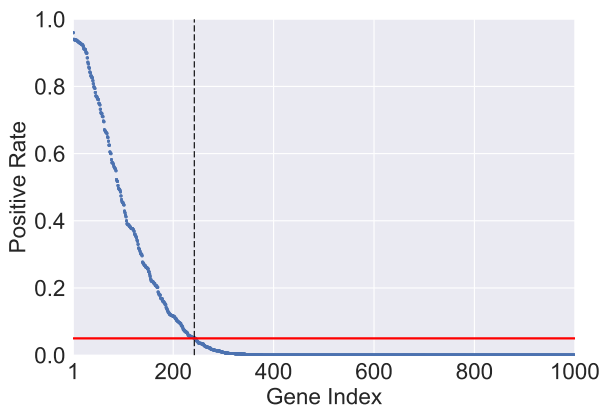


$n = 50, 100, 200, 500$

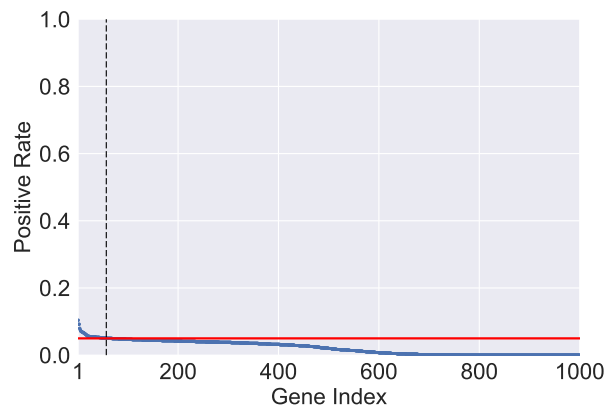


$d = 10, 20, 50, 100$

- 実データをランダムシャッフル

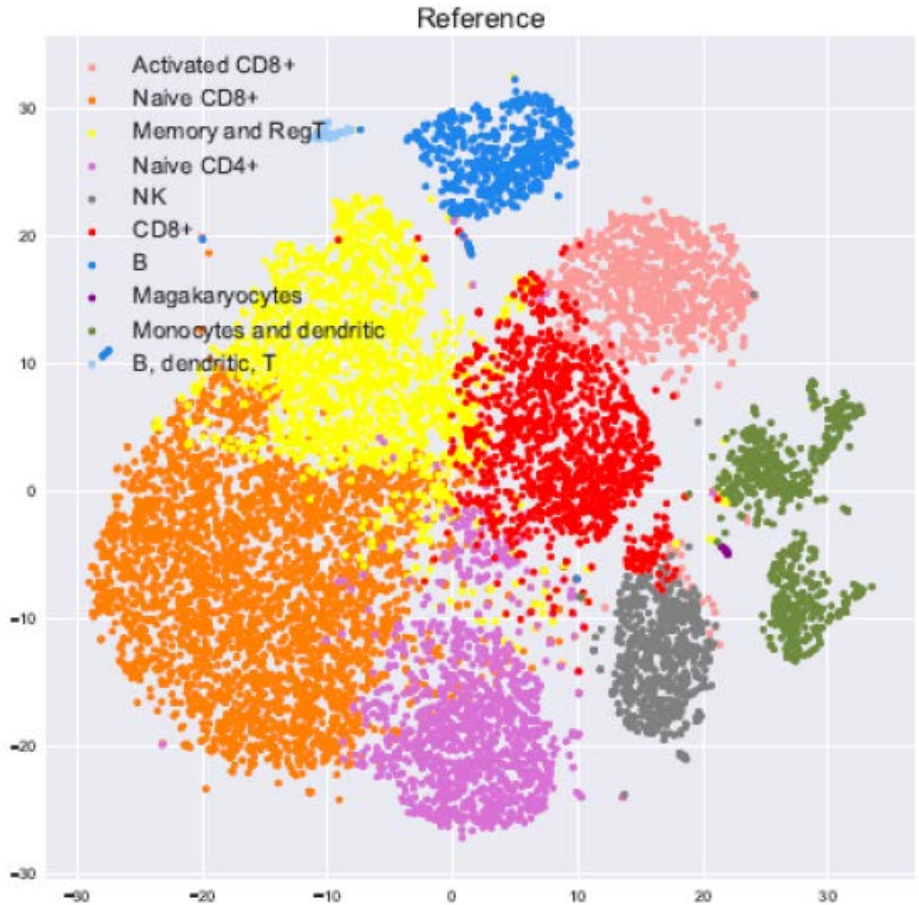


Naive (no bias correction)



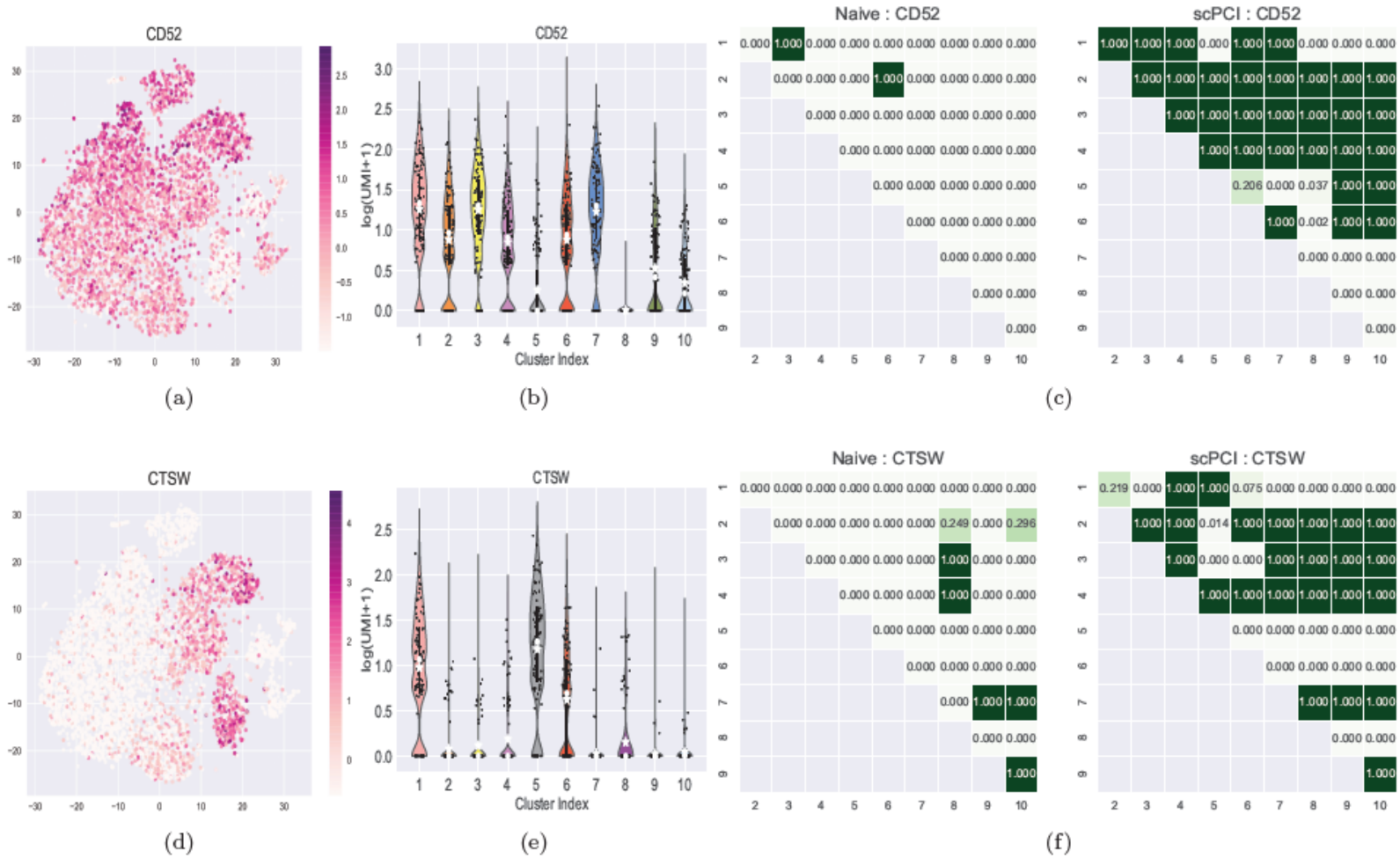
Post-clustering Inference

シングル細胞解析への応用

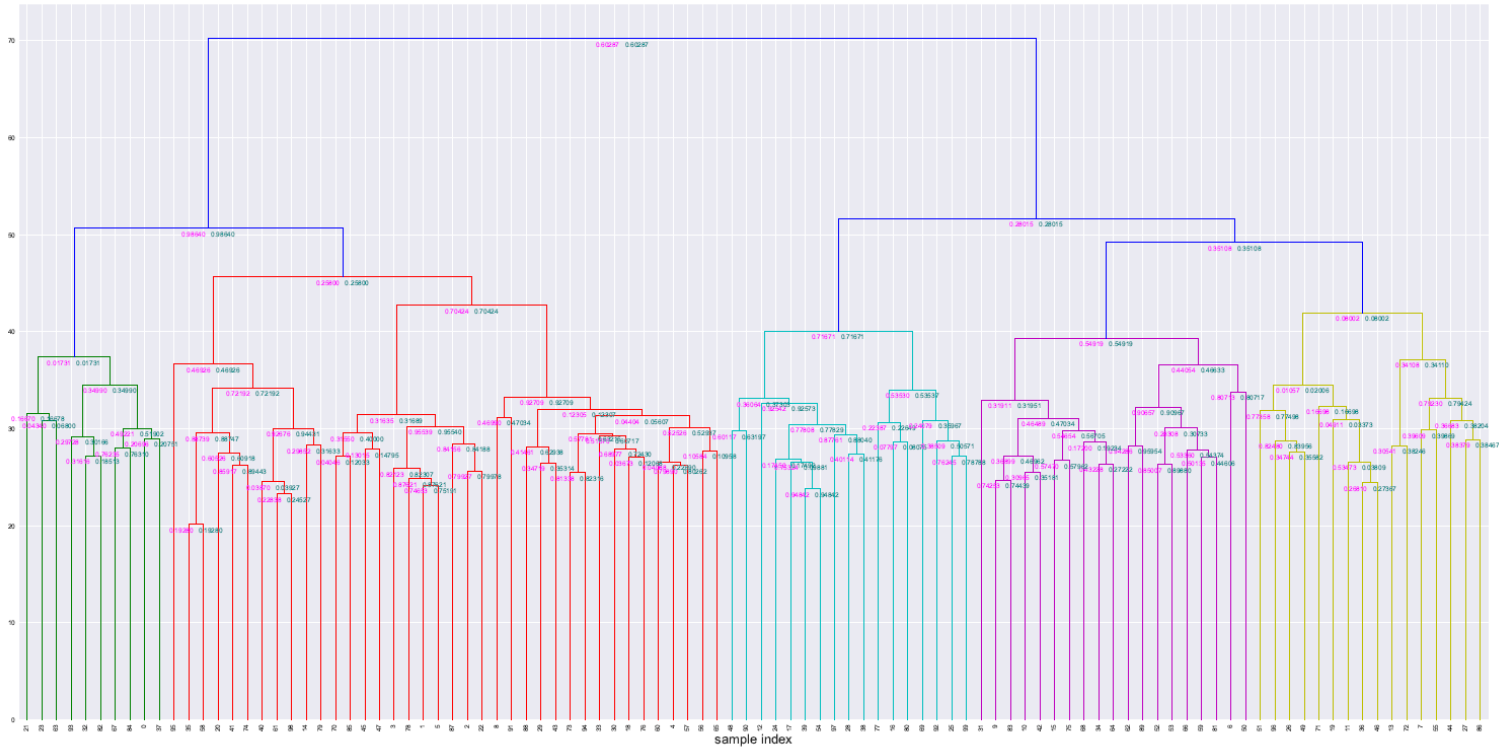


Zhang et al. (2017)

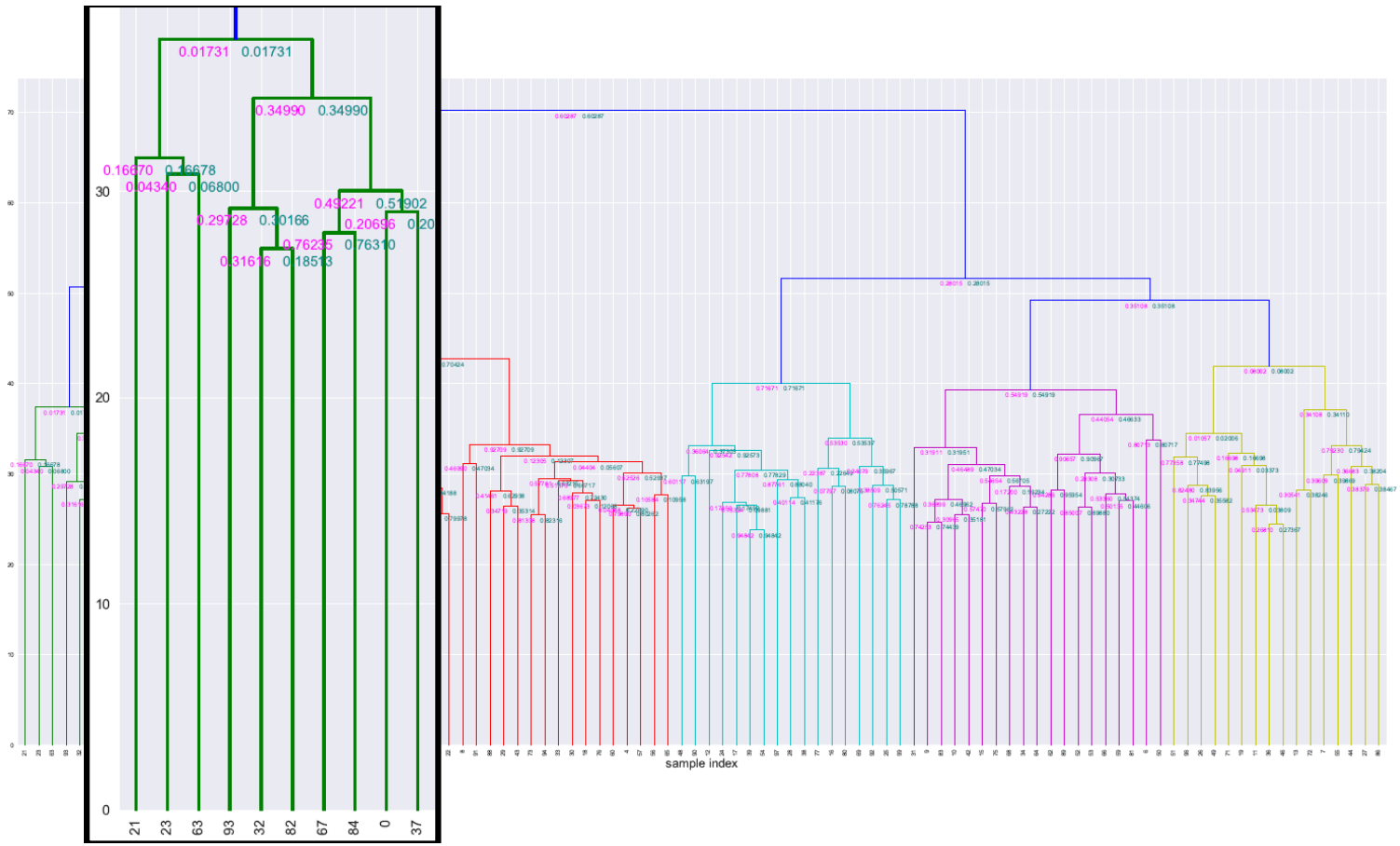
生物学的考察


CD52
log(UMI+1)
Cluster Index
(a)
(b)
Naive : CD52
scPCI : CD52
(c)
CTSW
(d)
(e)
Naive : CTSW
scPCI : CTSW
(f)

ポスト階層型クラスタリング推論も可能



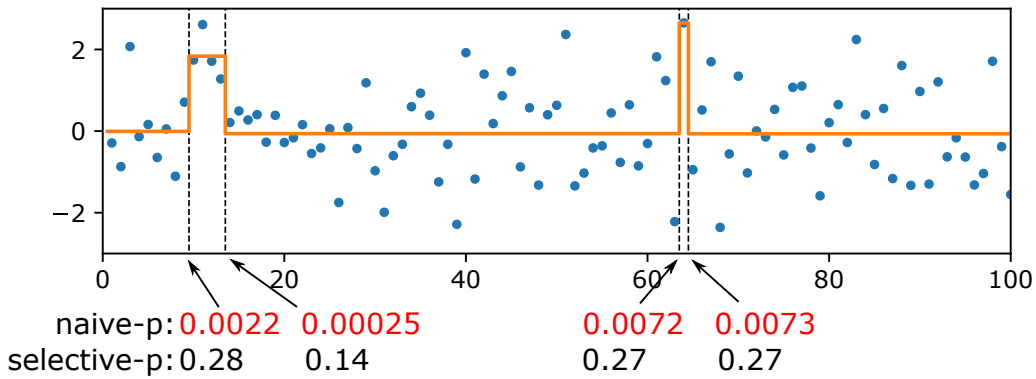
ポスト階層型クラスタリング推論も可能



ポストセグメンテーション推論

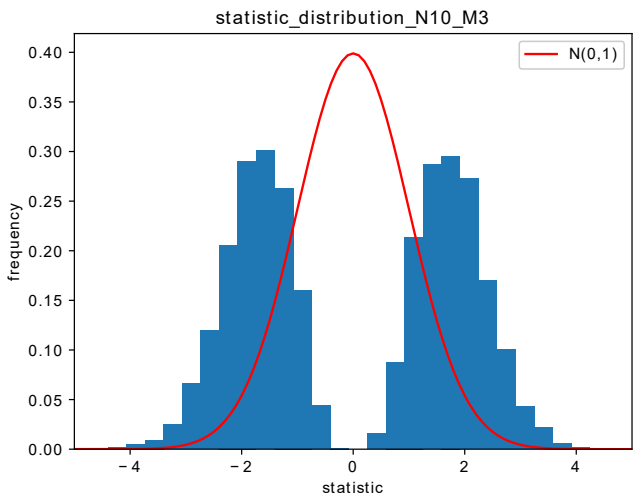
ポストセグメンテーション推論（動的計画法）

- Segmentation of null sequence



- Average difference between two segments:

$$\bar{X}_1 - \bar{X}_2 \sim N(0, \frac{1}{n_1} + \frac{1}{n_2})$$



- Sampling distribution of the statistic:

$$\tau := \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1)$$

二次形式による選択イベント

- Selection event

$$\mathcal{E} = \bigcap_{m=2}^{M} \bigcap_{n=m}^{m+N-M} \left\{ \hat{B}_n^{(m)} = B_n^{(m)} \right\}$$

$$= \bigcap_{m=2}^{M} \bigcap_{n=m}^{m+N-M} \bigcap_{h=m-1}^{n-1} \left\{ D_{B_n^{(m)}}^{(m-1)} + d_{B_n^{(m)}+1,n} \leq D_h^{(m-1)} + d_{h+1,n} \right\}$$

$$= \bigcap_{m=2}^{M} \bigcap_{n=m}^{m+N-M} \bigcap_{h=m-1}^{n-1} \left\{ \boldsymbol{x}^\top A_{nh}^{(m)} \boldsymbol{x} \leq 0 \right\}$$

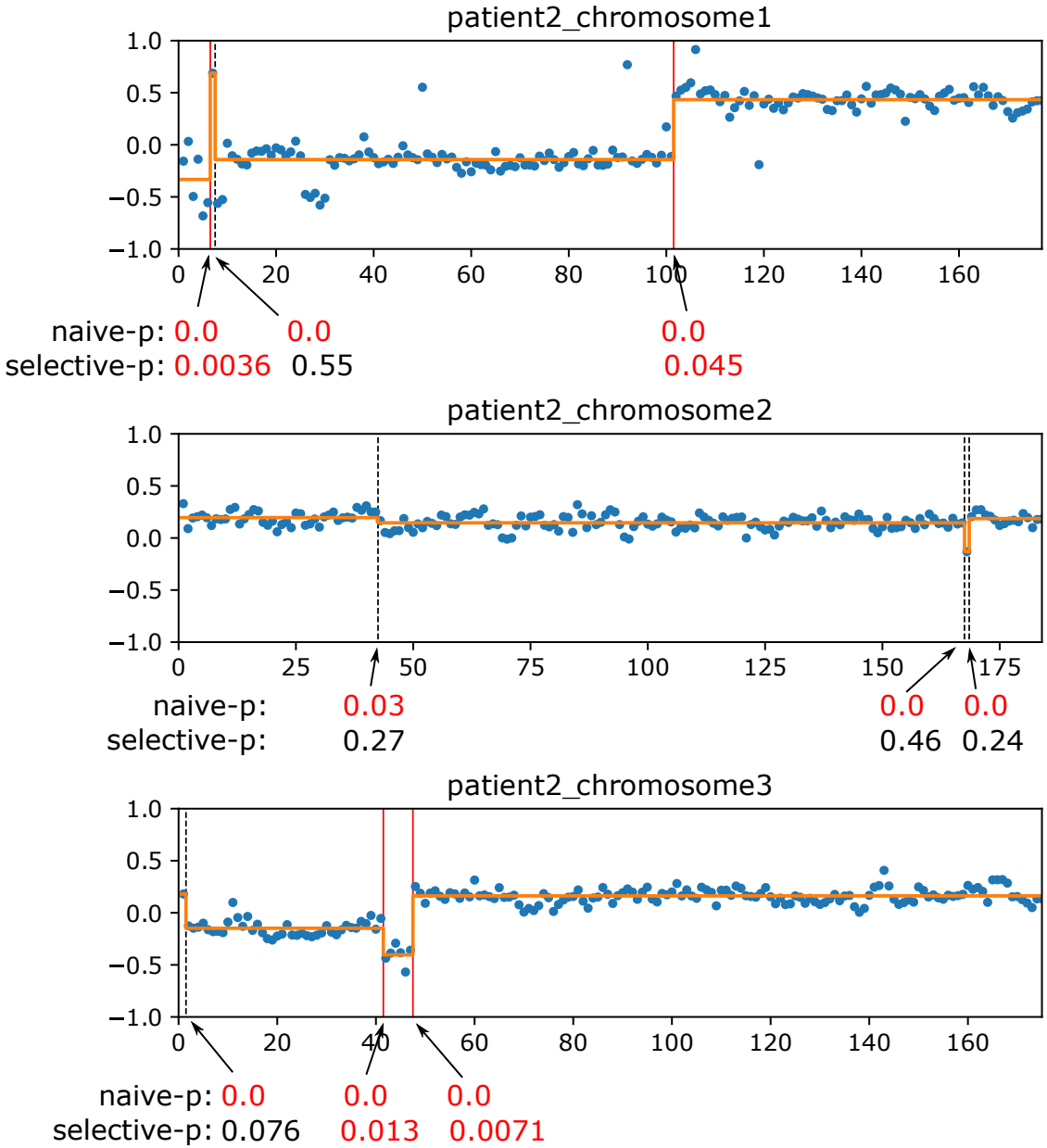
- Test statistic

$$\tau_a = \mathrm{sgn}(\boldsymbol{\delta}_a^\top \boldsymbol{x}) \boldsymbol{\delta}_a^\top \boldsymbol{x} = \boldsymbol{\eta}_a^\top \boldsymbol{x}$$

where

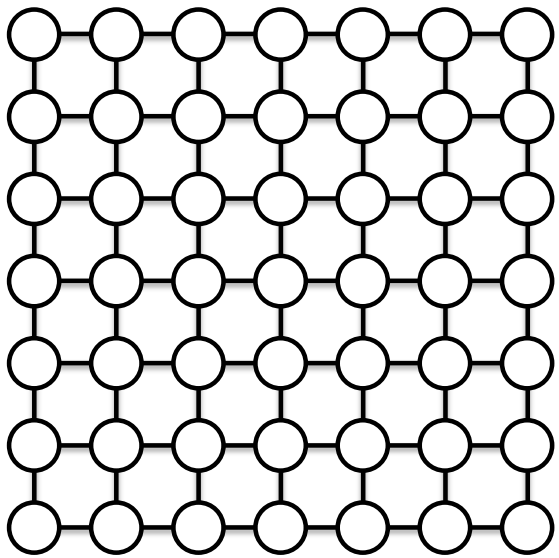
$$\boldsymbol{\delta}_a = \frac{1}{i_a - i_{a-1}} \mathbf{1}_{i_{a-1}+1,i_a} - \frac{1}{i_{a+1} - i_a} \mathbf{1}_{i_a+1,i_{a+1}}$$

ゲノムコピー数異常領域同定問題への応用結果



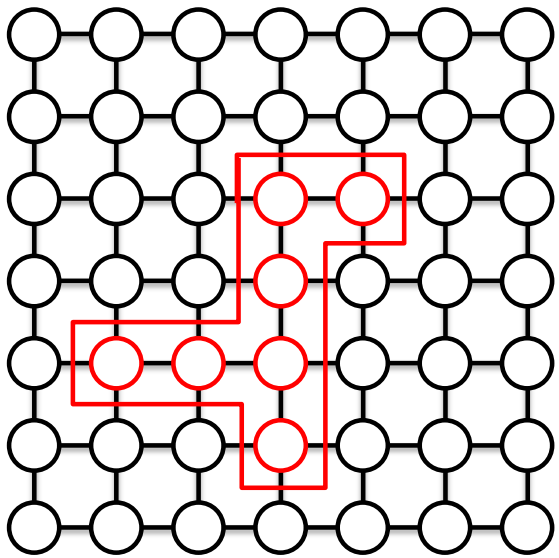
ポストセグメンテーション推論（グラフカット）

- グラフカットによる画像のセグメンテーション



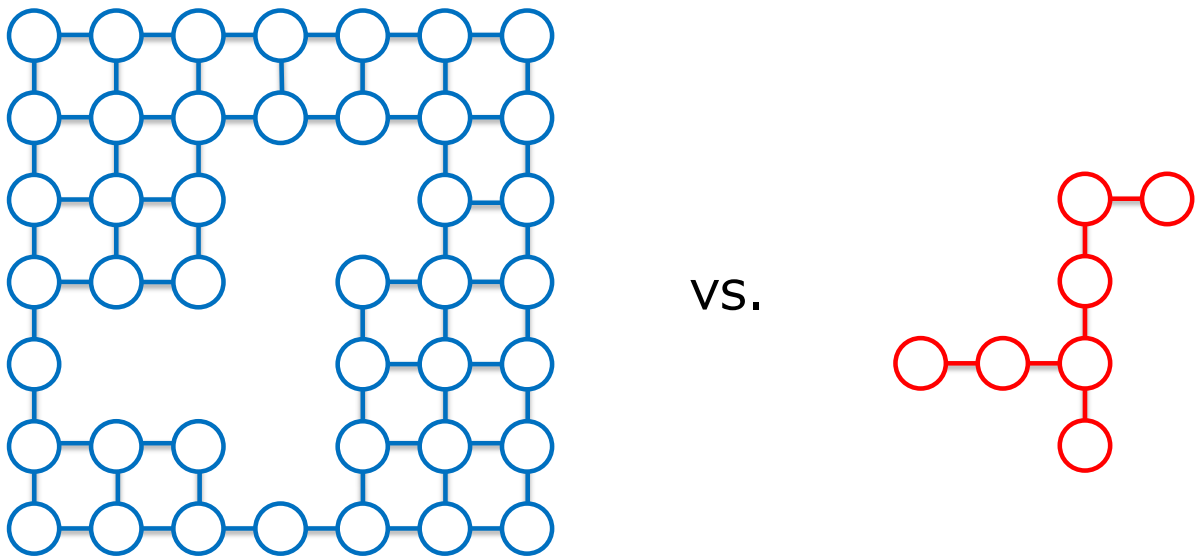
ポストセグメンテーション推論（グラフカット）

- グラフカットによる画像のセグメンテーション

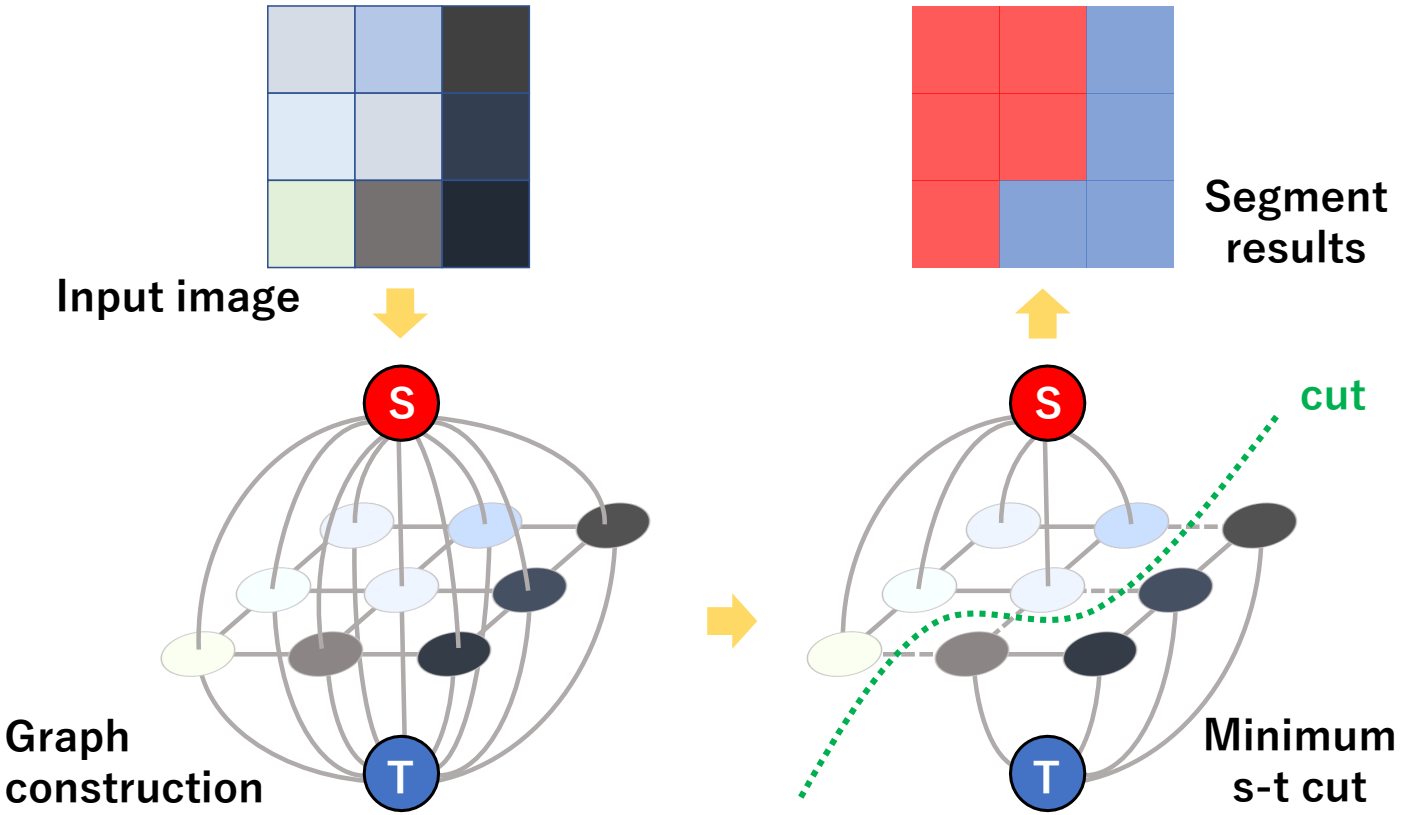


ポストセグメンテーション推論（グラフカット）

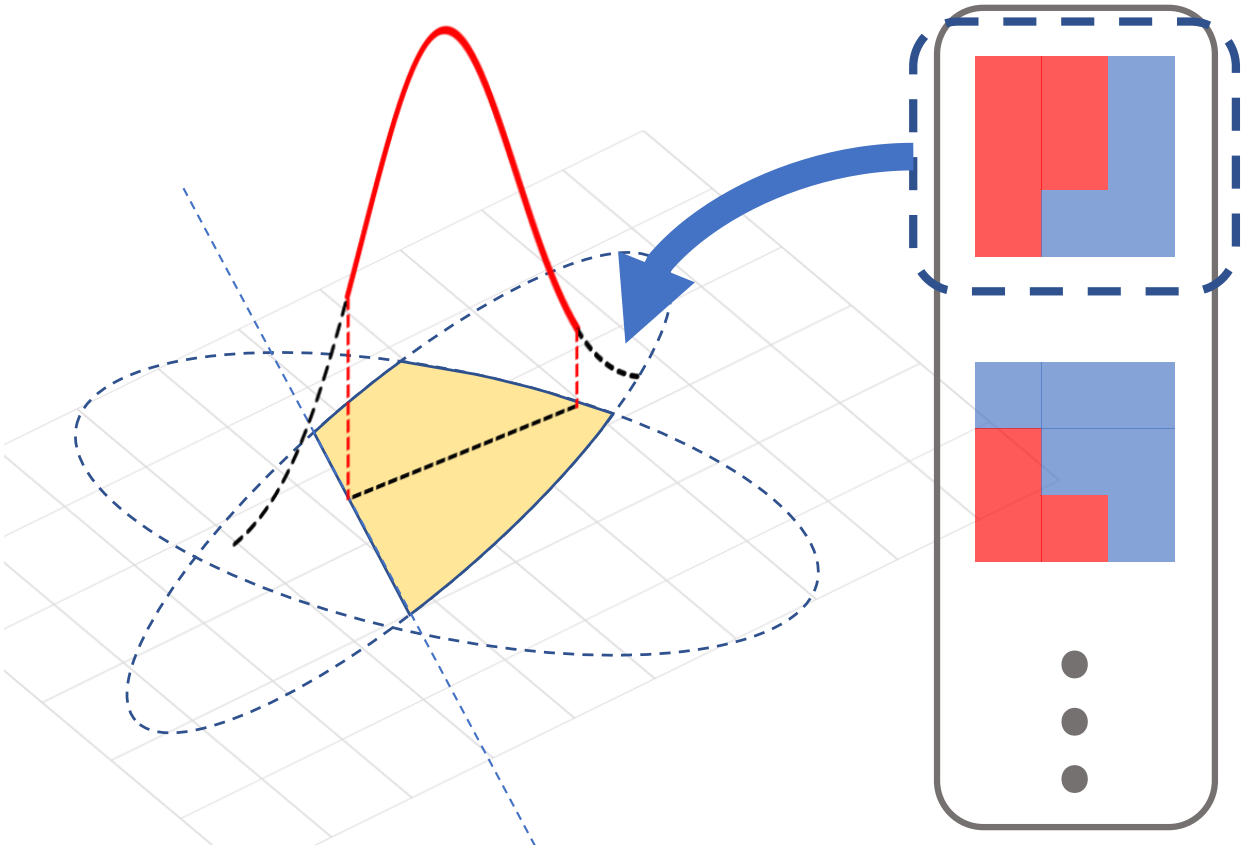
- グラフカットによる画像のセグメンテーション



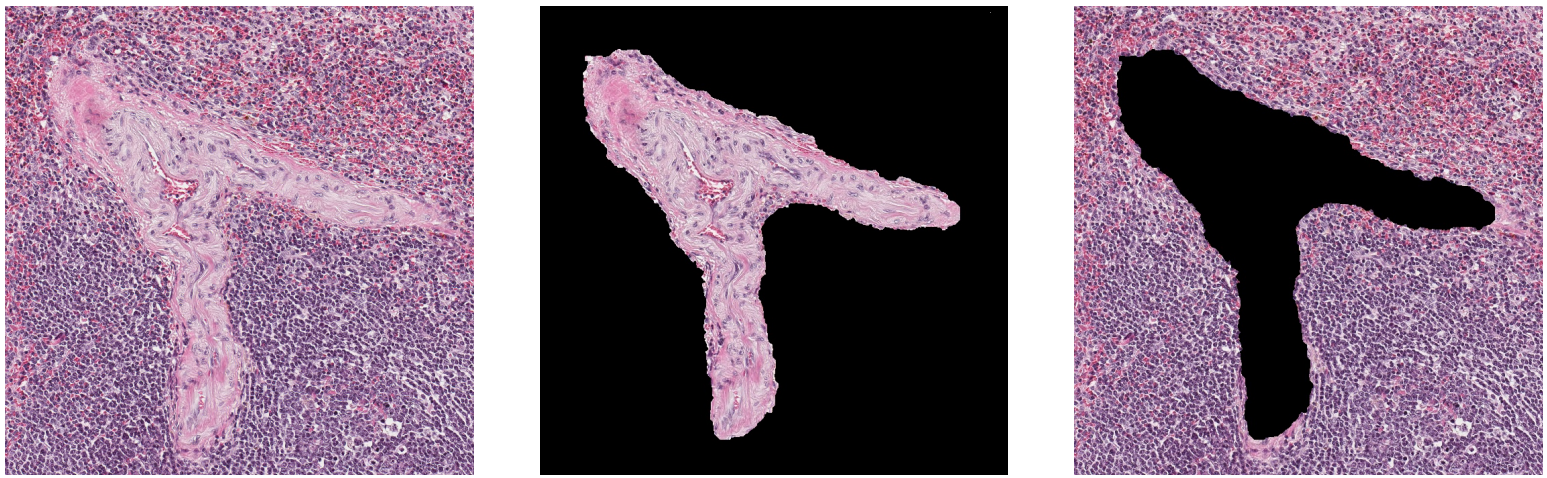
グラフカットによるセグメンテーション



グラフカットにおける選択的推論

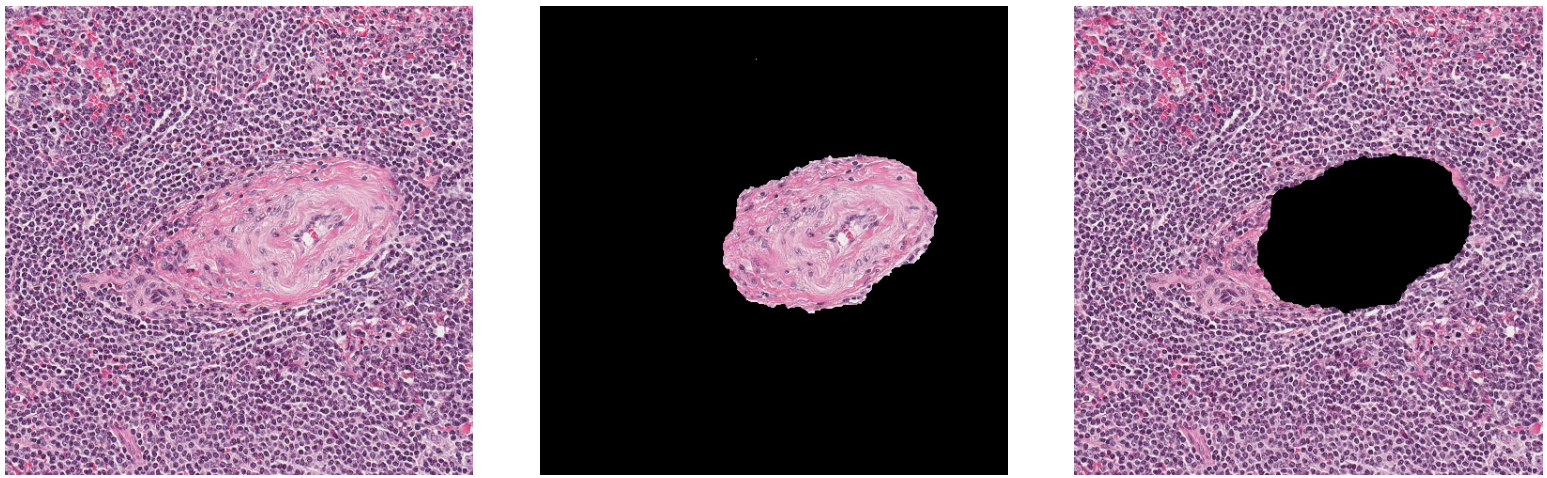


Application to Pathology 1



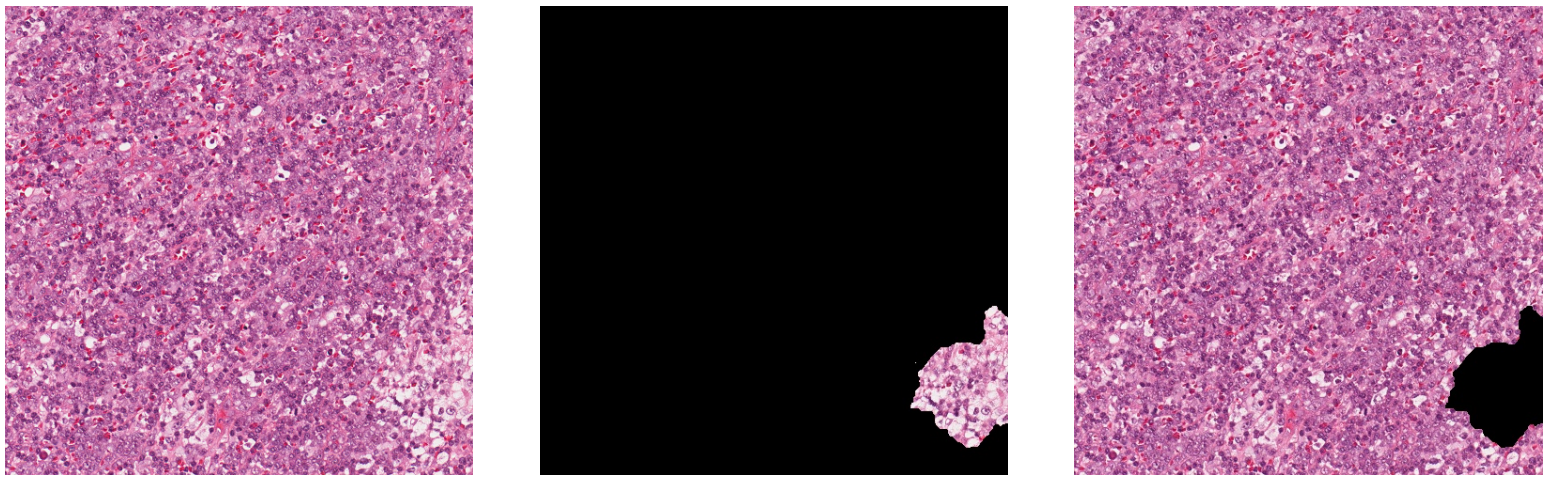
naive-p-value $= 0.000$, selective-p-value $= 0.000$

Application to Pathology 2



naive_p-value = 0.000, selective_p-value = 0.000

Application to Pathology 3

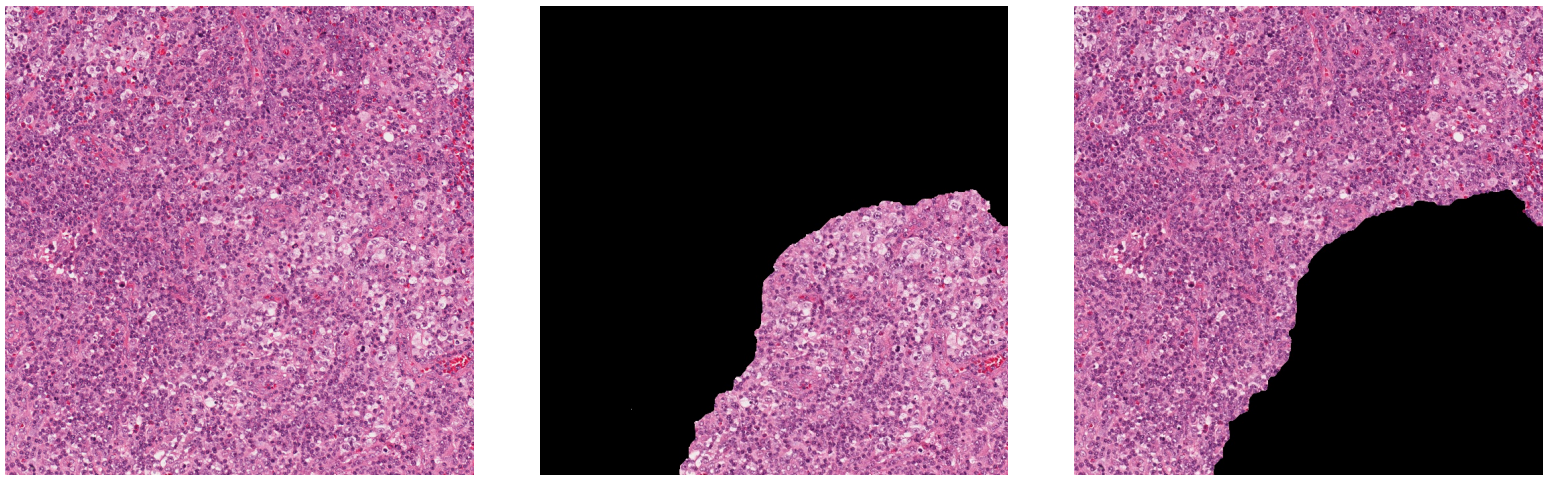


naive-p-value $= 0.000$, selective-p-value $= 0.351$

Application to Pathology 4

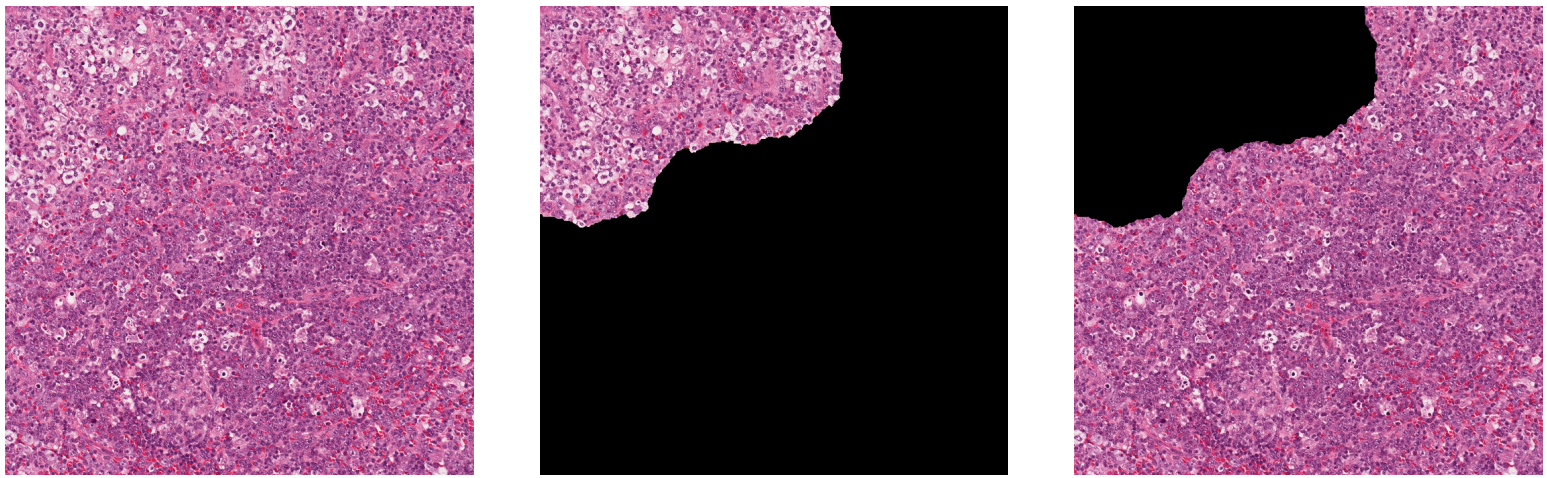
naive-p-value $= 0.000$, selective-p-value $= 0.800$

Application to Pathology 5



naive-p-value $= 0.000$, selective-p-value $= 1.000$

Application to Pathology 6



naive-p-value $= 0.000$, selective-p-value $= 0.024$

Part3のまとめ

- 選択的推論はLASSOなどの線形モデルの特徴選択問題以外にも幅広く利用可能
- ポストクラスタリング推論，ポストセグメンテーション推論（特に既存手法なし）に選択的推論が有効
- 一見複雑なアルゴリズムでも（頑張れば）仮説選択の逆像を二次不等式で記述（もしくは近似）可能

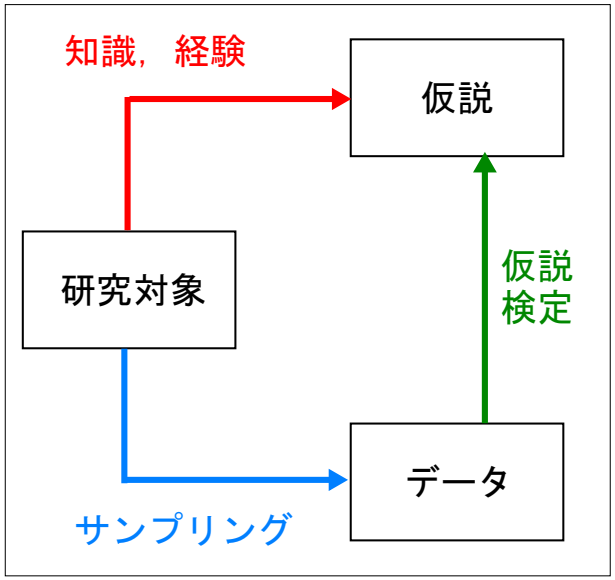
おわりに

- 機械学習により得られた知識の信頼性をいかに評価すればよいか $\Rightarrow$ 選択的推論は１つの有望なアプローチ
- 選択的推論は線形モデル特徴選択のための統計的推測法として研究されてきたが，より広い問題に適用可能
- 評価基準のないヒューリスティックなアルゴリズムや最適保証のないアルゴリズムに対しても利用可能（クラスタリングの例など）
- 正規性の仮定，分散既知の仮定などの制約を取り除くため，さらなる理論やアルゴリズムの研究が必要

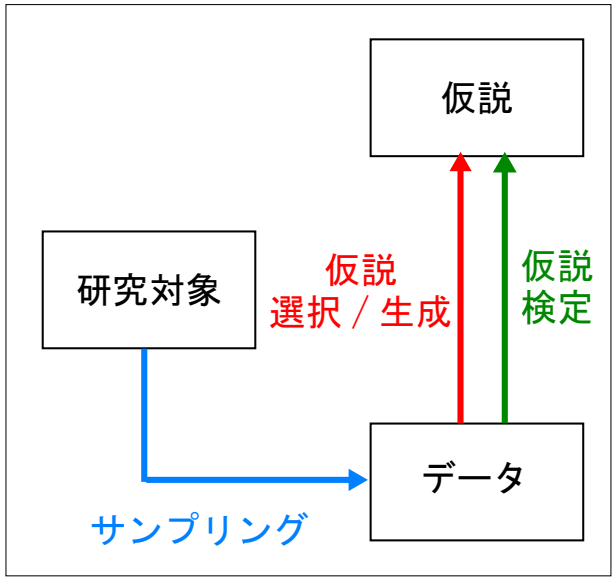
参考文献

- R. Barber, E. Candes. A knockoff filter for high-dimensional selective inference. arXiv, 2016.
- R. Berk et al. Valid post-selection inference. The Annals of Statistics, 2013.
- W. Fithian, D. Sun, J. Taylor. Optimal inference after model selection. arXiv, 2014.
- S. Hyun, M. G'Sell, R. Tibshirani. Exact Post-Selection Inference for Changepoint Detection and Other Generalized Lasso Problems. arXiv, 2016.
- J. Lee, Y. Sun, J. Taylor. Evaluating the statistical significance of biclusters. NIPS 2015.
- J. Lee, D. Sun, Y. Sun, J. Taylor. Exact post-selection inference, with application to the lasso. The Annals of Statistics, 2016.
- J. Lee, J. Taylor. Exact post model selection inference for marginal screening. NIPS 2014.
- R. Lockhart, J. Taylor, R. Tibshirani, R. Tibshirani. A significance test for the lasso. Annals of statistics. 2014.
- J. Taylor and R. Tibshirani. Post-selection inference for L1-penalized likelihood models. arXiv, 2016.
- X. Tian, J. Taylor. Asymptotics of selective inference. arXiv, 2015.
- F. Yang, R. Barber, P. Jain, J. Lafferty. Selective Inference for Group-Sparse Linear Models. arXiv, 2016.
- S. Suzumura, K. Nakagawa, Y. Umezu, K. Tsuda, I. Takeuchi. Selective Inference for Sparse Higher-Order Interaction Models. ICML 2017.
- M. Yamada, Y. Umezu, K. Fukumizu, I. Takeuchi. Post-selection inference with kernels. AISTATS 2019.

知識駆動型科学とデータ駆動型科学



知識駆動型科学における仮説検証



データ駆動型科学における仮説検証

データにより選択・生成された仮説には選択バイアスが生じる

"Collect data first, then ask questions" by E. Candes (2015)