Selective Inference for Deep Learning Model-driven Hypotheses

Ichiro Takeuchi

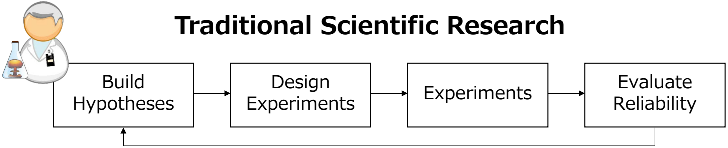
RIKEN AIP (Data-driven Biomedical Science Team)

This is joint work with V.N.L Duy, D. Miwa, and S.Iwazaki.

Brief self-introduction

- Professor at Nagoya Institute of Technology
- Team leader of data-driven biomedical science team at RIKEN AIP
- Mission: develop AI and ML methods for data-driven science and their applications to biology, medicine, and material science

Data-driven science



Traditional Scientific Research

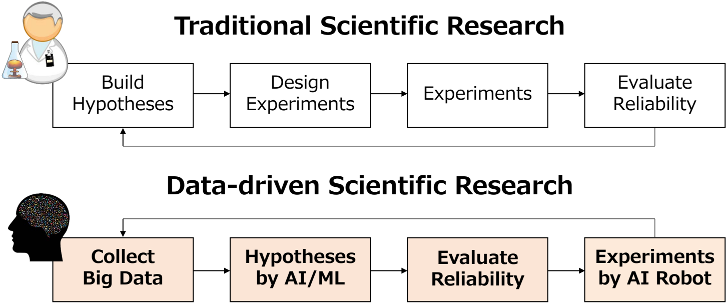
Build Hypotheses

Design Experiments

Experiments

Evaluate Reliability

Data-driven science



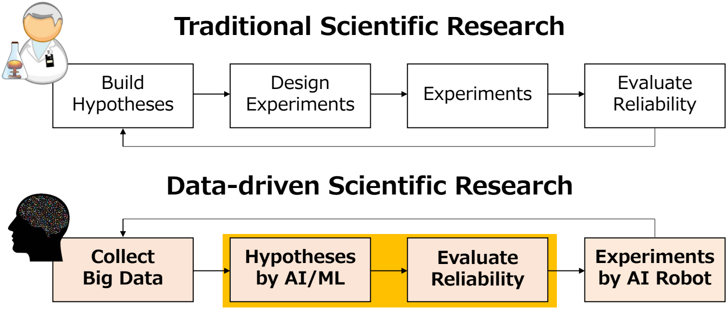
Traditional Scientific Research

Build Hypotheses → Design Experiments → Experiments → Evaluate Reliability

Data-driven Scientific Research

Collect Big Data → **Hypotheses by AI/ML** → **Evaluate Reliability** → **Experiments by AI Robot**

Data-driven science



Traditional Scientific Research

Build Hypotheses → Design Experiments → Experiments → Evaluate Reliability

Data-driven Scientific Research

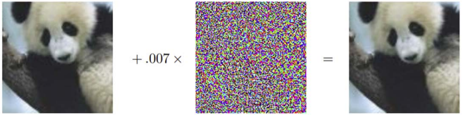
Collect Big Data → **Hypotheses by AI/ML** → **Evaluate Reliability** → **Experiments by AI Robot**

AI and ML in practice

- ELSI (Ethical, Legal and Social Issues)
    - Fairness
    - SDG
- Interpretability
    - Visualization
    - Rule Extraction
- Reliability
    - Robustness
    - Statistical Significance

Reliability in AI

- Robustness: the complexity of AI increases the risk that a small change in the data leads to a big change in the result.

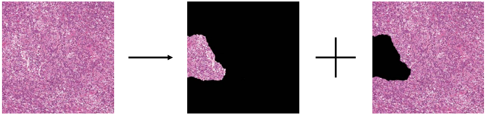


panda gibborn

Goodfellow et al. (ICLR2015) Fig.1

- Statistical significance: The flexibility of AI increases the risk of finding false positive (FP) results (which seems meaningful but is just an artifact).

An example of medical image segmentation

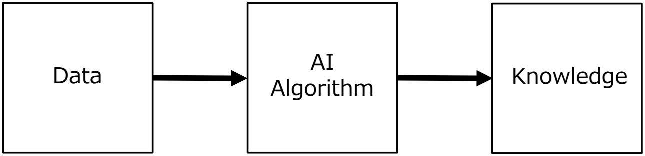


original image attention (object) region background region

(Traditional) Naive p-value $= 0.000$ (statistically significant)

Uncertainty quantification

- For evaluating statistical reliability of the knowledge obtained by AI, **uncertainty quantification** of the knowledge is needed.



Data

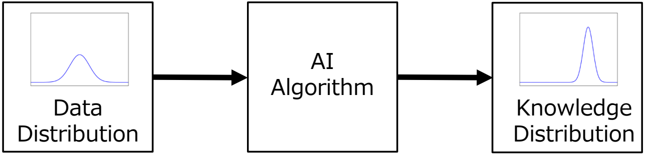
AI Algorithm

Knowledge

- Frequentist approach (sampling distribution)
  - Exact inference (deriving exact sampling distribution)
  - Randomized inference
  - Asymptotic inference
- Bayesian approach (posterior distribution)
  - Exact Bayesian inference
  - MCMC
  - Variational inference
- Uncertainty quantification approaches in deep neural network (DNN)
  - Dropout
  - Ensemble learning
  - Bayesian NN

Uncertainty quantification

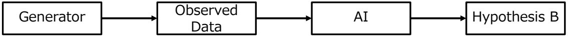
- For evaluating statistical reliability of the knowledge obtained by AI, **uncertainty quantification** of the knowledge is needed.



- Frequentist approach (sampling distribution)
    - Exact inference (deriving exact sampling distribution)
    - Randomized inference
    - Asymptotic inference
- Bayesian approach (posterior distribution)
    - Exact Bayesian inference
    - MCMC
    - Variational inference
- Uncertainty quantification approaches in deep neural network (DNN)
    - Dropout
    - Ensemble learning
    - Bayesian NN

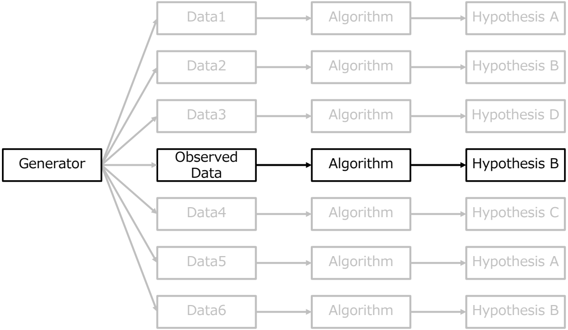
Probabilistic data generation model

- (Frequentist) statistical inference framework (parallel world interpretation)



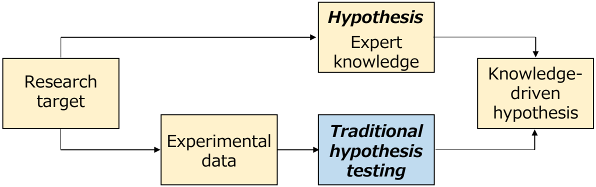
Probabilistic data generation model

- (Frequentist) statistical inference framework (parallel world interpretation)



Knowledge-driven science and data-driven science

(Traditional) Knowledge-driven science



Research target

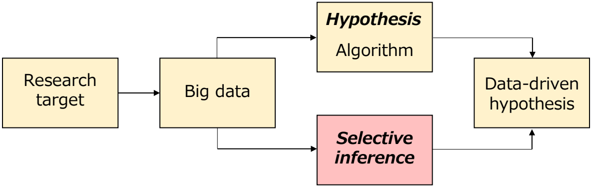
Experimental data

Hypothesis
Expert knowledge

Traditional hypothesis testing

Knowledge-driven hypothesis

Data-driven science



Research target

Big data

Hypothesis
Algorithm

Selective inference

Data-driven hypothesis

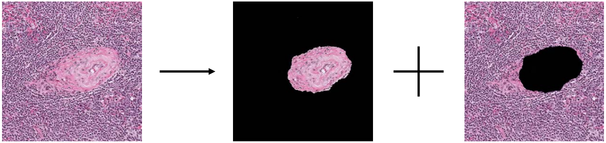
Outline

- Part 1: Hypothesis selection bias and multiple comparison
- Part 2: Conditional Selective Inference (SI)
- Part 3: Conditional SI for deep neural network (DNN)

Part 1: Hypothesis Selection Bias and Multiple Comparison

Problem 1: medical image segmentation

- Goal: Identify the attention (object) region in a medical image by segmentation



original image object background

- An image is represented as an n-dimensional random vector of pixel values $\boldsymbol{X} \in \mathbb{R}^n$ as

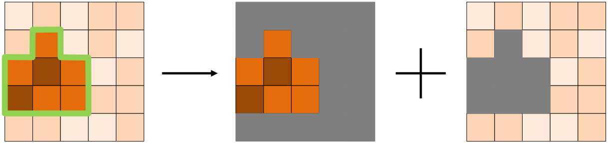
$$\underbrace{\boldsymbol{X}}_{\text{(random) vector}} = \underbrace{\boldsymbol{M}}_{\text{true vector}} + \underbrace{\boldsymbol{\varepsilon}}_{\text{noise vector}}, \qquad \underbrace{\boldsymbol{\varepsilon} \sim \mathbb{N}(\mathbf{0}, \Sigma)}_{\text{Normally-distributed noise}}$$

- Segmentation algorithm $\mathcal{A}$

$$\underbrace{\mathcal{A}}_{\text{algorithm}} : \underbrace{\boldsymbol{X}}_{\text{(random) image}} \mapsto \{ \underbrace{\mathcal{O}_{\boldsymbol{X}}}_{\text{pixels in object}}, \underbrace{\mathcal{B}_{\boldsymbol{X}}}_{\text{pixels in background}} \}$$

Problem 1: medical image segmentation

- Goal: Identify the attention (object) region in a medical image by segmentation



- An image is represented as an n-dimensional random vector of pixel values $\boldsymbol{X} \in \mathbb{R}^n$ as

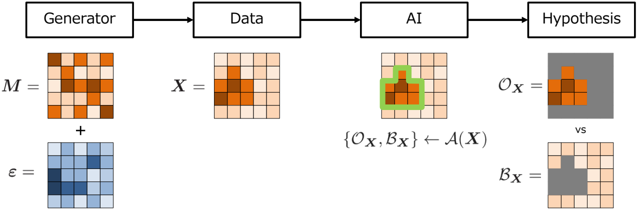
$$\underbrace{\boldsymbol{X}}_{\text{(random) vector}} = \underbrace{\boldsymbol{M}}_{\text{true vector}} + \underbrace{\boldsymbol{\varepsilon}}_{\text{noise vector}}, \quad \underbrace{\boldsymbol{\varepsilon} \sim \mathbb{N}(\mathbf{0}, \Sigma)}_{\text{Normally-distributed noise}}$$

- Segmentation algorithm $\mathcal{A}$

$$\underbrace{\mathcal{A}}_{\text{algorithm}} : \underbrace{\boldsymbol{X}}_{\text{(random) image}} \mapsto \{ \underbrace{\mathcal{O}_{\boldsymbol{X}}}_{\text{pixels in object}}, \underbrace{\mathcal{B}_{\boldsymbol{X}}}_{\text{pixels in background}} \}$$

Data-driven knowledge discovery

- Data-driven hypothesis



Statistical hypothesis testing

- Statistical hypothesis testing
    - Null hypothesis

$$\mathrm{H}_0 : \underbrace{\frac{1}{|\mathcal{O}_{\boldsymbol{X}}|} \sum_{i \in \mathcal{O}_{\boldsymbol{X}}} M_i}_{\text{mean pixel value in object}} = \underbrace{\frac{1}{|\mathcal{B}_{\boldsymbol{X}}|} \sum_{i \in \mathcal{B}_{\boldsymbol{X}}} M_i}_{\text{mean pixel value in background}}$$

    - Alternative hypothesis

$$\mathrm{H}_1 : \underbrace{\frac{1}{|\mathcal{O}_{\boldsymbol{X}}|} \sum_{i \in \mathcal{O}_{\boldsymbol{X}}} M_i}_{\text{mean pixel value in object}} \neq \underbrace{\frac{1}{|\mathcal{B}_{\boldsymbol{X}}|} \sum_{i \in \mathcal{B}_{\boldsymbol{X}}} M_i}_{\text{mean pixel value in background}}$$

- Test statistic: Difference of mean pixel values between object and background regions

$$\Delta_{\boldsymbol{X}} := \frac{1}{|\mathcal{O}_{\boldsymbol{X}}|} \sum_{i \in \mathcal{O}_{\boldsymbol{X}}} X_i - \frac{1}{|\mathcal{B}_{\boldsymbol{X}}|} \sum_{i \in \mathcal{B}_{\boldsymbol{X}}} X_i$$

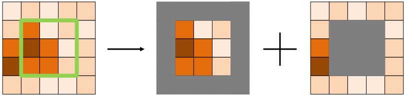
- Statistical significance (two-sided p-value)

$$p = \Pr\left(\underbrace{|\Delta_{\boldsymbol{X}}|}_{\text{random variable}} \geq \underbrace{|\Delta_{\boldsymbol{x}}|}_{\text{observation}} \right)$$

Knowledge-driven vs. data-driven hypotheses

- Knowledge-driven hypothesis: object/background regions do not depend on the data $\Rightarrow$ (traditional) z-test or t-test

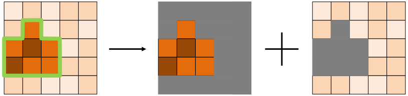
$$\underbrace{\frac{1}{|\mathcal{O}|}\sum_{i\in\mathcal{O}} M_i}_{\text{fixed mean pixel value in object}} \qquad = \text{ or } \neq \qquad \underbrace{\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} M_i}_{\text{fixed mean pixel value in background}}$$



original image　object　background

- Data-driven hypothesis: object/background regions are determined by the data $\Rightarrow$ data/algorithm dependent

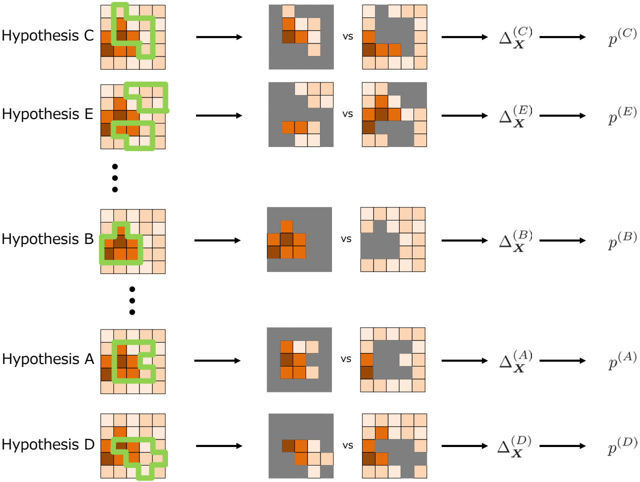
$$\underbrace{\frac{1}{|\mathcal{O}_{\boldsymbol{x}}|}\sum_{i\in\mathcal{O}_{\boldsymbol{x}}} M_i}_{\text{selected mean pixel value in object}} \qquad = \text{ or } \neq \qquad \underbrace{\frac{1}{|\mathcal{B}_{\boldsymbol{x}}|}\sum_{i\in\mathcal{B}_{\boldsymbol{x}}} M_i}_{\text{selected mean pixel value in background}}$$



original image　object　background

Multiple comparison, hypothesis selection, and selection bias

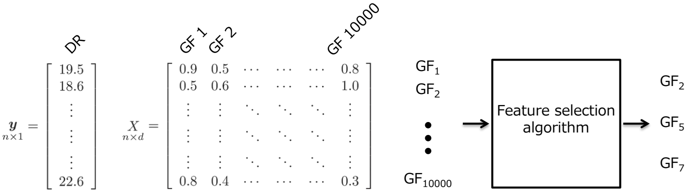
- The data-driven hypothesis is interpreted as the result of multiple comparison with all possible $2^{\#\mathrm{pixels}}$ segmentation results.



- Correction of the selection bias is indispensable in multiple comparison.

Problem 2: feature selection in linear models

- Goal: select a subset of 10,000 genes that are useful for predicting drug effects
- High-dimensional data and feature selection



$$\underset{n\times 1}{\boldsymbol{y}} = \left[\begin{array}{c} \text{DR} \\ 19.5 \\ 18.6 \\ \vdots \\ \vdots \\ \vdots \\ 22.6 \end{array}\right] \qquad \underset{n\times d}{X} = \left[\begin{array}{cccccc} \text{GF 1} & \text{GF 2} & & & & \text{GF 10000} \\ 0.9 & 0.5 & \cdots & \cdots & \cdots & 0.8 \\ 0.5 & 0.6 & \cdots & \cdots & \cdots & 1.0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0.8 & 0.4 & \cdots & \cdots & \cdots & 0.3 \end{array}\right]$$

GF_1
GF_2
$\vdots$
GF_{10000}

→ Feature selection algorithm →

GF_2
GF_5
GF_7

- Linear model with the selected features by least-square method

$$\hat{y}_i = \hat{\beta}_2 x_{i2} + \hat{\beta}_5 x_{i5} + \hat{\beta}_7 x_{i7},$$

where

$$\left[\begin{array}{c} \hat{\beta}_2 \\ \hat{\beta}_5 \\ \hat{\beta}_7 \end{array}\right] = \underset{\beta_2,\beta_5,\beta_7}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - (\beta_2 x_{i2} + \beta_5 x_{i5} + \beta_7 x_{i7})\right)^2$$

Problem 2: feature selection in linear models (problem formulation)

- Data ($n = 50$, $d = 10000$ in the example)

$$\boldsymbol{X} \in \mathbb{R}^{n \times d}, \boldsymbol{Y} \in \mathbb{R}^{n}$$

- Probabilistic model

$$\underbrace{\boldsymbol{Y}}_{\text{drug effect}} = \underbrace{\boldsymbol{\mu}(\boldsymbol{X})}_{\text{true drug effect}} + \underbrace{\boldsymbol{\varepsilon}}_{\text{noise}}, \qquad \underbrace{\boldsymbol{\varepsilon} \sim \mathbb{N}(\boldsymbol{0}, \Sigma)}_{\text{Normally-distributed noise}}$$

- Feature selection algorithm $\mathcal{A}$

$$\mathcal{A} : \boldsymbol{Y} \mapsto \mathcal{M}_{\boldsymbol{Y}},$$

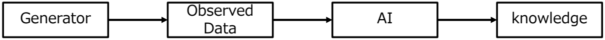
  where $\mathcal{M}_{\boldsymbol{Y}}$ is the set of selected features ($\mathcal{M}_{\boldsymbol{y}} = \{2, 5, 7\}$ in the example)

- Linear model with the selected features by least-square method

$$\hat{\boldsymbol{\beta}}_{\mathcal{M}_{\boldsymbol{Y}}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{|\mathcal{M}_{\boldsymbol{Y}}|}}{\operatorname{argmin}} \|\boldsymbol{Y} - \boldsymbol{X}_{\mathcal{M}_{\boldsymbol{Y}}}^{\top} \boldsymbol{\beta}\|^2 = (X_{\mathcal{M}_{\boldsymbol{Y}}}^{\top} X_{\mathcal{M}_{\boldsymbol{Y}}})^{-1} X_{\mathcal{M}_{\boldsymbol{Y}}}^{\top} \boldsymbol{Y}$$

AI and data-driven hypotheses

- Data-driven hypothesis



Generator → Observed Data → AI → knowledge

$\boldsymbol{\mu}(\boldsymbol{X}) + \varepsilon$ $\quad\quad$ $\boldsymbol{Y}$ $\quad\quad$ $\mathcal{M}_{\boldsymbol{Y}} \leftarrow \mathcal{A}(\boldsymbol{Y})$ $\quad\quad$ $\{\hat{\beta}_{\mathcal{M}_{\mathcal{Y}},j}\}_{j\in\mathcal{M}_{\boldsymbol{Y}}}$

Statistical hypothesis testing

- Statistical hypothesis testing
  - Population least-square solution

$$\boldsymbol{\beta}_{\mathcal{M}_{\boldsymbol{Y}}} = (X_{\mathcal{M}_{\boldsymbol{Y}}}^\top X_{\mathcal{M}_{\boldsymbol{Y}}})^{-1} X_{\mathcal{M}_{\boldsymbol{Y}}}^\top \underbrace{\boldsymbol{\mu}(\boldsymbol{X})}_{\text{true drug effect}}$$

  - Null hypothesis

$$\mathrm{H}_0 : \underbrace{\beta_{\mathcal{M}_{\boldsymbol{Y}},j}}_{\text{effect of the selected feature } j} = 0$$

  - Alternative hypothesis

$$\mathrm{H}_1 : \underbrace{\beta_{\mathcal{M}_{\boldsymbol{Y}},j}}_{\text{effect of the selected feature } j} \neq 0$$

- Test-statistic

$$\hat{\beta}_{\mathcal{M}_{\boldsymbol{Y}},j} = (X_{\mathcal{M}_{\boldsymbol{Y}}}^\top X_{\mathcal{M}_{\boldsymbol{Y}}})^{-1} X_{\mathcal{M}_{\boldsymbol{Y}}}^\top \underbrace{\boldsymbol{Y}}_{\text{observed drug effect}}$$

- Statistical significance (two-sided p-value)

$$p = \Pr\left(\underbrace{|\hat{\beta}_{\mathcal{M}_{\boldsymbol{Y}},j}|}_{\text{random var.}} \geq \underbrace{|\hat{\beta}_{\mathcal{M}_{\boldsymbol{y}},j}|}_{\text{observation}} \right)$$

Knowledge-driven hypotheses and data-driven hypotheses

- Knowledge-driven hypothesis: the set of features are selected without looking at the data $\Rightarrow$ (traditional) z-test or t-test

$$\underbrace{\beta_{\mathcal{M},j}}_{\text{effect of the selected feature } j \text{ for the fixed model}} = \text{ or } \neq 0$$

- Data-driven hypothesis: the set of features are selected by the data $\Rightarrow$ data/algorithm dependent

$$\underbrace{\beta_{\mathcal{M}_{\boldsymbol{Y}},j}}_{\text{effect of the selected feature } j \text{ for the selected model}} = \text{ or } \neq 0$$

Multiple comparison, hypothesis selection, and selection bias

- This data-driven hypothesis is interpreted as the result of multiple comparison with $2^{\#\text{features} \times \#\text{selected feature}}$ hypotheses.

$$\text{Hypothesis C} \quad \mathcal{M}^{(C)} = \{3, 7\} \longrightarrow \begin{matrix} \beta_{\mathcal{M}^{(C)},3} \neq 0 \\ \beta_{\mathcal{M}^{(C)},7} \neq 0 \end{matrix} \longrightarrow \begin{matrix} \hat{\beta}_{\mathcal{M}^{(C)},3} \\ \hat{\beta}_{\mathcal{M}^{(C)},7} \end{matrix} \longrightarrow \begin{matrix} p_3^{(C)} \\ p_7^{(C)} \end{matrix}$$

$$\text{Hypothesis E} \quad \mathcal{M}^{(E)} = \{2\} \longrightarrow \beta_{\mathcal{M}^{(E)},2} \neq 0 \longrightarrow \hat{\beta}_{\mathcal{M}^{(E)},2} \longrightarrow p_2^{(E)}$$

$$\vdots$$

$$\text{Hypothesis B} \quad \mathcal{M}^{(B)} = \{2, 5, 7\} \longrightarrow \begin{matrix} \beta_{\mathcal{M}^{(B)},2} \neq 0 \\ \beta_{\mathcal{M}^{(B)},5} \neq 0 \\ \beta_{\mathcal{M}^{(B)},7} \neq 0 \end{matrix} \longrightarrow \begin{matrix} \hat{\beta}_{\mathcal{M}^{(B)},2} \\ \hat{\beta}_{\mathcal{M}^{(B)},5} \\ \hat{\beta}_{\mathcal{M}^{(B)},7} \end{matrix} \longrightarrow \begin{matrix} p_2^{(B)} \\ p_5^{(B)} \\ p_7^{(B)} \end{matrix}$$

$$\vdots$$

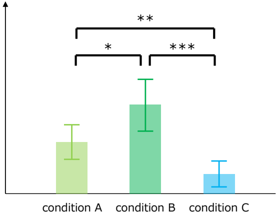
$$\text{Hypothesis A} \quad \mathcal{M}^{(A)} = \{9\} \longrightarrow \beta_{\mathcal{M}^{(A)},9} \neq 0 \longrightarrow \hat{\beta}_{\mathcal{M}^{(A)},9} \longrightarrow p_9^{(A)}$$

$$\text{Hypothesis D} \quad \mathcal{M}^{(D)} = \{3, 8\} \longrightarrow \begin{matrix} \beta_{\mathcal{M}^{(D)},3} \neq 0 \\ \beta_{\mathcal{M}^{(D)},8} \neq 0 \end{matrix} \longrightarrow \begin{matrix} \hat{\beta}_{\mathcal{M}^{(D)},3} \\ \hat{\beta}_{\mathcal{M}^{(D)},8} \end{matrix} \longrightarrow \begin{matrix} p_3^{(D)} \\ p_8^{(D)} \end{matrix}$$

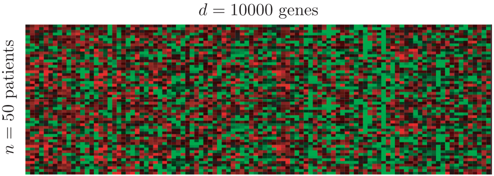
- Correction of the selection bias is indispensable in multiple comparison.

Multiple comparison

- In the context of traditional multiple hypothesis testing, only a handful of tests are considered.



- In the context of genetic data analysis (2000∼), large-scale multiple comparison with tens of thousands of hypotheses were considered.



- The number of all possible hypotheses that AI/ML can produce is much more than the existing methods can handle.

Three approaches for multiple comparison correction

- **Family-wise error rate (FWER) control**: controlling the probability of finding a false positive (FP) $< \alpha$ (e.g., 0.05)
- **False discover rate (FDR)**: controlling the expected proportion of discoveries that are false $< \alpha$ (e.g., 0.05)
- **Conditional selective inference (SI)**: controlling the probability of finding a FP conditional on the hypothesis selection event $< \alpha$ (e.g., 0.05)

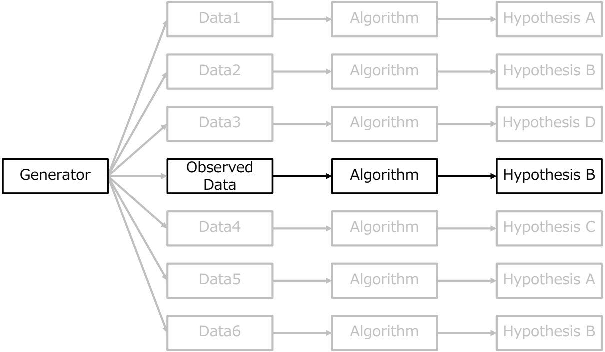
Summary of part 1

- Knowledge obtained by AI/ML algorithm is considered as data-driven hypotheses.
- Statistical reliability of data-driven hypotheses cannot be properly evaluated with traditional statistical inference due to the selection bias.
- This problem can be interpreted as a huge-scale multiple comparison problem where the one is selected from all possible hypotheses that AI/ML can produce.

Part 2: Conditional Selective Inference (SI)

Basic idea of conditional SI

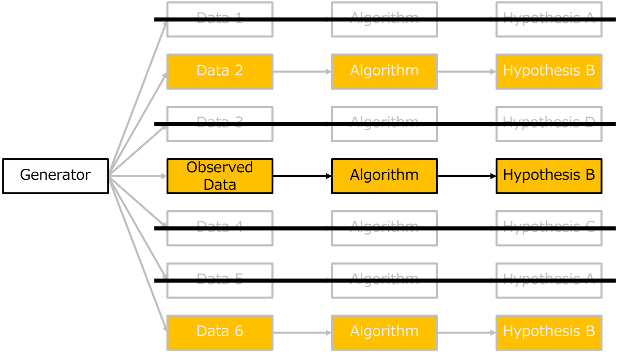
- The key idea of conditional SI is to consider only the cases (parallel worlds) where the same hypothesis is selected.



- Intuitively, by considering only the randomness where the same hypothesis is selected, the hypothesis selection bias disappears.

Basic idea of conditional SI

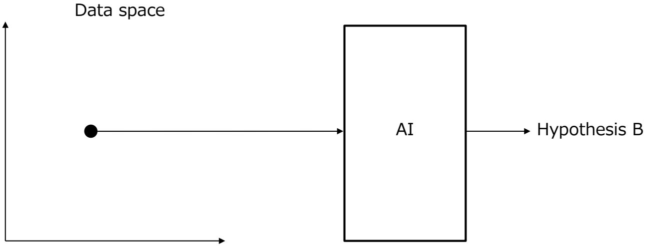
- The key idea of conditional SI is to consider only the cases (parallel worlds) where the same hypothesis is selected.



- Intuitively, by considering only the randomness where the same hypothesis is selected, the hypothesis selection bias disappears.

Basic idea of conditional SI

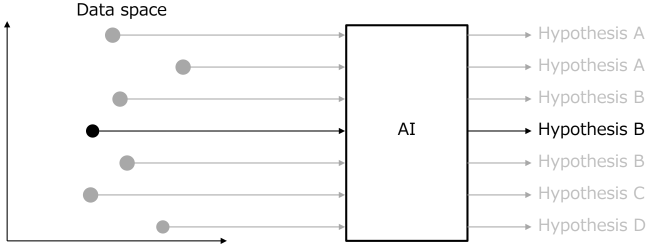
- The key idea of conditional SI is to consider only the cases (parallel worlds) where the same hypothesis is selected.



- Intuitively, by considering only the randomness in the subset of the data space, the hypothesis selection bias disappears.

Basic idea of conditional SI

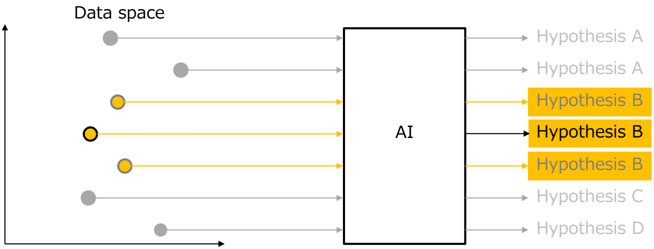
- The key idea of conditional SI is to consider only the cases (parallel worlds) where the same hypothesis is selected.



- Intuitively, by considering only the randomness in the subset of the data space, the hypothesis selection bias disappears.

Basic idea of conditional SI

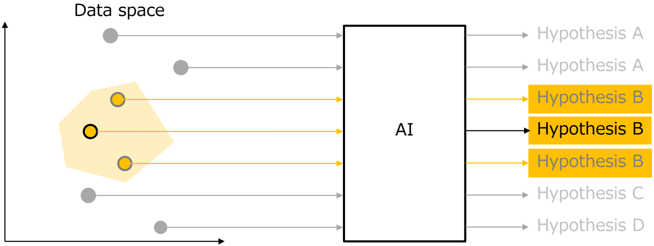
- The key idea of conditional SI is to consider only the cases (parallel worlds) where the same hypothesis is selected.



- Intuitively, by considering only the randomness in the subset of the data space, the hypothesis selection bias disappears.

Basic idea of conditional SI

- The key idea of conditional SI is to consider only the cases (parallel worlds) where the same hypothesis is selected.



- Intuitively, by considering only the randomness in the subset of the data space, the hypothesis selection bias disappears.

Conditional SI for medical image segmentation problem

- Ordinary statistical significance (p-value)

$$p = \Pr\left(\underbrace{|\Delta_{\boldsymbol{X}}|}_{\text{random var.}} \geq \underbrace{|\Delta_{\boldsymbol{x}}|}_{\text{observation}} \right)$$

- Conditional statistical significance (selective p-value)

$$p = \Pr\left(\underbrace{|\Delta_{\boldsymbol{X}}|}_{\text{random var.}} \geq \underbrace{|\Delta_{\boldsymbol{x}}|}_{\text{observation}} \;\middle|\; \underbrace{\{\mathcal{O}_{\boldsymbol{X}}, \mathcal{B}_{\boldsymbol{X}}\} = \{\mathcal{O}_{\boldsymbol{x}}, \mathcal{B}_{\boldsymbol{x}}\}}_{\text{the same object/background are selected}} \right)$$

- The main challenge of conditional SI is to characterize the selection event and compute the conditional probability.

A simple example of conditional SI for image segmentation

- A simple segmentation algorithm based on a threshold θ

$$\begin{aligned}\mathcal{O}_{\boldsymbol{X}} &= \{X_i \geq \theta\}, \\ \mathcal{B}_{\boldsymbol{X}} &= \{X_i < \theta\}\end{aligned}$$

- Selection event

$$\begin{aligned}&X_i \geq \theta, i \in \mathcal{O}_{\boldsymbol{x}}, \\ &X_i < \theta, i \in \mathcal{B}_{\boldsymbol{x}}\end{aligned}$$

- Selective p-value

$$\begin{aligned}p &= \Pr\left(\underbrace{|\Delta_{\boldsymbol{X}}|}_{\text{random var.}} \geq \underbrace{|\Delta_{\boldsymbol{x}}|}_{\text{observation}} \;\middle|\; \underbrace{\{\mathcal{O}_{\boldsymbol{X}}, \mathcal{B}_{\boldsymbol{X}}\} = \{\mathcal{O}_{\boldsymbol{x}}, \mathcal{B}_{\boldsymbol{x}}\}}_{\text{the same object/background are selected}}\right) \\ &= \Pr\left(\underbrace{|\Delta_{\boldsymbol{X}}|}_{\text{random var.}} \geq \underbrace{|\Delta_{\boldsymbol{x}}|}_{\text{observation}} \;\middle|\; \underbrace{X_i \geq \theta, i \in \mathcal{O}_{\boldsymbol{x}}, X_i < \theta, i \in \mathcal{B}_{\boldsymbol{x}}}_{\text{the same object/background are selected}}\right)\end{aligned}$$

Conditional SI for feature selection

- Naive p-value:

$$p = \Pr\left(\underbrace{|\hat{\beta}_{\mathcal{M}_{\boldsymbol{Y}},j}|}_{\text{random var.}} \geq \underbrace{|\hat{\beta}_{\mathcal{M}_{\boldsymbol{y}},j}|}_{\text{observation}} \right)$$

- Selective p-value:

$$p = \Pr\left(\underbrace{|\hat{\beta}_{\mathcal{M}_{\boldsymbol{Y}},j}|}_{\text{random var.}} \geq \underbrace{|\hat{\beta}_{\mathcal{M}_{\boldsymbol{y}},j}|}_{\text{observation}} \;\middle|\; \underbrace{\mathcal{M}_{\boldsymbol{Y}} = \mathcal{M}_{\boldsymbol{y}}}_{\text{the same set of features are selected}} \right)$$

A simple example of conditional SI for feature selection

- Marginal screening: select k features whose correlation between $\boldsymbol{Y} \in \mathbb{R}^n$ and $\boldsymbol{x}_j \in \mathbb{R}^n$ are large:

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

  Note that the correlation is represented as an inner product when variables are standardized.

- Hypothesis selection event

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

- Selective p-value

$$p = \Pr\left(\underbrace{|\hat{\beta}_{\mathcal{M}_{\boldsymbol{Y}},j}|}_{\text{random var.}} \ge \underbrace{|\hat{\beta}_{\mathcal{M}_{\boldsymbol{y}},j}|}_{\text{observation}} \;\middle|\; \underbrace{\mathcal{M}_{\boldsymbol{Y}} = \mathcal{M}_{\boldsymbol{y}}}_{\text{the same set of features are selected}} \right)$$

$$= \Pr\left(\underbrace{|\hat{\beta}_{\mathcal{M}_{\boldsymbol{Y}},j}|}_{\text{random var.}} \ge \underbrace{|\hat{\beta}_{\mathcal{M}_{\boldsymbol{y}},j}|}_{\text{observation}} \;\middle|\; \underbrace{\left\{ {\boldsymbol{x}_{(\ell)}}^\top \boldsymbol{Y} \ge {\boldsymbol{x}_{(m)}}^\top \boldsymbol{Y} \right\}_{(\ell,m)\in\{1,\ldots,k\}\times\{k+1,\ldots,d\}}}_{\text{the same set of features are selected}} \right)$$

Hypothesis selection event

Segmentation problem

Hypothesis selection event

$$X_i \geq \theta \text{ if } x_i \in \mathcal{O}_{\boldsymbol{x}}, X_i < \theta \text{ if } x_i \in \mathcal{B}_{\boldsymbol{x}}$$

Feature selection problem

Hypothesis selection event

$$\left\{ {\boldsymbol{x}_{(\ell)}}^\top \boldsymbol{Y} \geq {\boldsymbol{x}_{(m)}}^\top \boldsymbol{Y} \right\}_{(\ell,m)\in\{1,\ldots,k\}\times\{k+1,\ldots,d\}}$$

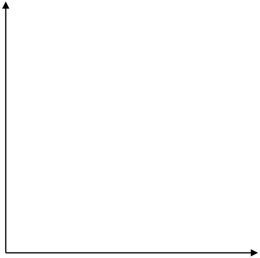
Hypothesis selection event

Segmentation problem

Hypothesis selection event

$$X_i \geq \theta \text{ if } x_i \in \mathcal{O}_{\boldsymbol{x}}, X_i < \theta \text{ if } x_i \in \mathcal{B}_{\boldsymbol{x}}$$

Data space $\boldsymbol{X} \in \mathbb{R}^n$

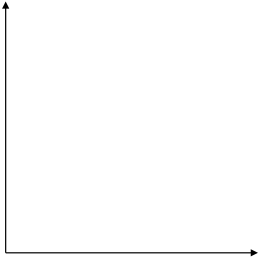


Feature selection problem

Hypothesis selection event

$$\left\{ {\boldsymbol{x}_{(\ell)}}^\top \boldsymbol{Y} \geq {\boldsymbol{x}_{(m)}}^\top \boldsymbol{Y} \right\}_{(\ell,m) \in \{1,\ldots,k\} \times \{k+1,\ldots,d\}}$$

Data space $\boldsymbol{Y} \in \mathbb{R}^n$



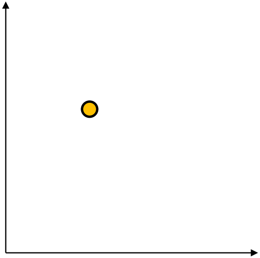
Hypothesis selection event

<u>Segmentation problem</u>

Hypothesis selection event

$X_i \geq \theta \text{ if } x_i \in \mathcal{O}_{\boldsymbol{x}}, X_i < \theta \text{ if } x_i \in \mathcal{B}_{\boldsymbol{x}}$

Data space $\boldsymbol{X} \in \mathbb{R}^n$

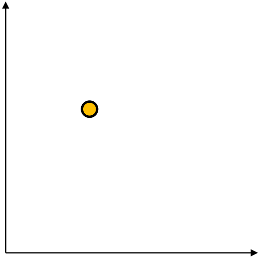


<u>Feature selection problem</u>

Hypothesis selection event

$\left\{ {\boldsymbol{x}_{(\ell)}}^\top \boldsymbol{Y} \geq {\boldsymbol{x}_{(m)}}^\top \boldsymbol{Y} \right\}_{(\ell,m)\in\{1,\ldots,k\}\times\{k+1,\ldots,d\}}$

Data space $\boldsymbol{Y} \in \mathbb{R}^n$



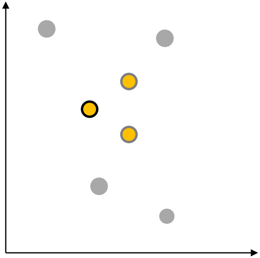
Hypothesis selection event

Segmentation problem

Hypothesis selection event

$X_i \geq \theta \text{ if } x_i \in \mathcal{O}_{\boldsymbol{x}}, X_i < \theta \text{ if } x_i \in \mathcal{B}_{\boldsymbol{x}}$

Data space $\boldsymbol{X} \in \mathbb{R}^n$

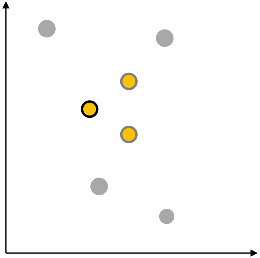


Feature selection problem

Hypothesis selection event

$\left\{ \boldsymbol{x}_{(\ell)}{}^\top \boldsymbol{Y} \geq \boldsymbol{x}_{(m)}{}^\top \boldsymbol{Y} \right\}_{(\ell,m)\in\{1,\dots,k\}\times\{k+1,\dots,d\}}$

Data space $\boldsymbol{Y} \in \mathbb{R}^n$



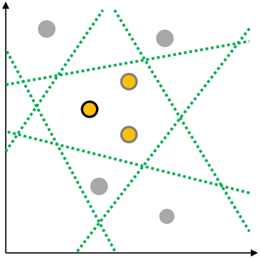
Hypothesis selection event

Segmentation problem

Hypothesis selection event

$$X_i \geq \theta \text{ if } x_i \in \mathcal{O}_{\boldsymbol{x}}, X_i < \theta \text{ if } x_i \in \mathcal{B}_{\boldsymbol{x}}$$

Data space $\boldsymbol{X} \in \mathbb{R}^n$

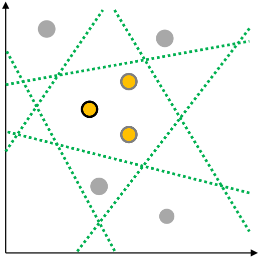


Feature selection problem

Hypothesis selection event

$$\left\{ {\boldsymbol{x}_{(\ell)}}^\top \boldsymbol{Y} \geq {\boldsymbol{x}_{(m)}}^\top \boldsymbol{Y} \right\}_{(\ell, m) \in \{1, \ldots, k\} \times \{k+1, \ldots, d\}}$$

Data space $\boldsymbol{Y} \in \mathbb{R}^n$



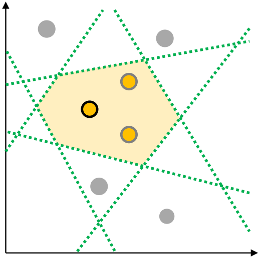
Hypothesis selection event

<u>Segmentation problem</u>

Hypothesis selection event

$$X_i \geq \theta \text{ if } x_i \in \mathcal{O}_{\boldsymbol{x}}, X_i < \theta \text{ if } x_i \in \mathcal{B}_{\boldsymbol{x}}$$

Data space $\boldsymbol{X} \in \mathbb{R}^n$

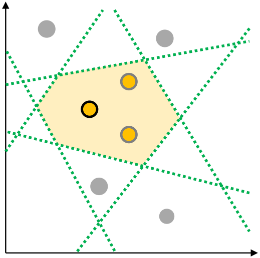


<u>Feature selection problem</u>

Hypothesis selection event

$$\left\{ {\boldsymbol{x}_{(\ell)}}^\top \boldsymbol{Y} \geq {\boldsymbol{x}_{(m)}}^\top \boldsymbol{Y} \right\}_{(\ell,m)\in\{1,\dots,k\}\times\{k+1,\dots,d\}}$$

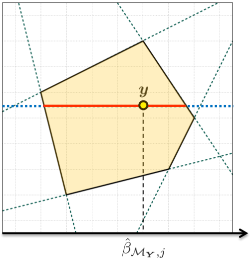
Data space $\boldsymbol{Y} \in \mathbb{R}^n$

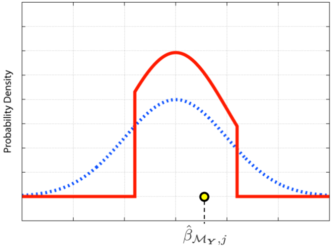


Polyhedral lemma (Lee+16)

- Conditional SI has been actively studied after the seminal work by (Lee+2016)
- Conditional SI for Lasso feature selection was studied in (Lee+2016)
- Brief summary of polyhedral lemma: If
    1. the selection event is represented by a polyhedron (a set of linear inequalities) in the data space, and
    2. the test-statistic is linear function of the data,

  then the exact selective p-values can be computed based on truncated Normal distribution.
- By further conditioning on the sufficient statistic of the nuisance component, we have the sampling distribution on a line in the data space truncated by a polyhedron.





Conditional SI for the selected features by Lasso (Lee+16)

- Consider Lasso is an algorithm to select a set of features and their signs

$$\mathcal{A}^{\text{Lasso}} : \boldsymbol{Y} \mapsto \{\mathcal{M}, \boldsymbol{s}\},$$

  where $\mathcal{M}$ is the set of selected features and $\boldsymbol{s}$ is the set of their signs.

- Test-statistic of the selected features

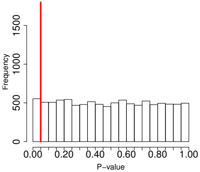
$$\hat{\beta}_{\mathcal{M}_{\boldsymbol{Y}},j} = (X_{\mathcal{M}_{\boldsymbol{Y}}}^{\top} X_{\mathcal{M}_{\boldsymbol{Y}}})^{-1} X_{\mathcal{M}_{\boldsymbol{Y}}}^{\top} \boldsymbol{Y} = \boldsymbol{\eta}^{\top} \boldsymbol{Y}$$

- Selective p-value

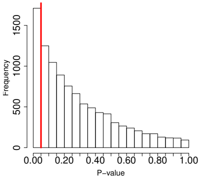
$$p = \Pr\left(|\hat{\beta}_{\mathcal{M}_{\boldsymbol{Y}},j}| \geq |\hat{\beta}_{\mathcal{M}_{\boldsymbol{y}},j}| \;\middle|\; \underbrace{\mathcal{M}_{\boldsymbol{Y}} = \mathcal{M}_{\boldsymbol{y}}}_{\text{features}}, \underbrace{\boldsymbol{s}_{\boldsymbol{Y}} = \boldsymbol{s}_{\boldsymbol{y}}}_{\text{signs}}, \underbrace{P_{\boldsymbol{\eta}}^{\perp} \boldsymbol{Y} = P_{\boldsymbol{\eta}}^{\perp} \boldsymbol{y}}_{\text{nuisance component}} \right)$$

- Selective p-values follow uniform distribution

$$\Pr_{\mathrm{H}_0} (p \leq \alpha \mid \mathcal{M}_{\boldsymbol{Y}} = \mathcal{M}_{\boldsymbol{y}}, \boldsymbol{s}_{\boldsymbol{Y}} = \boldsymbol{s}_{\boldsymbol{y}}) = \alpha \quad \forall \alpha \in (0, 1)$$



Selective p-value distribution



Naive p-value distribution

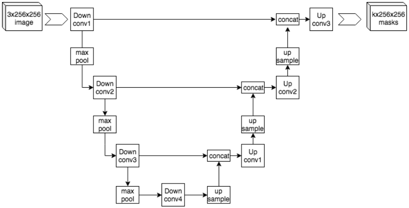
Summary of part 2

- Conditional SI has been actively studied as a promising approach for hypothesis selection bias correction.
- By polyhedral lemma, if the selection event is represented by a polyhedron, the selective p-values can be computed.
- Selection event depends on each algorithm — it is challenging to apply conditional SI to complicated algorithms such as DNN.

Part 3: Conditional SI for DNN-driven Hypotheses

Image segmentation by DNN

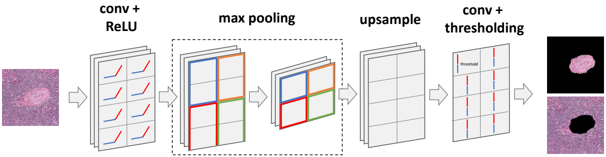
- U-net is one of the most-commonly used CNN for image segmentation task:



Basic structure of U-Net (Wikipedia)

U-net is fully convolutional network and has U-shape.

- Basic components of CNN



- CNN is a complicated function as a whole, but it consists of several basic simple components.

Selection event for DNN

- CNN-based segmentation algorithm

$$\mathcal{A}^{\mathrm{CNN}} : \boldsymbol{X} \mapsto \{\mathcal{O}_{\boldsymbol{X}}, \mathcal{B}_{\boldsymbol{X}}\}$$

- Conditional SI for CNN-based segmentation

$$p = \Pr\left(|\Delta_{\boldsymbol{X}}| \geq |\Delta_{\boldsymbol{x}}| \;\middle|\; \underbrace{\mathcal{A}^{\mathrm{CNN}}(\boldsymbol{X}) = \mathcal{A}^{\mathrm{CNN}}(\boldsymbol{x})}_{\text{the same CNN outputs are selected}}, \underbrace{P_{\boldsymbol{\eta}}^{\perp}\boldsymbol{X} = P_{\boldsymbol{\eta}}^{\perp}\boldsymbol{x}}_{\text{nuisance component}} \right)$$

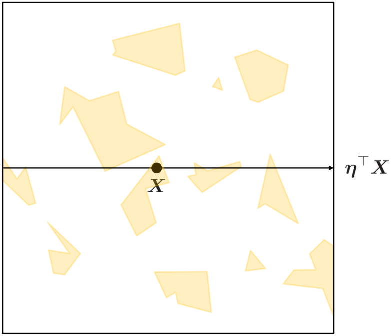
- Q. Can we characterize the complicated selection event of CNN?

$$\{\mathcal{O}_{\boldsymbol{X}}, \mathcal{B}_{\boldsymbol{X}}\} = \{\mathcal{O}_{\boldsymbol{x}}, \mathcal{B}_{\boldsymbol{x}}\} \quad \Leftrightarrow \quad \mathcal{A}^{\mathrm{CNN}}(\boldsymbol{X}) = \mathcal{A}^{\mathrm{CNN}}(\boldsymbol{x})$$

Unfortunately, the selection event cannot be represented as a polyhedron.

Parametric programming approach

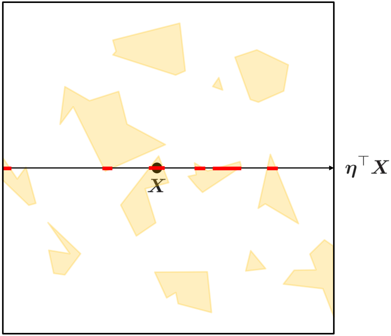
- The selection event of CNN segmentation algorithm is complicated:



- Our idea is to consider solving a sequence of segmentation problems for a parametrized data in the direction of the test-statistic:
  1. Consider multiple finer selection events with additional conditions;
  2. Run the segmentation algorithm for each finer selection event on the line;
  3. Identify the truncation region at which the same result is obtained;
  4. Combine the probability mass of multiply truncated Normal distributions;

Parametric programming approach

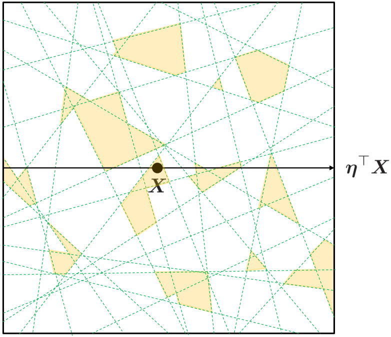
- The selection event of CNN segmentation algorithm is complicated:



- Our idea is to consider solving a sequence of segmentation problems for a parametrized data in the direction of the test-statistic:
  1. Consider multiple finer selection events with additional conditions;
  2. Run the segmentation algorithm for each finer selection event on the line;
  3. Identify the truncation region at which the same result is obtained;
  4. Combine the probability mass of multiply truncated Normal distributions;

Parametric programming approach

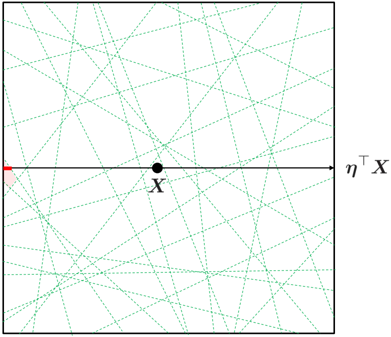
- The selection event of CNN segmentation algorithm is complicated:



- Our idea is to consider solving a sequence of segmentation problems for a parametrized data in the direction of the test-statistic:
  1. Consider multiple finer selection events with additional conditions;
  2. Run the segmentation algorithm for each finer selection event on the line;
  3. Identify the truncation region at which the same result is obtained;
  4. Combine the probability mass of multiply truncated Normal distributions;

Parametric programming approach

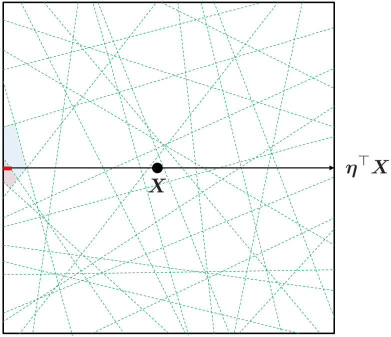
- The selection event of CNN segmentation algorithm is complicated:



- Our idea is to consider solving a sequence of segmentation problems for a parametrized data in the direction of the test-statistic:
  1. Consider multiple finer selection events with additional conditions;
  2. Run the segmentation algorithm for each finer selection event on the line;
  3. Identify the truncation region at which the same result is obtained;
  4. Combine the probability mass of multiply truncated Normal distributions;

Parametric programming approach

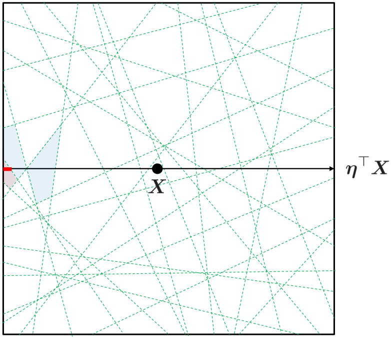
- The selection event of CNN segmentation algorithm is complicated:



- Our idea is to consider solving a sequence of segmentation problems for a parametrized data in the direction of the test-statistic:
  1. Consider multiple finer selection events with additional conditions;
  2. Run the segmentation algorithm for each finer selection event on the line;
  3. Identify the truncation region at which the same result is obtained;
  4. Combine the probability mass of multiply truncated Normal distributions;

Parametric programming approach

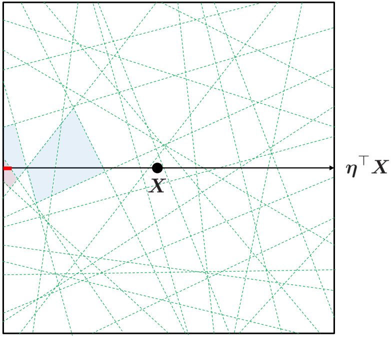
- The selection event of CNN segmentation algorithm is complicated:



- Our idea is to consider solving a sequence of segmentation problems for a parametrized data in the direction of the test-statistic:
  1. Consider multiple finer selection events with additional conditions;
  2. Run the segmentation algorithm for each finer selection event on the line;
  3. Identify the truncation region at which the same result is obtained;
  4. Combine the probability mass of multiply truncated Normal distributions;

Parametric programming approach

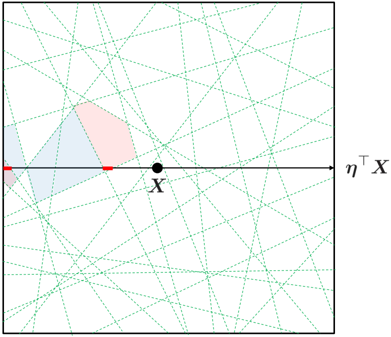
- The selection event of CNN segmentation algorithm is complicated:



- Our idea is to consider solving a sequence of segmentation problems for a parametrized data in the direction of the test-statistic:
  1. Consider multiple finer selection events with additional conditions;
  2. Run the segmentation algorithm for each finer selection event on the line;
  3. Identify the truncation region at which the same result is obtained;
  4. Combine the probability mass of multiply truncated Normal distributions;

Parametric programming approach

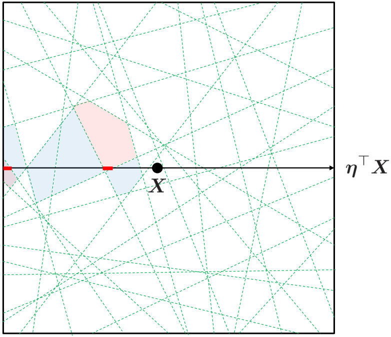
- The selection event of CNN segmentation algorithm is complicated:



- Our idea is to consider solving a sequence of segmentation problems for a parametrized data in the direction of the test-statistic:
  1. Consider multiple finer selection events with additional conditions;
  2. Run the segmentation algorithm for each finer selection event on the line;
  3. Identify the truncation region at which the same result is obtained;
  4. Combine the probability mass of multiply truncated Normal distributions;

Parametric programming approach

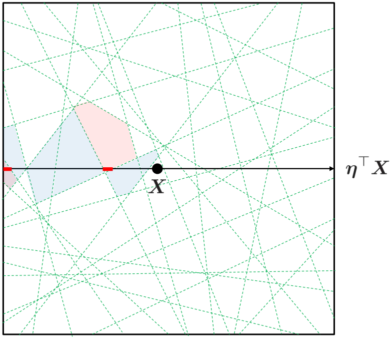
- The selection event of CNN segmentation algorithm is complicated:



- Our idea is to consider solving a sequence of segmentation problems for a parametrized data in the direction of the test-statistic:
  1. Consider multiple finer selection events with additional conditions;
  2. Run the segmentation algorithm for each finer selection event on the line;
  3. Identify the truncation region at which the same result is obtained;
  4. Combine the probability mass of multiply truncated Normal distributions;

Parametric programming approach

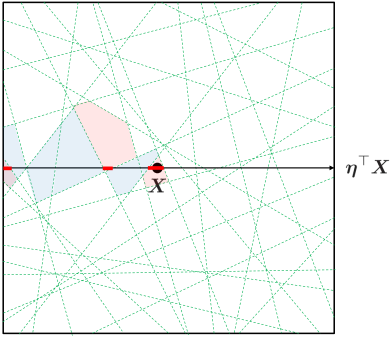
- The selection event of CNN segmentation algorithm is complicated:



- Our idea is to consider solving a sequence of segmentation problems for a parametrized data in the direction of the test-statistic:
  1. Consider multiple finer selection events with additional conditions;
  2. Run the segmentation algorithm for each finer selection event on the line;
  3. Identify the truncation region at which the same result is obtained;
  4. Combine the probability mass of multiply truncated Normal distributions;

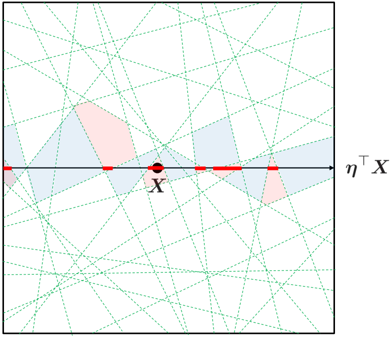
Parametric programming approach

- The selection event of CNN segmentation algorithm is complicated:



- Our idea is to consider solving a sequence of segmentation problems for a parametrized data in the direction of the test-statistic:
  1. Consider multiple finer selection events with additional conditions;
  2. Run the segmentation algorithm for each finer selection event on the line;
  3. Identify the truncation region at which the same result is obtained;
  4. Combine the probability mass of multiply truncated Normal distributions;

- The selection event of CNN segmentation algorithm is complicated:

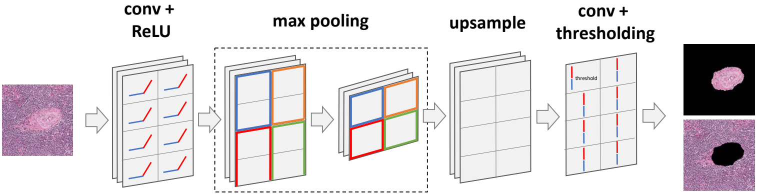


- Our idea is to consider solving a sequence of segmentation problems for a parametrized data in the direction of the test-statistic:
  1. Consider multiple finer selection events with additional conditions;
  2. Run the segmentation algorithm for each finer selection event on the line;
  3. Identify the truncation region at which the same result is obtained;
  4. Combine the probability mass of multiply truncated Normal distributions;

Additional conditioning by finer events regarding basic components of CNN

- By additionally conditioning on finer selection events regarding the basic components of CNN, the selection event is characterized as a union of polhedra.

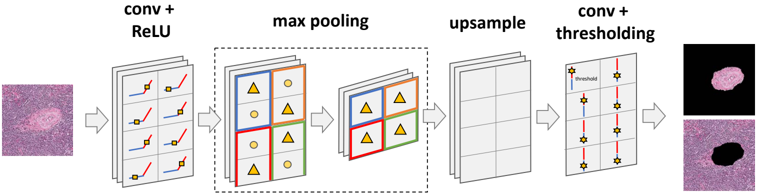
Component	Operations
Convolution	linear
ReLU transfer function	piecewise-linear
Max-pooling	comparison
Upsampling	linear
Thresholding	comparison



Additional conditioning by finer events regarding basic components of CNN

- By additionally conditioning on finer selection events regarding the basic components of CNN, the selection event is characterized as a union of polhedra.

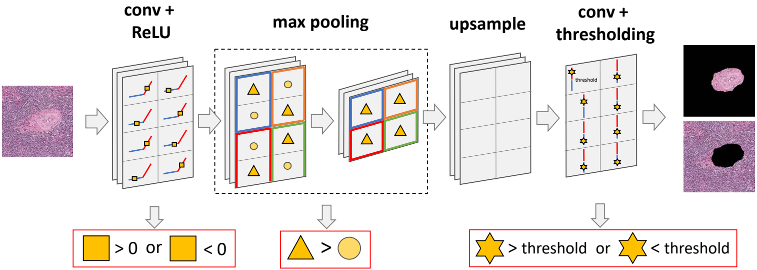
Component	Operations
Convolution	linear
ReLU transfer function	piecewise-linear
Max-pooling	comparison
Upsampling	linear
Thresholding	comparison



Additional conditioning by finer events regarding basic components of CNN

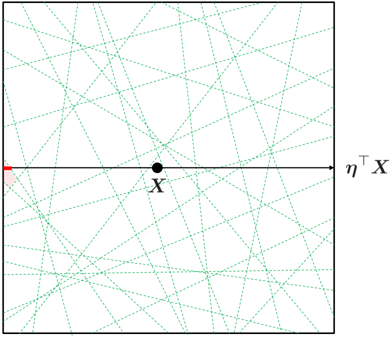
- By additionally conditioning on finer selection events regarding the basic components of CNN, the selection event is characterized as a union of polhedra.

Component	Operations
Convolution	linear
ReLU transfer function	piecewise-linear
Max-pooling	comparison
Upsampling	linear
Thresholding	comparison



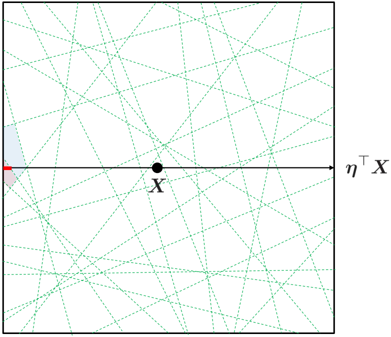
A summary of the divide-and-conquer approach

- Parametric programming approach looks like:



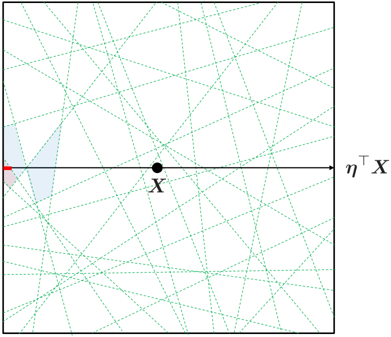
A summary of the divide-and-conquer approach

- Parametric programming approach looks like:



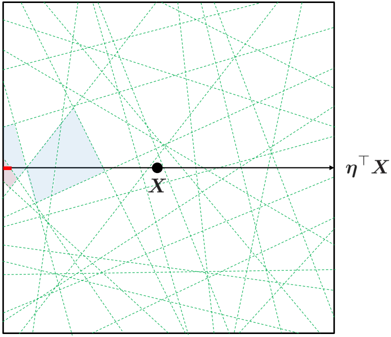
A summary of the divide-and-conquer approach

- Parametric programming approach looks like:



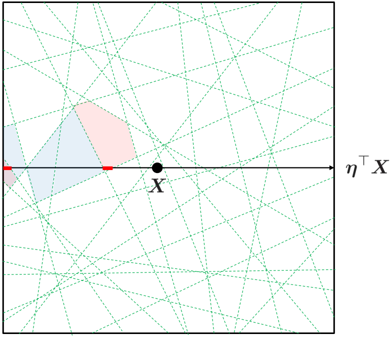
A summary of the divide-and-conquer approach

- Parametric programming approach looks like:



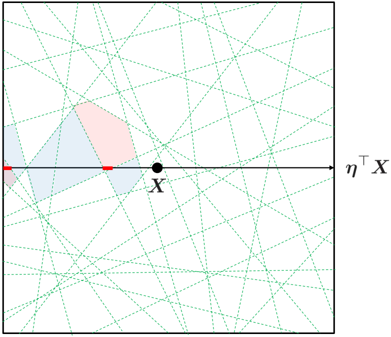
A summary of the divide-and-conquer approach

- Parametric programming approach looks like:



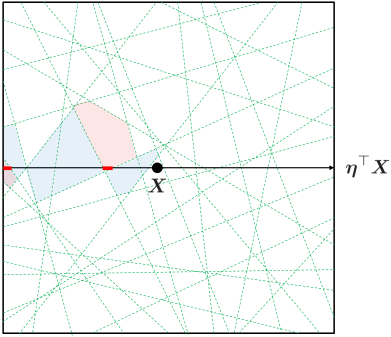
A summary of the divide-and-conquer approach

- Parametric programming approach looks like:



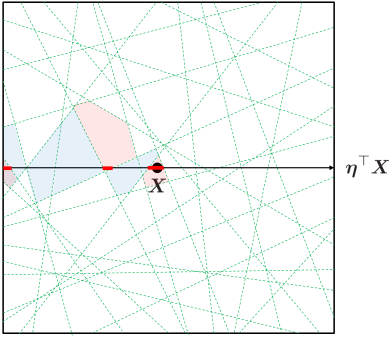
A summary of the divide-and-conquer approach

- Parametric programming approach looks like:



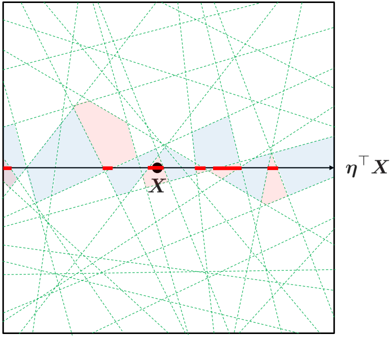
A summary of the divide-and-conquer approach

- Parametric programming approach looks like:



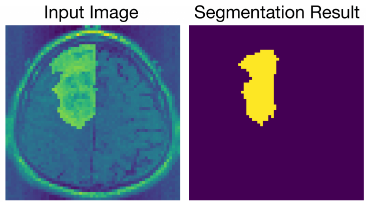
A summary of the divide-and-conquer approach

- Parametric programming approach looks like:

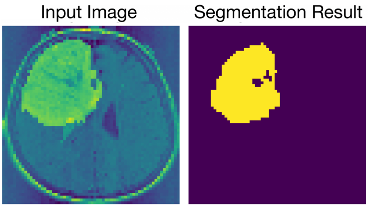


Examples of brain tumor image segmentation

- Positive cases (with true brain tumors)

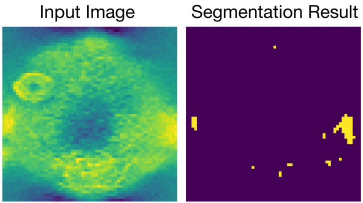


naive p: $\underbrace{0.000}_{\text{true pos.}}$ selective p: $\underbrace{0.000}_{\text{true pos.}}$

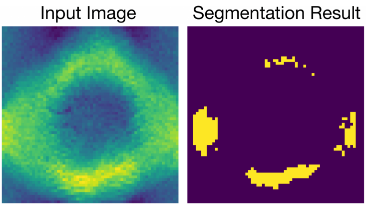


naive p: $\underbrace{0.000}_{\text{true pos.}}$ selective p: $\underbrace{0.000}_{\text{true pos.}}$

- Negative cases (without true brain tumors)



naive p: $\underbrace{0.000}_{\text{false pos.}}$ selective p: $\underbrace{0.670}_{\text{true neg.}}$



naive p: $\underbrace{0.000}_{\text{false pos.}}$ selective p: $\underbrace{0.451}_{\text{true neg.}}$

Summary

- Knowledge obtained by AI is a data-driven hypothesis, which cannot be properly evaluated by traditional statistical inference.
- Conditional SI is a promising approach for exact inference for data-driven hypotheses.
- The main technical challenge of conditional SI is how to characterize the selection event.
- Polyhedral lemma enables us to handle selection event represented as a polyhedron.
- Our parametric programming approach can be used for more complicate selection event such as DNN-driven hypotheses.
- We applied this parametric programming approach to several other problems such change point detection, outlier detection, clustering etc.

References

- J. Lee et al. Exact post-selection inference, with application to the lasso. Annals of Statistics 2016.
- R. Tibshirani et al. Exact post-selection inference for sequential regression procedures. Journal of American Statistical Association 2016.
- W. Fithian et al. Optimal inference after model selection. arXiv 2014.
- J. Taylor et al. Statistical learning and selective inference. PNAS 2015.
- X. Tian, J. Taylor. Asymptotics of selective inference. arXiv 2015.
- S. Suzumura et al. Selective inference for sparse higher-order interaction models. ICML2017.
- K. Tanizaki et al. Computing valid p-values for image segmentation by selective inference. CVPR2020.
- VNL. Duy et al. Computing valid p-value for optimal changepoint by selective inference using dynamic programming. NeurIPS2020.
- VNL. Duy et al. Parametric programming approach for more powerful and general Lasso selective inference. AISTATS2021.
- VNL. Duy et al. Quantifying Statistical Significance of Neural Network-based Image Segmentation by Selective Inference. arXiv 2020.
- K. Sugiayam et al, More powerful and general selective inference for stepwise feature selection using homotopy method. ICML2021.
- D. Das et al. Fast and more powerful selective inference for sparse high-order interaction model. AAAI2022.