# Computing Valid $p$-values for Image Segmentation by Selective Inference

Kosuke Tanizaki[1], Noriaki Hashimoto[1], Yu Inatsu[2], Hidekata Hontani[1], Ichiro Takeuchi[1,2*]
[1]Nagoya Institute of Technology, [2]RIKEN

## Abstract

*Image segmentation is one of the most fundamental tasks in computer vision. In many practical applications, it is essential to properly evaluate the reliability of individual segmentation results. In this study, we propose a novel framework for quantifying the statistical significance of individual segmentation results in the form of $p$-values by statistically testing the difference between the object region and the background region. This seemingly simple problem is actually quite challenging because the difference — called segmentation bias — can be deceptively large due to the adaptation of the segmentation algorithm to the data. To overcome this difficulty, we introduce a statistical approach called selective inference, and develop a framework for computing valid $p$-values in which segmentation bias is properly accounted for. Although the proposed framework is potentially applicable to various segmentation algorithms, we focus in this paper on graph-cut- and threshold-based segmentation algorithms, and develop two specific methods for computing valid $p$-values for the segmentation results obtained by these algorithms. We prove the theoretical validity of these two methods and demonstrate their practicality by applying them to the segmentation of medical images.*

## 1. Introduction

Image segmentation is one of the most fundamental tasks in computer vision. Segmentation algorithms are usually formulated as a problem of optimizing a certain loss function. For example, in threshold (TH)-based segmentation algorithms [24, 38], the loss functions are defined based on similarity within a given region and dissimilarity between different regions. In graph cut (GC)-based segmentation algorithms [5, 2], the loss functions are defined based on the similarity of adjacent pixels in a given region and dissimilarity of adjacent pixels at the boundaries. Depending on the problem and the properties of the target images, an appropriate segmentation algorithm must be selected.

*Correspondence to I.T. (e-mail: takeuchi.ichiro@nitech.ac.jp).

In many practical non-trivial applications, there may be the risk of obtaining incorrect segmentation results. In practical problems in which segmentation results are used for high-stake decision-making or as a component of a larger system, it is essential to properly evaluate their reliability. For example, when segmentation results are used in a computer-aided diagnosis system, it should be possible to appropriately quantify the risk that the obtained individual segments are false positive findings. If multiple segmentation results with ground truth annotations are available, it is possible to quantify the expected proportion of the overall false positive findings (e.g., by receiver operating characteristic (ROC) curve analysis). On the other hand, it is challenging to quantitatively evaluate the reliability of individual segmentation results when the ground-truth segmentation result is not available,

In this study, we propose a novel framework called Post-Segmentation Inference (PSegI) for quantifying the statistical significance of individual segmentation results in the form of $p$-values when only null images (images in which we know that there exist no objects) are available. For simplicity, we focus only on segmentation problems in which an image is divided into an object region and a background region. To quantify the reliability of individual segmentation results, we utilize a statistical hypothesis test for determining the difference between the two regions (see (2) in §2). If the difference is sufficiently large and the probability of observing such a large difference is sufficiently small in a null image (i.e., one that contains no specific objects), it indicates that the segmentation result is statistically significant. The $p$-value of the statistical hypothesis test can be used as a quantitative reliability metric of individual segmentation results; i.e., if the $p$-value is sufficiently small, it implies that a segmentation result is reliable.

Although this problem seems fairly simple, computing a valid $p$-value for the above statistical hypothesis test is challenging because the difference between the object and background regions can be deceptively large even in a null image that contains no specific objects since the segmentation algorithm divides the image into two regions so that they are as different as possible. We refer to this deceptive difference in segmentation results as segmentation bias.

It can be interpreted that segmentation bias arises because the image data are used twice: once for dividing the object and background regions with a segmentation algorithm, and again for testing the difference in the average intensities between the two regions. Such data analysis is often referred to as double-dipping [19] in statistics, and it has been recognized that naively computed $p$-values in double-dipping data analyses are highly biased. Figure 1 illustrates segmentation bias in a simple simulation. In the proposed PSegI framework, we overcome this difficulty by introducing a recently developed statistical approach called selective inference (SI). Selective Inference has been mainly studied for the statistical analysis of linear model coefficients after feature selection, which can be interpreted as an example of double-dipping [12, 31, 33, 22, 20, 36, 29, 34, 32, 10] *.

To the best of our knowledge, due to the difficulty in accounting for segmentation bias, statistical testing approaches have never been successfully used for evaluating the reliability of segmentation results. Our paper has three main contributions. First, we propose the PSegI framework in which the problem of quantifying the reliability of individual segmentation results is formulated as an SI problem, making the framework potentially applicable to a wide range of existing segmentation algorithms. Second, we specifically study the GC-based segmentation algorithm [5, 2] and the TH-based segmentation algorithm [24, 38] as examples, and develop two specific PSegI methods, called PSegI-GC and PSegI-TH, for computing valid $p$-values for the segmentation results obtained with these two respective segmentation algorithms. Finally, we apply the PSegI-GC and PSegI-TH methods to medical images to demonstrate their efficacy.

**Related work.** A variety of image segmentation algorithms with different losses have been developed by incorporating various properties of the target images [21, 11, 41]. The performance of a segmentation algorithm is usually measured based on a human-annotated ground-truth dataset. One commonly used evaluation criterion for segmentation algorithms is the area under the curve (AUC). Unfortunately, criteria such as AUC cannot be used to quantify the reliability of individual segmentation results. The segmentation problem can also be viewed as a two-class classification problem that classifies pixels into object and background classes. Many two-class classification algorithms can provide some level of confidence that a given pixel belongs to the object or the background, e.g., by estimating the posterior probabilities [18, 17, 25, 15]. Although confidence measures can be used to assess the relative reliability of a given pixel, they do not quantify the statistical significance of the segmentation result. In emphcontrario approach, similar discussion to this study has taken place regarding the reliability of objects detected from a noisy image [8, 7, 26, 37]. Unfortunately, however, the contrario approach does not properly account for segmentation bias, and the reliability measure discussed in [26] cannot be used as a $p$-value.

*The main idea of the SI approach was first developed in [20] for computing the $p$-values of the coefficients of LASSO [35]. This problem can be interpreted as an instance of double-dipping data analysis since the training set is used twice: once for selecting features with $L_1$ penalized fitting, and again for testing the statistical significances of the coefficients of the selected features.

**Notation.** We use the following notation in the rest of the paper. For a scalar $s$, $\mathrm{sgn}(s)$ is the sign of $s$, i.e., $\mathrm{sgn}(s) = 1$ if $s \geq 0$ and $-1$ otherwise. For a condition $c$, $\mathbf{1}\{c\}$ is the indicator function, which returns 1 if $c$ is true and 0 otherwise. For natural number $j < n$, $\boldsymbol{e}_j$ is a vector of length $n$ whose $j^{\mathrm{th}}$ element is 1 and whose other elements are 0. Similarly, for a set $\mathcal{S} \subseteq \{1, \ldots, n\}$, $\boldsymbol{e}_{\mathcal{S}}$ is an $n$-dimensional vector whose elements in $\mathcal{S}$ are 1 and 0 otherwise. For a natural number $n$, $I_n$ indicates the $n$-by-$n$ identity matrix.

## 2. Problem Setup

Consider an image with $n$ pixels. We denote the preprocessed pixel values after appropriate filtering operations as $x_1, \ldots, x_n \in \mathbb{R}$, i.e., the $n$-dimensional vector $\boldsymbol{x} := [x_1, \ldots, x_n]^\top \in \mathbb{R}^n$ represents the preprocessed image. For simplicity, we only study segmentation problems in which an image is divided into two regions†. We call these two regions the object region and the background region for clarity. After a segmentation algorithm is applied, $n$ pixels are classified into one of the two regions. We denote the set of pixels classified into the object and the background regions as $\mathcal{O}$ and $\mathcal{B}$, respectively. With this notation, a segmentation algorithm $\mathcal{A}$ is considered to be a function that maps an image $\boldsymbol{x}$ into the two sets of pixels $\mathcal{O}$ and $\mathcal{B}$, i.e., $\{\mathcal{O}, \mathcal{B}\} = \mathcal{A}(\boldsymbol{x})$.

### 2.1. Testing Individual Segmentation Results

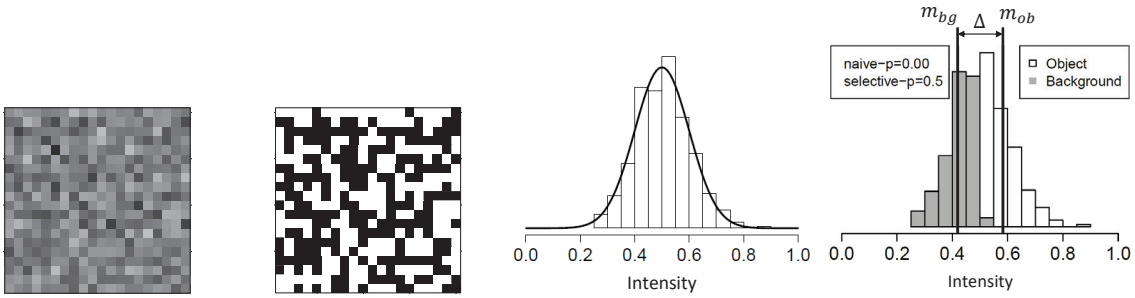
To quantify the reliability of individual segmentation results, consider a score $\Delta$ that represents how much the object and the background regions differ. The PSegI framework can be applied to any scores if it is written in the form of $\Delta = \boldsymbol{\eta}^\top \boldsymbol{x}$ where $\boldsymbol{\eta} \in \mathbb{R}^n$ is any $n$-dimensional vector. For example, if we denote the average pixel values of the object and the background regions as

$$m_{\mathrm{ob}} = \frac{1}{|\mathcal{O}|} \sum_{p \in \mathcal{O}} x_p, \; m_{\mathrm{bg}} = \frac{1}{|\mathcal{B}|} \sum_{p \in \mathcal{B}} x_p,$$

and define $\boldsymbol{\eta}$ as

$$\eta_i = \begin{cases} \mathrm{sgn}(m_{\mathrm{ob}} - m_{\mathrm{bg}})/|\mathcal{O}|, & \text{if } i \in \mathcal{O}, \\ \mathrm{sgn}(m_{\mathrm{bg}} - m_{\mathrm{ob}})/|\mathcal{B}|, & \text{if } i \in \mathcal{B}, \end{cases}$$

†The proposed PSegI framework can be easily extended to cases where an image is divided into more than two regions.




(a) Original image (b) Segmentation result (c) Pixel distribution (d) Ob. and Bg. histograms

Figure 1: Schematic illustration of segmentation bias that arises when the statistical significance of the difference between the object and background regions obtained with a segmentation algorithm is tested. (a) Randomly generated image with $n = 400$ pixels from $N(0.5, 0.1^2)$. (b) Segmentation result obtained with the local threshold-based segmentation algorithm in [38]. (c) Distribution and histogram of pixel intensities. (d) Histograms of pixel intensities in the object region (white) and the background region (gray). Note that even for an image that contains no specific objects, the pixel intensities of the object and background regions are clearly different. Thus, if we naively compute the statistical significance of the difference, the $p$-value (naive-$p$ in (d)) would be very small, indicating that it cannot be used for properly evaluating the reliability of the segmentation result. In this paper, we present a novel framework for computing valid $p$-values (selective-$p$ in (d)), which properly account for segmentation bias.

then the score $\Delta$ represents the absolute average difference in the pixel values between object and background region

$$\Delta = |m_{\mathrm{ob}} - m_{\mathrm{bg}}|. \tag{1}$$

In what follows, for simplicity, we assume that the score $\Delta$ is in the form of (1), but any other scores in the form of $\boldsymbol{\eta}^\top \boldsymbol{x}$ can be employed. Other specific examples are discussed in supplement A.

If the difference $\Delta$ is sufficiently large, it implies that the segmentation result is reliable. As discussed in §1, it is non-trivial to properly evaluate the statistical significance of the difference $\Delta$ since it can be deceptively large due to the effect of segmentation bias. To quantify the statistical significance of the difference $\Delta$, we consider a statistical hypothesis test with the following null hypothesis $\mathrm{H}_0$ and alternative hypothesis $\mathrm{H}_1$:

$$\mathrm{H}_0 : \mu_{\mathrm{ob}} = \mu_{\mathrm{bg}} \quad \text{vs.} \quad \mathrm{H}_1 : \mu_{\mathrm{ob}} \neq \mu_{\mathrm{bg}}, \tag{2}$$

where $\mu_{\mathrm{ob}}$ and $\mu_{\mathrm{bg}}$ are the true means of the pixel intensities in the object and background regions, respectively. Under the null hypothesis $\mathrm{H}_0$, we assume that an image consists only of background information, and that the statistical variability of the background information can be represented by an $n$-dimensional normal distribution $N(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu} \in \mathbb{R}^n$ is the unknown mean vector and $\Sigma \in \mathbb{R}^{n \times n}$ is the covariance matrix, which is known or estimated from independent data. In practice, we estimate the covariance matrix $\Sigma$ from a null image in which we know that there exists no object. For example, in medical image analysis, it is not uncommon to assume the availability of such null images.

In a standard statistical test, the $p$-value is computed based on the null distribution $\mathbb{P}_{\mathrm{H}_0}(\Delta)$, i.e., the sampling distribution of the test statistic $\Delta$ under the null hypothesis. On the one hand, if we naively compute the $p$-values from the pixel intensities in $\mathcal{O}$ and $\mathcal{B}$ without considering that $\{\mathcal{O}, \mathcal{B}\}$ was obtained with a segmentation algorithm, these naive $p$-values will be highly underestimated due to segmentation bias, and hence the probability of finding incorrect segmentation results cannot be properly controlled.

**Selective inference.** SI is a type of conditional inference, in which a statistical test is conducted based on a conditional sampling distribution of the test statistic under the null hypothesis. In our problem, to account for segmentation bias, we consider testing the difference $\Delta$ conditional on the segmentation results. Specifically, the *selective p-value* is defined as

$$\mathfrak{p} = \mathbb{P}_{\mathrm{H}_0}\left(\Delta > \Delta^{\mathrm{obs}} \;\middle|\; \begin{array}{c} \mathcal{A}(\boldsymbol{x}) = \mathcal{A}(\boldsymbol{x}^{\mathrm{obs}}) \\ \boldsymbol{z}(\boldsymbol{x}) = \boldsymbol{z}(\boldsymbol{x}^{\mathrm{obs}}) \end{array}\right), \tag{3}$$

where the superscript $^{\mathrm{obs}}$ indicates the observed quantity of the corresponding random variable. In (3), the first condition $\mathcal{A}(\boldsymbol{x}) = \mathcal{A}(\boldsymbol{x}^{\mathrm{obs}})$ indicates that we only consider the case where the segmentation result $(\mathcal{O}, \mathcal{B})$ is the same as what we observed from the actual image data. The second condition $\boldsymbol{z}(\boldsymbol{x}) = \boldsymbol{z}(\boldsymbol{x}^{\mathrm{obs}})$ indicates that the component that is (unconditionally) independent of the test statistic $\Delta$ for a random variable $\boldsymbol{x}$ is the same as the observed one [‡].

[‡]In the unconditional case, the condition $\boldsymbol{z}(\boldsymbol{x}) = \boldsymbol{z}(\boldsymbol{x}^{\mathrm{obs}})$ does not change the sampling distribution since $\boldsymbol{\eta}^\top \boldsymbol{x}$ and $\boldsymbol{z}(\boldsymbol{x})$ are (marginally)

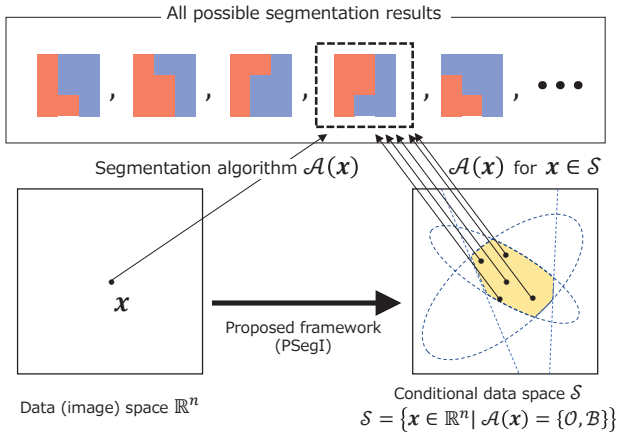



Figure 2: Schematic illustration of the basic idea used in the proposed PSegI framework. By applying a segmentation algorithm $\mathcal{A}$ to an image $\boldsymbol{x}$ in the data space $\mathbb{R}^n$, a segmentation result $\{\mathcal{O}, \mathcal{B}\}$ is obtained. In the PSegI framework, the statistical inference is conducted conditional on the subspace $\mathcal{S} = \{\boldsymbol{x} \in \mathbb{R}^n \mid \mathcal{A}(\boldsymbol{x}) = \{\mathcal{O}, \mathcal{B}\}\}$; i.e., the subspace is selected such that an image taken from the subspace has the same segmentation result $\{\mathcal{O}, \mathcal{B}\}$.

The component $\boldsymbol{z}(\boldsymbol{x})$ is written as

$$\boldsymbol{z}(\boldsymbol{x}) = (I_n - \boldsymbol{c}\boldsymbol{\eta}^\top)\boldsymbol{x} \text{ with } \boldsymbol{c} = \Sigma\boldsymbol{\eta}(\boldsymbol{\eta}^\top\Sigma\boldsymbol{\eta})^{-1}.$$

Figure 2 shows a schematic illustration of the basic idea used in the proposed PSegI framework.

## 2.2. Graph-cut-based Segmentation

As one of the two examples of segmentation algorithm $\mathcal{A}$, we consider the GC-based segmentation algorithm in [5, 2]. In GC-based segmentation algorithms, the target image is considered to be a directed graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ and $\mathcal{E}$ are the sets of nodes and edges, respectively. Let $\mathcal{P}$ be the set of all $n$ pixels, and $\mathcal{N}$ be the set of all directed edges from each pixel to its eight adjacent nodes (each pixel is connected to its horizontally, vertically, and diagonally adjacent pixels). Furthermore, consider two terminal nodes $S$ and $T$. Then, $\mathcal{V}$ and $\mathcal{E}$ of the graph $\mathcal{G}$ are defined as $\mathcal{V} = \mathcal{P} \cup \{S, T\}$ and $\mathcal{E} = \mathcal{N} \cup \bigcup_{p \in \mathcal{P}} \{(S, p), (p, T)\}$, where for the two nodes $p$ and $q$, $(p, q)$ indicates the directed edge from $p$ to $q$. At each edge of the graph $(p, q) \in \mathcal{E}$, non-negative weights $w_{(p,q)}$ are defined based on the pixel intensities (see §3.2).

In GC-based segmentation algorithms, the segmentation into object and background regions is conducted by cutting the graph into two parts. Let us write the ordered partition of the graph as $(\mathcal{V}_s, \mathcal{V}_t)$, where $\mathcal{V}_s$ and $\mathcal{V}_t$ constitute a partition of $\mathcal{V}$. If $S \in \mathcal{V}_s$ and $T \in \mathcal{V}_t$, the ordered partition $(\mathcal{V}_s, \mathcal{V}_t)$ is called an $s$-$t$ cut. Let $\mathcal{E}_{\text{cut}} \subset \mathcal{E}$ be the set of edges $(p, q) \in \mathcal{E}$ such that $p$ belongs to $\mathcal{V}_s$ and $q$ belongs to $\mathcal{V}_t$. The cost function of an $s$-$t$ cut $(\mathcal{V}_s, \mathcal{V}_t)$ is defined as $L_{\text{cut}}(\mathcal{V}_s, \mathcal{V}_t) = \sum_{(p,q) \in \mathcal{E}_{\text{cut}}} w_{(p,q)}$. The GC-based segmentation algorithm is formulated as the optimization problem for finding the optimal $s$-$t$ cut:

$$(\mathcal{V}_s^*, \mathcal{V}_t^*) = \arg \min_{(\mathcal{V}_s, \mathcal{V}_t)} L_{\text{cut}}(\mathcal{V}_s, \mathcal{V}_t). \tag{4}$$

Then, the segmentation result $\{\mathcal{O}, \mathcal{B}\}$ is obtained as $\mathcal{O} \leftarrow \mathcal{V}_s^* \setminus \{S\}$ and $\mathcal{B} \leftarrow \mathcal{V}_t^* \setminus \{T\}$. The minimum cut problem (4) is known to be a dual problem of the maximum flow problem, for which polynomial time algorithms exist [13, 14, 9]. Among the several implementations of the maximum flow problem, we employed the one presented in [4], in which three stages, called the growth stage, the augmentation stage, and the adoption stage, are iterated. Briefly, a path from $S$ to $T$ is obtained in the growth stage. The edge with the minimum weight in the path is selected and all the weights of the path are reduced by the minimum weight to account for the flow in the augmentation stage. The data structure of the updated graph is reconstructed in the adoption stage (see [4] for details).

## 2.3. Threshold-based Segmentation

Next, we briefly describe TH-based segmentation algorithms [27]. In TH-based segmentation algorithms, pixels are simply classified into either the object or background class depending on whether their intensity is greater or smaller than a certain threshold. According to the application and the features of the target images, various approaches for determining the threshold have been proposed. In the following, we first describe a global TH algorithm in which a single threshold is used for the entire image, and then present a local TH algorithm in which different thresholds are used for different pixels.

In the method proposed by Otsu [24], a global threshold is selected to maximize the dispersion between the object and background pixels. Here, dispersion is defined so that the between-region variance is maximized and the within-region variance is minimized. Since the sum of these two variances is the total variance and does not depend on the threshold, we can only maximize the former. Denoting the number, mean, and variance of the pixels with values greater (resp. smaller) than the threshold $t$ as $\overline{n}_t$, $\overline{\mu}_t$, and $\overline{\sigma}_t^2$ (resp. $\underline{n}_t$, $\underline{\mu}_t$, and $\underline{\sigma}_t^2$), respectively, the between-region variance with the threshold $t$ is written as $\sigma_{\text{bet}}^2(t) = \frac{\overline{n}(t)\underline{n}(t)(\overline{\mu}(t) - \underline{\mu}(t))^2}{n^2}$. The global threshold is then determined as $t^* = \arg\max_t \sigma_{\text{bet}}^2(t)$. Although this algorithm is simple, it is used in many practical applications.

The local thresholding approach allows more flexible

independent. On the other hand, under the condition with $\mathcal{A}(\boldsymbol{x}) = \mathcal{A}(\boldsymbol{x}^{\text{obs}})$, $\boldsymbol{\eta}^\top\boldsymbol{x}$ and $\boldsymbol{z}(\boldsymbol{X})$ are not conditionally independent. See [12, 20] for the details.

segmentation since the threshold is determined per pixel. The method proposed by White and Rohrer [38] determines the pixel-wise threshold by comparing the pixel intensity with the average pixel intensity of its neighbors. Here, neighbors are defined by a square window around a pixel, and the local threshold for the pixel is determined as $t_p^* = |\mathcal{W}_p|^{-1} \sum_{q \in \mathcal{W}_p} x_q / \theta$, where $\mathcal{W}_p$ is the set of pixels within the window around the pixel $p$, and $\theta$ is a scalar value specified based on the properties of the image.

## 3. Post-segmentation Inference

In this section, we consider the problem of how to provide a valid $p$-value for the observed difference in average pixel intensities between the object and background regions $\Delta$ when the two regions are obtained by applying a segmentation algorithm $\mathcal{A}$ to an image $\boldsymbol{x}$. In the proposed PSegI framework, we solve this problem by considering the sampling distribution of $\Delta$ conditional on the event $\mathcal{A}(\boldsymbol{x}) = \{\mathcal{O}, \mathcal{B}\}$ under the null hypothesis that $\mu_{\rm ob} = \mu_{\rm bg}$; i.e., the actual mean pixel intensities of the object and background regions are the same. By conditioning on the segmentation result $\{\mathcal{O}, \mathcal{B}\}$, the effect of segmentation bias is properly corrected.

By definition, a valid $p$-value must be interpreted as an upper bound on the probability that the obtained segmentation result $\{\mathcal{O}, \mathcal{B}\}$ is incorrect[§]. To this end, in our conditional inference, a valid $p$-value $\mathfrak{p}$ must satisfy

$$\mathbb{P}_{\rm H_0}\left(\mathfrak{p} \leq \alpha \,\middle|\, \begin{array}{c} \mathcal{A}(\boldsymbol{x}) = \mathcal{A}(\boldsymbol{x}^{\rm obs}) \\ \boldsymbol{z}(\boldsymbol{x}) = \boldsymbol{z}(\boldsymbol{x}^{\rm obs}) \end{array}\right) = \alpha \;\; \forall\, \alpha \in [0,1]. \quad (5)$$

This property is satisfied if and only if $\mathfrak{p}$ is uniformly distributed in $[0,1]$. Therefore, our problem is cast into the problem of computing a function of the test statistic $\Delta$ that follows $\mathrm{Unif}[0,1]$ when the test statistic follows the conditional sampling distribution in (3).

In §3.1, we first present our main result for the proposed PSegI framework. Here, we show that if the event $\mathcal{A}(\boldsymbol{x}) = \{\mathcal{O}, \mathcal{B}\}$ is characterized by a finite set of quadratic inequalities on $\boldsymbol{x}$, then a valid $p$-value can be exactly computed. In §3.2 and §3.3, as examples of the proposed PSegI framework, we develop concrete methods of, respectively, the PSegI framework for a GC-based segmentation algorithm [5, 2] and a TH-based segmentation algorithm [24, 38]. Our key finding is that the event $\mathcal{A}(\boldsymbol{x}) = \{\mathcal{O}, \mathcal{B}\}$ can be characterized by a finite set of quadratic inequalities on $\boldsymbol{x}$ for these segmentation algorithms and thus valid $p$-values can be computed by using the result in §3.1.

### 3.1. Selective Inference for Segmentation Results

The following theorem is the core of the proposed PSegI framework.

[§]Naive $p$-values do not satisfy this property due to segmentation bias.

**Theorem 1.** *Suppose that an image $\boldsymbol{x}$ of size $n$ is drawn from an $n$-dimensional normal distribution $N(\boldsymbol{\mu}, \Sigma)$ with unknown $\boldsymbol{\mu}$ and known or independently estimable $\Sigma$. If the event $\mathcal{A}(\boldsymbol{x}) = \{\mathcal{O}, \mathcal{B}\}$ is characterized by a finite set of quadratic inequalities on $\boldsymbol{x}$ of the form*

$$\boldsymbol{x}^\top A_j \boldsymbol{x} + \boldsymbol{b}_j^\top \boldsymbol{x} + c_j \leq 0, \;\; j = 1, 2, \ldots, \quad (6)$$

*with certain $A_j \in \mathbb{R}^{n \times n}$, $\boldsymbol{b}_j \in \mathbb{R}^n$, and $c_j \in \mathbb{R}$, $j = 1, 2, \ldots$, then*

$$\mathfrak{p} = 1 - F^{E(\boldsymbol{z}(\boldsymbol{x}^{\rm obs}))}_{0, \boldsymbol{\eta}^\top \Sigma \boldsymbol{\eta}}(|m_{\rm ob} - m_{\rm bg}|) \quad (7)$$

*is a valid $p$-value in the sense that it satisfies (5), where $F^E_{m,s^2}$ is a cumulative distribution function of a truncated normal distribution with mean $m$, variance $s^2$, and truncation intervals $E$. Here, the truncation interval is written as*

$$E(\boldsymbol{z}(\boldsymbol{x})) = \bigcap_j \left\{ \Delta > 0 \,\middle|\, \begin{array}{c} (\boldsymbol{z}(\boldsymbol{x}) + \Delta \boldsymbol{y})^\top A_j (\boldsymbol{z}(\boldsymbol{x}) + \Delta \boldsymbol{y}) \\ + \boldsymbol{b}_j^\top (\boldsymbol{z}(\boldsymbol{x}) + \Delta \boldsymbol{y}) + c_j \leq 0 \end{array} \right\},$$

*where $\boldsymbol{y} = \Sigma \boldsymbol{\eta}^\top \boldsymbol{\eta} / (\boldsymbol{\eta}^\top \Sigma \boldsymbol{\eta})$.*

The proof of Theorem 1 is presented in supplement B. This theorem is an adaptation of Theorem 5.2 in [20] and Theorem 3.1 in [22], in which SI on the selected features of a linear model was studied. Note that the normality assumption in Theorem 1 does NOT mean that $n$ pixel values are normally distributed; it means that the noise (deviation from the unknown true mean value) of each pixel value is normally distributed.

### 3.2. Valid $p$-values for GC-based segmentation

As briefly described in §2-2, GC-based segmentation is conducted by solving the maximum flow optimization problem on the directed graph. Basically, all the operations in this optimization process can be decomposed into additions, subtractions, and comparisons of the weights $w_{(p,q)}$ of the directed graph. This suggests that as long as each weight $w_{(p,q)}$ is written as a quadratic function of the image $\boldsymbol{x}$, the event that the GC-based segmentation algorithm produces the segmentation result $\{\mathcal{O}, \mathcal{B}\}$ can be fully characterized by a finite set of quadratic inequalities in the form of (6). In the following, we explain how to set the weights $w_{(p,q)}$ for each edge $(p,q) \in \mathcal{E}$. For properly defining the weights, it is necessary to introduce seed pixels for the object and background regions. The pixels known or highly plausible to be in the object or background regions are set as the seed pixels, denoted as $\mathcal{O}^{\rm se}, \mathcal{B}^{\rm se} \subset \mathcal{P}$, respectively. The seed pixels may be specified by human experts, or the pixel with the largest or smallest intensity may be specified as the object or background seed pixel, respectively.

When the two pixel nodes $p, q \in \mathcal{P}$ are adjacent to each other, the weight $w_{(p,q)}$ is determined based on the similarity of their pixel intensities and the distance between them.

Pixel similarity is usually defined based on the properties of the target image. To provide flexibility in the choice of the similarity function, we employ a quadratic spline approximation, which allows one to specify the desired similarity function with arbitrary approximation accuracy. For example, Figure 1 in supplement C shows an example of the quadratic spline approximation of commonly used weights $w_{(p,q)} = \exp(-(x_p - x_q)^2/(2\sigma^2))\mathrm{dist}(p,q)^{-1}$, where $\mathrm{dist}(p,q)$ is the distance between the two nodes.

The weight between the terminal node $S$ and the general pixel node $p \in \mathcal{P} \setminus (\mathcal{O}^{\mathrm{se}} \cup \mathcal{B}^{\mathrm{se}})$ is usually determined based on the negative log-likelihood of the pixel in the object region. Under the normality assumption, it is written as $w_{S,p} = -\log \mathbb{P}(x_p \mid p \in \mathcal{O}) \simeq \log(2\pi\sigma^2 + (x_p - m_{\mathrm{ob}}^{\mathrm{se}})^2/(2\sigma^2))$, where $m_{\mathrm{ob}}^{\mathrm{se}} = \sum_{i\in\mathcal{O}^{\mathrm{se}}} x_i/|\mathcal{O}^{\mathrm{se}}|$ is the estimate of the mean pixel intensity in the object region from the object seed pixel intensities. The weight between the terminal node $S$ and an object seed pixel node $p \in \mathcal{O}^{\mathrm{se}}$ should be sufficiently large. It is usually determined as $w_{S,p} = 1 + \max_{q\in\mathcal{P}} \sum_{r:(q,r)\in\mathcal{N}} w_{(q,r)}$. The weight between the terminal node $S$ and a background seed node $p \in \mathcal{B}^{\mathrm{se}}$ is set to zero. The weights between the terminal node $T$ and pixel nodes are determined in the same way.

Since all the weights for the edges are represented by quadratic equations and quadratic constraints on $\boldsymbol{x}$, all the operations for solving the minimum cut (or maximum flow) optimization problem (4) can be fully characterized by a finite set of quadratic inequalities in the form of (6). Thus, valid $p$-values of the segmentation result obtained with the GC-based segmentation algorithm can be computed using the PSegI framework. In supplement C, we present all the matrices $A_j$, vectors $\boldsymbol{b}_j$, and scalars $c_j$ needed for characterizing a GC-based segmentation event.

### 3.3. Valid $p$-values for TH-based segmentation

Both the global and local TH algorithms fit into the PSegI framework. First, consider the global TH algorithm. For simplicity, consider selecting a global threshold $t^*$ from 256 values $t \in \{0, 1, \ldots, 255\}$. An event that the global threshold $t^*$ is selected can be simply written as

$$\sigma^2_{\mathrm{bet}}(t^*) \geq \sigma^2_{\mathrm{bet}}(t), t \in \{0, \ldots, 255\}. \tag{8}$$

Let $\overline{\boldsymbol{u}}(t)$ and $\underline{\boldsymbol{u}}(t)$ be $n$-dimensional vectors whose elements are defined as

$$\overline{u}(t)_p = \begin{cases} 1 & \text{if } x_p \geq t, \\ 0 & \text{otherwise;} \end{cases} \quad \underline{u}(t)_p = \begin{cases} 0 & \text{if } x_p \geq t, \\ 1 & \text{otherwise.} \end{cases}$$

Then, since the between-region variance $\sigma^2_{\mathrm{bet}}(t)$ is written as the quadratic function

$$\boldsymbol{x}^\top \left( \frac{\underline{n}(t)}{\overline{n}(t)} \overline{\boldsymbol{u}}(t)\overline{\boldsymbol{u}}(t)^\top + \frac{\overline{n}(t)}{\underline{n}(t)} \underline{\boldsymbol{u}}(t)\underline{\boldsymbol{u}}(t)^\top - 2\overline{\boldsymbol{u}}(t)\underline{\boldsymbol{u}}(t)^\top \right) \boldsymbol{x},$$

the event in (8) is represented by 255 quadratic inequalities on $\boldsymbol{x}$. Furthermore, it is necessary to specify whether pixels are in the object or background region at each threshold $t \in \{0, \ldots, 255\}$. To this end, consider conditioning on the order of pixel intensities, which is represented by a set of $n-1$ linear inequalities:

$$\boldsymbol{e}_{(i)}^\top \boldsymbol{x} \leq \boldsymbol{e}_{(i+1)}^\top \boldsymbol{x}, i = 1, \ldots, n-1, \tag{9}$$

where $(1), (2), \ldots, (n)$ is the sequence of pixel IDs such that $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Since the conditions (8) and (9) are represented by sets of quadratic and linear inequalities on $\boldsymbol{x}$, valid $p$-values of the segmentation result obtained with the global TH algorithm can be computed using the PSegI framework.

Next, consider the local threshold approach. The conditions under which the $p^{\mathrm{th}}$ pixel is classified into the object or background region are simply written as a set of linear inequalities on $\boldsymbol{x}$ as

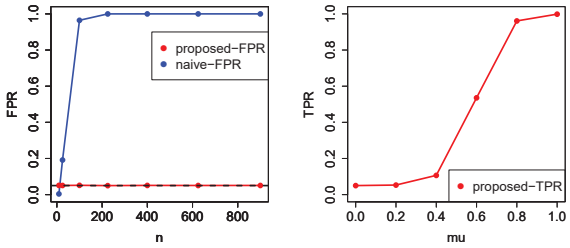
$$x_p \geq (|\mathcal{W}_p|^{-1} \sum_{q\in\mathcal{W}_p} x_q)/\theta \;\Leftrightarrow\; \boldsymbol{e}_p^\top \boldsymbol{x} \geq |\mathcal{W}_p|^{-1} \boldsymbol{e}_{\mathcal{W}_p}^\top \boldsymbol{x},$$
$$x_p \leq (|\mathcal{W}_p|^{-1} \sum_{q\in\mathcal{W}_p} x_q)/\theta \;\Leftrightarrow\; \boldsymbol{e}_p^\top \boldsymbol{x} \leq |\mathcal{W}_p|^{-1} \boldsymbol{e}_{\mathcal{W}_p}^\top \boldsymbol{x},$$

respectively. Thus, valid $p$-values of the segmentation result obtained with the local TH-based algorithm can be computed using the PSegI framework.

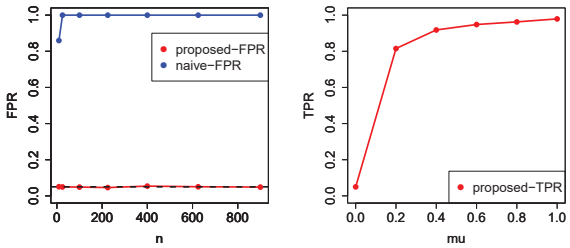
## 4. Experiments

We confirm the validity of the proposed method by numerical experiments. First, we evaluated the false positive rate (FPR) and the true positive rate (TPR) of the proposed method using artificial data. Then, we applied the proposed method to medical images as a practical application. We compared the proposed method with the naive method, which assumes that $\Delta \sim N(0, \tilde{\sigma}^2)$, where $\tilde{\sigma}^2$ is computed based on the segmentation result without considering segmentation bias. We denote the $p$-values obtained using the proposed method and the naive method as selective-$p$ and naive-$p$, respectively.

**Experiments using artificial data.** In the artificial data experiments, Monte Carlo simulation was conducted $10^5$ times. The significance level was set to $\alpha = 0.05$ and the FPRs and TPRs were estimated as $10^{-5} \sum_{i=1}^{10^5} \mathbf{1}\{\mathfrak{p}_i < \alpha\}$, where $\mathfrak{p}_i$ is the $p$-value at the $i^{\mathrm{th}}$ Monte Carlo trial. Data were generated with the range of pixel values $x \in [0, 1]$. The maximum and minimum values were used as the seeds for the object and background regions, respectively. Note that these seed selections were incorporated as selection events. In the experiment for FPR, the data were randomly generated as $\boldsymbol{x} \sim \mathrm{N}(\mathbf{0.5}_n, 0.5I_{n\times n})$ for




(a) FPR for GC (b) TPR for GC




(c) FPR for TH (d) TPR for TH

Figure 3: Results of artificial data experiments using GC- and TH-based segmentation algorithms. (a) and (c) show that the FPRs of the proposed method are properly controlled at the desired significance level. In contrast, the naive method totally failed to control the FPRs. (b) and (d) show that the proposed method successfully identified the correct segmentation results.

$n = 9, 25, 100, 225, 400, 625, 900$. Next, in the experiment for TPR, data were randomly generated as $\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{\mu}, 0.1^2 I_{n\times n})$. Here, $\boldsymbol{\mu}$ is an $n$-dimensional vector that contains $100 \times 100$ elements for which the upper left submatrix with size $50 \times 50$ has a mean value $\mu_S$ and for which the remaining values have mean value $\mu_T$. Cases with $\mu = \mu_S - \mu_T = 0.0, 0.2, \ldots, 1.0$ were investigated. The results are shown in Figure 3. Figures 3a-b and c-d show the results for the GC- and TH-based segmentation algorithms, respectively. As shown in Figures 3a and c, the proposed method controlled the FPRs at the desired significance level, whereas the naive method could not. The FPR of the naive method increased with image size $n$ since the deceptive difference in the mean value between the two regions increased. Figures 3b and d show that the TPR of the proposed method increased as the difference between the two regions $\mu$ increased.

**Experiments using medical images.** In this section, we applied the proposed method and the naive method to pathological images and computed tomography (CT) images. For pathological images, the GC-based segmentation algorithm was employed to extract fibrous tissue regions in pathological tissue specimens. The quantitative analysis of pathological images is useful for computer-aided diagnosis, and the extraction of specified areas is practically important [6, 30, 39]. The pathological images were obtained by scanning tissue specimens of the spleen and cervical lymph node stained with hematoxylin and eosin at Nagoya University Hospital. As a scanning equipment Aperio ScanScope XT (Leica Biosystems, Germany) was utilized and the glass slides were scanned at 20x magnification. From the scanned whole-slide images, several region-of-interest (ROI) images were manually extracted with and without fibrous regions at 5x magnification. The GC-based segmentation algorithm was applied to the above images and a significance test was performed for the segmented regions. Variance was set to $\Sigma = \hat{\sigma}^2 I_{n\times n}$, where $\hat{\sigma}^2$ was estimated from independent data with the maximum likelihood method. In this experiment, the seed regions were manually selected, but the effect of the manual selection was not considered when computing $p$-values. Figures 4 and 5 show the results. It can be observed that the $p$-values obtained with the proposed method are smaller than $\alpha = 0.05$ only when there are actually fibrous regions in the images. In contrast, the naive method always gives $p$-values that are zero, even for images that do not contain fibrous regions.

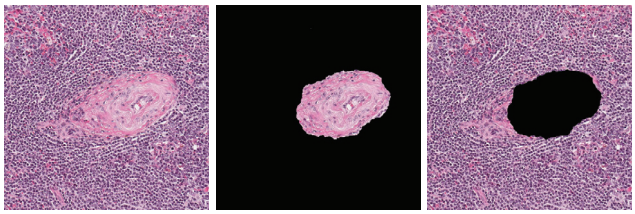
In experiments with CT images, we aimed to extract the tumor region in the liver [23, 40, 1, 3, 16, 28]. In the experiments, we used CT images from the 2017 MICCAI Liver Tumor Segmentation Challenge. Here, the local TH-based segmentation algorithm was employed for identifying liver tumor regions since CT values in tumor regions are lower than those in surrounding organ regions. Before applying the local TH algorithm, original images were blurred with Gaussian filtering with a filter size of $11\times11$. The parameters for local thresholding were a window size of 50 and $\theta = 1.1$. The results of local thresholding for CT images are shown in Figures 6 and 7. It can be observed that the $p$-values obtained with the proposed method (selective-$p$) are smaller than the significance level $\alpha = 0.05$ only when there are actually tumor regions in the images. In contrast, the naive method always gives $p$-values that are zero, even for images that do not contain tumor regions.

## 5. Conclusions

In this paper, we proposed a novel framework called PSegI for providing a reliability metric for individual segmentation results by quantifying the statistical significance of the difference between the object and background regions in the form of $p$-values.

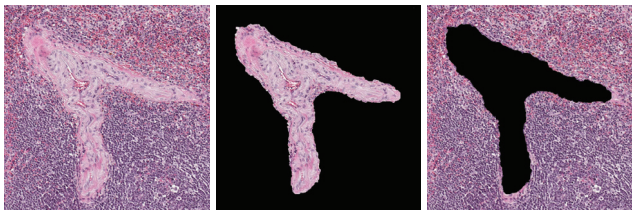

## Acknowledgment

This work was partially supported by MEXT KAKENHI 17H00758, 16H06538 to I.T., JST CREST JPMJCR1502 to I.T., and RIKEN Center for Advanced Intelligence Project to I.T.



(a) Original (b) Object (c) Background

(naive-$p$ = **0.00** and selective-$p$ = **0.00**)



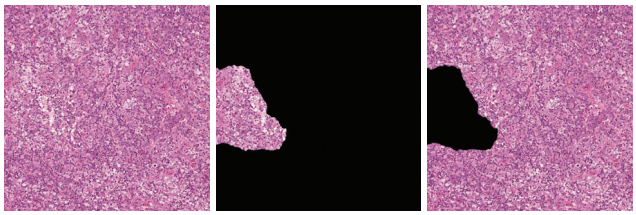
(d) Original (e) Object (f) Background

(naive-$p$ = **0.00** and selective-$p$ = **0.00**)

Figure 4: Segmentation results for pathological images with fibrous regions. The $p$-values obtained with the proposed method (selective-$p$) are smaller than $\alpha = 0.05$, indicating that these segmentation results correctly identified the fibrous regions.

(a) Original (b) Object (c) Background

(naive-$p$ = **0.00** and selective-$p$ = 0.35)



(d) Original (e) Object (f) Background

(naive-$p$ = **0.00** and selective-$p$ = 0.73)

Figure 5: Segmentation results for pathological images without fibrous regions. The $p$-values obtained with the proposed method (selective-$p$) are greater than $\alpha = 0.05$, indicating that the differences between the two regions in these images are deceptively large due to segmentation bias. It is obvious that these images do not contain specific objects.

(a) Original (b) Blurred (c) Binarized

(naive-$p$ = **0.00** and selective-$p$ = **0.00**)

(d) Original (e) Blurred (f) Binarized

(naive-$p$ = **0.00** and selective-$p$ = **0.00**)

Figure 6: Segmentation results for CT images with tumor regions. The $p$-values obtained with the proposed method (selective-$p$) are smaller than $\alpha = 0.05$. These images contain ground-truth tumor regions, which were successfully identified by the segmentation algorithm.

(a) Original (b) Blurred (c) Binarized

(naive-$p$ = **0.00** and selective-$p$ = 0.21)

(d) Original (e) Blurred (f) Binarized

(naive-$p$ = **0.00** and selective-$p$ = 0.77)

Figure 7: Segmentation results for CT images without tumor regions. The $p$-values obtained with the proposed method (selective-$p$) are greater than $\alpha = 0.05$. These images do not contain any ground-truth tumor regions. The differences between the two regions in these images are deceptively large due to segmentation bias.

## References


[1] K. T. Bae, M. L. Giger, C.-T. Chen, and C. E. Kahn. Automatic segmentation of liver structure in ct images. *Medical physics*, 20(1):71–78, 1993. 7

[2] Y. Boykov and G. Funka-Lea. Graph cuts and efficient nd image segmentation. *International journal of computer vision*, 70(2):109–131, 2006. 1, 2, 4, 5

[3] Y. Boykov and M.-P. Jolly. Interactive organ segmentation using graph cuts. In *International conference on medical image computing and computer-assisted intervention*, pages 276–286. Springer, 2000. 7

[4] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (9):1124–1137, 2004. 4

[5] Y. Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, volume 1, pages 105–112. IEEE, 2001. 1, 2, 4, 5

[6] D. Comaniciu and P. Meer. Cell image segmentation for diagnostic pathology. pages 541–558, 2002. 7

[7] A. Desolneux, L. Moisan, and J.-M. More. A grouping principle and four applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(4):508–513, 2003. 2

[8] A. Desolneux, L. Moisan, and J.-M. Morel. Meaningful alignments. *International Journal of Computer Vision*, 40(1):7–23, 2000. 2

[9] E. A. Dinic. Algorithm for solution of a problem of maximum flow in networks with power estimation. In *Soviet Math. Doklady*, volume 11, pages 1277–1280, 1970. 4

[10] V. N. L. Duy, H. Toda, R. Sugiyama, and I. Takeuchi. Computing valid p-value for optimal changepoint by selective inference using dynamic programming. *arXiv preprint arXiv:2002.09132*, 2020. 2

[11] A. Elnakib, G. Gimelfarb, J. S. Suri, and A. El-Baz. Medical image segmentation: a brief survey. In *Multi Modality State-of-the-Art Medical Image Segmentation and Registration Methodologies*, pages 1–39. Springer, 2011. 2

[12] W. Fithian, D. Sun, and J. Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014. 2, 4

[13] L. R. Ford and D. R. Fulkerson. Maximal flow through a network. In *Classic papers in combinatorics*, pages 243–248. Springer, 2009. 4

[14] A. V. Goldberg and R. E. Tarjan. A new approach to the maximum-flow problem. *Journal of the ACM (JACM)*, 35(4):921–940, 1988. 4

[15] L. Grady. Random walks for image segmentation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11):1768–1783, 2006. 2

[16] Y. Gu, V. Kumar, L. O. Hall, D. B. Goldgof, C.-Y. Li, R. Korn, C. Bendtsen, E. R. Velazquez, A. Dekker, H. Aerts, et al. Automated delineation of lung tumors from ct images using a single click ensemble segmentation approach. *Pattern recognition*, 46(3):692–702, 2013. 7

[17] T. Hershkovitch and T. Riklin-Raviv. Model-dependent uncertainty estimation of medical image segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1373–1376. IEEE, 2018. 2

[18] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017. 2

[19] N. Kriegeskorte, W. K. Simmons, P. S. Bellgowan, and C. I. Baker. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, 12(5):535, 2009. 2

[20] J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. Exact post-selection inference, with application to the lasso. *Annals of Statistics*, 44(3):907–927, 2016. 2, 4, 5

[21] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma. A survey of content-based image retrieval with high-level semantics. *Pattern recognition*, 40(1):262–282, 2007. 2

[22] J. R. Loftus and J. E. Taylor. Selective inference in regression models with groups of variables. *arXiv preprint arXiv:1511.01478*, 2015. 2, 5

[23] L. Massoptier and S. Casciaro. A new fully automatic and robust algorithm for fast segmentation of liver tissue and tumors from ct scans. *European radiology*, 18(8):1658, 2008. 7

[24] N. Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979. 1, 2, 4, 5

[25] K. Qin, K. Xu, F. Liu, and D. Li. Image segmentation based on histogram analysis utilizing the cloud model. *Computers & Mathematics with Applications*, 62(7):2824–2833, 2011. 2

[26] F. Rousseau, F. Blanc, J. de Seze, L. Rumbach, and J.-P. Armspach. An a contrario approach for outliers segmentation: Application to multiple sclerosis in mri. In *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 9–12. IEEE, 2008. 2

[27] M. Sezgin and B. Sankur. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic imaging*, 13(1):146–166, 2004. 4

[28] A. Shimizu, R. Ohno, T. Ikegami, H. Kobatake, S. Nawano, and D. Smutek. Segmentation of multiple organs in non-contrast 3d abdominal ct images. *International journal of computer assisted radiology and surgery*, 2(3-4):135–142, 2007. 7

[29] S. Suzumura, K. Nakagawa, Y. Umezu, K. Tsuda, and I. Takeuchi. Selective inference for sparse high-order interaction models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3338–3347. JMLR. org, 2017. 2

[30] V.-T. Ta, O. Lézoray, A. Elmoataz, and S. Schüpp. Graph-based tools for microscopic cellular image segmentation. *Pattern Recognition*, 42(6):1113–1125, 2009. 7

[31] J. Taylor, R. Lockhart, R. J. Tibshirani, and R. Tibshirani. Exact post-selection inference for forward stepwise and least angle regression. *arXiv preprint arXiv:1401.3889*, 7:10–1, 2014. 2

[32] J. Taylor and R. Tibshirani. Post-selection inference for-penalized likelihood models. *Canadian Journal of Statistics*, 46(1):41–61, 2018. 2

[33] J. Taylor and R. J. Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015. 2

[34] X. Tian, J. Taylor, et al. Selective inference with a randomized response. *The Annals of Statistics*, 46(2):679–710, 2018. 2

[35] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. 2

[36] R. J. Tibshirani, J. Taylor, R. Lockhart, and R. Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016. 2

[37] R. G. Von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall. Lsd: a line segment detector. *Image Processing On Line*, 2:35–55, 2012. 2

[38] J. M. White and G. D. Rohrer. Image thresholding for optical character recognition and other applications requiring character image extraction. *IBM Journal of research and development*, 27(4):400–411, 1983. 1, 2, 3, 5

[39] Y. Xu, J.-Y. Zhu, I. Eric, C. Chang, M. Lai, and Z. Tu. Weakly supervised histopathology cancer image segmentation and classification. *Medical image analysis*, 18(3):591–604, 2014. 7

[40] X. Ye, G. Beddoe, and G. Slabaugh. Automatic graph cut segmentation of lesions in ct using mean shift superpixels. *Journal of Biomedical Imaging*, 2010:19, 2010. 7

[41] B. Zhao, J. Feng, X. Wu, and S. Yan. A survey on deep learning-based fine-grained object classification and semantic segmentation. *International Journal of Automation and Computing*, 14(2):119–135, 2017. 2