A framework for statistically-sound customer segment search

Sihem Amer-Yahia, Laure Berti-Equille, Abdelouahab Chibah

A Framework for Statistically-Sound Customer Segment Search Authors' Copy

Sihem Amer-Yahia
CNRS, University Grenoble Alpes, France
sihem.amer-yahia@univ-grenoble-alpes.fr

Laure Berti-Equille
IRD, ESPACE-DEV, France
laure.berti@ird.fr

Abdelouahab Chibah
CNRS, University Grenoble Alpes, France
abdelouahab.chibah@univ-grenoble-alpes.fr

***Abstract*—We develop *S4*, a Statistically-Sound Segment Search framework that combines principled data partitioning and sound statistical testing to verify common hypotheses in retail data and return interpretable customer data segments. Our framework accommodates one-sample, two-sample, and multiple-sample testing, to provide various aggregations and comparisons of customer transactions. To control the proportion of false discoveries in multiple hypothesis testing, we enforce an FDR-controlling procedure and formulate a unified optimization problem that returns customer data segments that satisfy the test for a given significance level, maximize coverage of the input data, and are within a risk capital. We develop a greedy algorithm to explore different data partitions and test multiple hypotheses in a sound manner. Our extensive experiments on four retail data sets examine the interaction between significance, risk and coverage, and demonstrate the expressivity, usefulness, and scalability of *S4* in practice.**

I. Introduction

The Internet of Behaviors (IoB) trend[1] calls for developing expressive and robust methods for exploring data about people to capture changes in their behaviors. Understanding customer behavior enables actionable insights for social science studies and marketing campaigns. While several approaches in Marketing and Econometrics were proposed to test hypotheses from customer data, their applicability is limited as it relies on carefully choosing groups of customers for the hypothesis testing, which in turn requires domain expertise. The availability of large amounts of customer data today constitutes a great opportunity to develop powerful means to test and compare multiple hypotheses on customer behaviors and preferences. In this work, we develop ***S4***, a framework that seamlessly integrates principled data partitioning and hypothesis testing to find and compare relevant customer groups in a statistically-sound manner.

Illustrative example. Consider a marketing analyst interested in running new promotions for *customer segments*, e.g., *Young males whose purchase average increased in the month following a promotion*. Figure 1 illustrates this scenario in 3 steps. In **Step 1**, the analyst seeks to explore how responsive customers were to previous promotions. A two-sample test identifies segments *whose weekly purchase average* remained the same after a promotion. For each qualifying segment, the analyst further explores its demographics in **Step 2** using a two-sample test, to compare whom among *males or females* have a *lower average*. This results in female segments which are used in a one-sample test (**Step 3**) to verify if the proportion of young females among them is higher than 50% of the overall population. The returned segments are a good target for new promotions.

Challenges. Realizing our example requires to address two challenges: (i) identify customer segments flexibly and exhaustively, and (ii) conduct rigorous multiple hypothesis testing at each step. The first challenge lies in the lack of a principled way to explore different data partitionings. The common practice is to perform data dredging[2], a tedious error-prone process. The second challenge lies in the number and variety of statistical tests (with requirements on data normality and independence) to verify if the customer behavior supports the null or the alternative hypothesis. There are precise criteria for excluding or not a null hypothesis at a certain significance level [1], [2]. Those criteria depend on the type of test (e.g., one-sample, two-sample, or multiple-sample), the aggregation function (e.g., mean, variance, proportion), the sample sizes, whether the samples are paired (same subjects), etc. Existing work on statistically-sound pattern mining [3], [4] falls short in addressing those needs as it does not provide an expressive yet simple way to identify, combine, and test relevant data partitions.

Our solution. We propose ***S4***, our framework for expressing **S**tatistically-**S**ound **S**egment **S**earch. ***S4*** combines powerful data partitioning with multiple hypothesis testing. Data partitioning is achieved via pivoting and segmentation. We define *promotion-based and demographics-based pivoting* as a form of hold-out evaluation that divides the input data into an exploratory set and a hold-out set [5], [6]. Segmentation is then applied to each set to generate candidate partitions. The common practice in marketing is to segment customers based on who they are (e.g., demographics such as age), or on what they do, i.e., their behavior, such as how much they spend

[1] https://www.gartner.com/smarterwithgartner/gartner-top-strategic-technology-trends-for-2021/

[2] https://en.wikipedia.org/wiki/Data_dredging

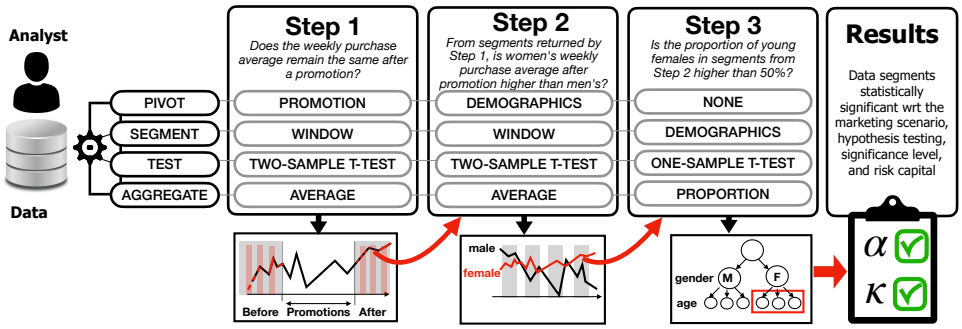


Fig. 1: An illustrative scenario of a marketing analyst in 3 steps: (1) explore how responsive customers were to previous promotions, (2) compare the purchase amounts from male or female customers to find which group has a lower average, and (3) verify if the proportion of young customers in this group is higher than 50% of the overall population.

and how often [3]. We adopt this practice and propose two segmentation modalities: a *demographics-split* and a time-based or point-based *window-split* [7]. The combination of pivoting and segmentation yields candidate data partitions which are fed to multiple hypothesis testing to perform one-sample, two-sample, and multiple-sample tests and return segments that pass the test.

Overcoming the problems of multiple hypothesis testing. When multiple hypotheses are tested, the chance of observing a rare event increases, and hence, the likelihood of incorrectly rejecting a null hypothesis (i.e., making a Type I error [8]) increases. A typical way to mitigate this problem is to adjust the significance threshold of each hypothesis [1], [3], [4]. In our work, we use the Benjamini-Yekutieli False Discovery Rate (FDR) procedure [9] that was designed to adjust the significance of dependent hypotheses. FDR-controlling procedures provide less stringent control of errors compared to Family-Wise Error Rate (FWER) procedures (e.g., the Bonferroni correction [10]), which controls the probability of at least one Type I error. FDR not only reduces the number of false positives, but it also minimizes the number of false negatives. This makes it better suited to our context where different segment searches are composed in a scenario and the result of one identifies "candidate positives" for subsequent ones in the pipeline.

The degradation of p-values into significant and non-significant, is not well-adapted to a large number of hypotheses [11] and has been shown to prevent the acceptance of many true discoveries [12]. To mitigate that, we propose to use a risk capital that is computed as the sum of p-values of hypotheses that pass the test, and to seek segments that cover as much as possible the input data set. We formulate the ***S4*** Problem, an optimization problem that admits a data set, a pivot, a segmentation, a significance level, and an upper-bound on risk (i.e., a risk capital), and returns segments that satisfy the test and are within the capital. This formulation unifies one-sample tests where a single segment is compared against a reference value, two-sample tests where two segments are compared (e.g., before and after a promotion), as well as multiple-sample tests where the statistic measure (e.g. mean, variance, proportion) of a segment is compared to several others. The ***S4*** problem is similar to the 0-1 Knapsack problem [13], making it NP-hard. To solve it, we develop a greedy algorithm. Our algorithm makes use of a primitive *nextCandidate()* that returns the next best candidate hypothesis to test. This primitive has one of two semantics: *p-value scan* that returns the candidate with the next smallest p-value and reflects what is commonly used in multiple hypothesis testing, and *coverage scan* that returns the candidate with the next highest coverage of the input data set. This allows us to explore the relationship between significance, risk and coverage.

Our experiments on four real-world data sets `RETAIL`, `TAFENG`, `SALES`, and `AMAZON`, clearly demonstrate that leveraging coverage for hypotheses filtering is useful and scalable. We show that the number of results is reduced by 1 to 3 orders of magnitude while preserving coverage of the input data. This has the additional benefit of being highly scalable, thereby enabling interactive times and the composition of segment searches in a pipeline.

Our contributions.

1) We develop the ***S4*** framework that combines pivoting, segmentation, and multiple hypothesis testing in a principled manner. To enhance the reliability of our hypothesis testing, we leverage the powerful Benjamini-Yekutieli FDR procedure to control false discoveries for dependent hypotheses.
2) We formalize a unified ***S4*** problem that captures one-sample, two-sample, and multiple-sample tests and returns segments that satisfy the test (i.e, accept the null hypothesis at a given significance level), cover the input data, and are within a user-defined risk capital.
3) We develop the ***S4*** algorithm that scans hypotheses either in increasing p-values or in decreasing coverage.
4) We run extensive experiments showing that ***S4*** addresses today's information needs in advanced data analytics scenarios while mitigating the risk of multiple hypothesis testing.

Organization. We provide a summary of the related work

[3]https://www.qualtrics.com/experience-management/brand/what-is-market-segmentation/

in Section II. Section III introduces the ***S4*** framework and defines the ***S4*** Problem. Section IV contains our solution. Our experiments are described in Section V and a discussion follows in Section V-F. The conclusion is provided in Section VI.

II. Related work

Existing work on combining hypothesis testing with powerful customer segment discovery [3], [4] benefits from the computational and statistical aspects of pattern mining with an emphasis on controlling the risk of false discoveries, i.e., patterns found in the data sample that do not hold in the entire population. The first step is to find genuine patterns that are likely to reflect properties of the underlying population and hold also in the data samples. A variety of statistical tests have been used to filter out patterns that are unlikely to be useful, removing uninformative variations of key patterns in the data [5].

A statistical hypothesis test compares two models (the null hypothesis and the alternative hypothesis) and deems the comparison statistically significant if, according to a significance threshold, the data is very unlikely to have occurred under the null hypothesis. Deciding a significance threshold is hard. In particular, when a large number of hypotheses are tested, raw p-values often yield very few significant hypotheses [1]. This requires the use of powerful p-value correction methods [2], [11] such as Family-Wise Error Rate (FWER) or False Discovery Rate (FDR) [14]. The idea of controlling risk within a budget was introduced in [12].

Our work proposes a framework that combines expressive data partitioning and multiple hypothesis testing, to generate interpretable customer segments. Additionally, it explores the intricate relationship between setting a significance level, aiming to cover the input data, and capping risk. To the best of our knowledge, this is the first approach of this kind.

III. The ***S4*** framework

We start with examples to illustrate the types of hypotheses we support. We then present the ***S4*** model and formalize our problem.

A. Motivating Examples

Our purpose is to develop a powerful framework to verify hypotheses on customer transaction data. A hypothesis is verified when *customer segments* that are sets of transactions identified by particular *filters*, have *similar, higher, lower aggregates* with respect to some *pivot*. An aggregate can be the mean, variance, or count of purchases. A pivot can be *promotion-based* or *demographics-based*. It separates an input data set D into an exploratory subset D_E and a hold out set D_H such that $D_E \cup D_H = D$ and $D_E \cap D_H = \emptyset$. The two sets D_E and D_H are further partitioned using one of two segmentations: *demographics-split* or *window-split*, that generate subsets to be compared via hypothesis testing.

Figure 2 presents as examples various questions a data analyst may ask and they will be used throughout the paper to illustrate the analysis scenarios we consider. Segments are highlighted in green, pivots in red, aggregates in blue, and operators in orange. Figure 3 shows a 2D representation of segment search with the pivots and segmentation types we consider. We discuss the three examples highlighted in the figure.

Our first example (#Q1 in the figure) is the simple case where there is no pivot. Segmentation type is based on 2-day time windows that also include week-ends. The average number of purchases is compared to a reference value (300). Here, segments corresponding to week-ends and satisfying the null hypothesis (using a one-sample t-test) are searched in the data set returned.

Our second example (#Q4) is a marketing analyst who examines the impact of a promotion (pivot) on male and female customers (demographics segmentation). The null hypothesis is "women's purchase average is the same as men's after promotion". A (two-sample t-test will compare the average number of purchases of men and women before and after the promotion period which will result in accepting or rejecting the null hypothesis.

Our third example (#Q9)) is a social scientist seeking to compare the number of transactions by males and females on week-ends (the null hypothesis) or different. This requires a demographics pivot on gender coupled with a window-based segmentation that compares 2-day periods (using a two-proportion Z-test).

B. Our Model

Our data is represented as time series in the form of a sequence of pairs $\mathcal{D} = [(d_1, s_1), \ldots, (d_n, s_n)]$ $(s_1 < s_2 < \ldots s_{n-1} < s_n)$, where each d_i is a triplet $\langle r_i, c_i, p_i \rangle$ where r_i is a unique transaction identifier, c_i is a customer identifier, and p_i is a product purchased by customer c_i at time s_i, the timestamp at which d_i occurs. Each customer is defined by a set of demographic attributes $A = \{a_1, \ldots, a_n\}$ such as age, location and gender.

1) Pivots and Segments: We define the notion of pivot V in the spirit of hold-out evaluation [4]. The pivot breaks down a data set $D \subseteq \mathcal{D}$ into an exploratory set D_E and a hold out set D_H. We introduce two types of pivots: **promotion-based** where D_E is the subset of D containing transactions that precede the promotion period, and D_H is the subset of D that contains transactions following the promotion period. Other types of events could be considered such as Christmas and Chinese New Year; and **demographics-based**, e.g., gender, where for instance, D_E is the subset of D where every d_i contains a male customer c_i, and D_H is the subset of D where every d_i contains a female customer c_i.

Following marketing practices, we define two customer segmentation types that produce non-overlapping partitions of D_E and D_H: **demographics-split** where each partition represents transactions generated by customers with the same values for conjunction of attributes in A. For instance, a

$\#Q_i$	Questions
#Q1	Is the **average** number of purchases during week-ends *greater* than 300?
#Q2	Does the **average** number of purchases of females and males *differ* ?
#Q3	Is the **average** number of weekly purchases the *same* before or after promotion?
#Q4	Does the women's purchase **average** *exceed* men's after promotion ?
#Q5	Is the **average** number of purchases overall *higher* after promotion?
#Q6	Is there a *difference* between the **average** number of purchases of young males, older males, and females?
#Q7	Is the **number** of purchases by men and women *higher* depending on location ?
#Q8	Is the overall purchase **average** *greater* than 300?
#Q9	Are the **number** of purchases of males and females *different* ?

Fig. 2: Examples of questions in marketing handled in ***S4*** with pivot, segmentation, **aggregate**, and *operator*

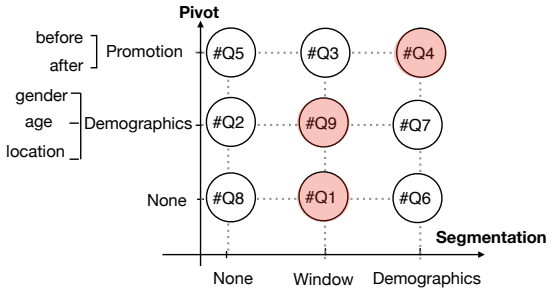


Fig. 3: Grid of ***S4*** pivots (Promotions or Demographics) and segmentation types (Demographics and Window) with the questions of our example

AGGREGATE	STATISTICAL TEST	HYPOTHESIS AND TEST DEFINITION		EXAMPLE
Mean μ	Test about a mean: **One-sample t-test**	$H_0 : \mu = \mu_0$ $H_a : \mu \neq \mu_0$ or $\mu > \mu_0$ or $\mu < \mu_0$	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ with $\bar{x}$ and s , the sample mean and standard deviation, and μ_0 the reference mean	#Q1.
	Test to compare two means: **Two-sample t-test** (or Welch's t-test for unequal variances)	$H_0 : \mu_1 = \mu_2$ $H_a : \mu_1 \neq \mu_2$ or $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$	$t = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{\eta}$ with $\bar{x}_1 - \bar{x}_2$ the difference between 2 sample means, μ_0 the reference mean, s the pooled standard deviation, and $\eta = \begin{cases} s\sqrt{2/n} \text{ for } n = n_1 = n_2 \\ s\sqrt{1/n_1 + 1/n_2} \text{ for } n_1 \neq n_2 \text{ and } s = s_1 = s_2 \\ \sqrt{s_1/n_1 + s_2/n_2} \text{ for } n_1 \neq n_2 \text{ and } s_1 \neq s_2 \end{cases}$	#Q2. #Q3. #Q4.
	Test about a mean with paired data: **Paired difference t-test**	$H_0 : \mu_D = 0$ $H_a : \mu_D \neq 0$ or $\mu_D > 0$ or $\mu_D < 0$	$t = \frac{\bar{d} - d_0}{s_d/\sqrt{n}}$ with $\bar{d}$ and s_d the average and standard deviation of the differences between all pairs, and d_0 the reference difference	#Q5.
	Test to compare multiple means: **F-test for one way ANOVA**	$H_0 : \mu_1 = \mu_2 = \ldots = \mu_K$ H_a : not all the means are equal	$F = \frac{MST}{MSE}$ with $n = n_1 + \cdots + n_K$, $MST = \frac{\sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x})^2}{K-1}$, $\bar{x} = \frac{\sum_{i=1}^{K} x_i}{n}$, and $MSE = \frac{\sum_{i=1}^{K} (n_i - 1)s_i^2}{n-K}$	#Q6.
Variance σ	Test to compare two population: variances: **F-test**	$H_0 : \sigma_1^2 = \sigma_2^2$ $H_0 : \sigma_1^2 \neq \sigma_2^2$	$F = s_1^2/s_2^2$ with σ_1^2 and σ_2^2 , the 2 population variances and s_1^2 and s_2^2 , the sample variances of the 2 populations	#Q7.
Proportion p	Test about a proportion: **One proportion Z-test**	$H_0 : p = p_0$ $H_a : p \neq p_0$ or $p > p_0$ or $p < p_0$	$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$ with p_0 the reference population, $\hat{p}$ the sample proportion	#Q8.
	Test to compare two proportions: **Two proportion Z-test**	$H_0 : p_1 = p_2$ $H_a : p_1 \neq p_2$ or $p_1 > p_2$ or $p_1 < p_2$	$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$ with $p_1 - p_2$ the difference in 2 population proportions, $\hat{p}_1 - \hat{p}_2$ the difference in 2 proportions of the samples x_1 and x_2 , and $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$	#Q9.

Fig. 4: Summary of statistical tests considered in ***S4***

partitioning with attributes gender and location generates non-overlapping subsets of transactions with all combinations of values of gender and location (e.g., males in the north); and **window-split** that partitions data into equal-length segments of consecutive transactions. This is commonly used in drift detection [7] and can be either time-based resulting in windows of the same length, or point-based resulting in windows with the same size (in our case, the same number of transactions).

2) Hypotheses and Segment Search: The choice of a pivot V determines the two data subsets for which a hypothesis is tested: the exploratory set D_E and the hold-out set D_H. The choice of a segmentation S splits D_E (resp., D_H) into non-overlapping segments s.t. $\bigcup_{s_E} = D_E$ (resp., $\bigcup_{s_H} = D_H$).

A hypothesis H is defined as a quadruple $(H_0, H_a, \texttt{AGG}, \texttt{OP})$ where H_0 is the null hypothesis, H_a the alternative hypothesis, AGG is an aggregate measure applied to customer purchases (average, variance, proportion), and OP the operator used to compare aggregates $(=, <>, >, and <)$. Figure 4 summarizes the aggregates and statistical tests considered in this work, along with examples from Figure 2.

In our framework, a statistically-sound segment search Q is a quadruple (H, D, V, S) that explores all data segments, noted $allSegments$ to test the hypothesis. Table I provides the exact definition of $allSegments$ for each test type used in Q.

3) Statistical Testing: The ***S4*** framework is aimed to be generic and accommodate various tests. Different statistical tests qualify depending on sample size, the subjects they contain (paired or unpaired), and the aggregation function AGG. Normality and independence of candidate segment pairs, noted $Candidates$ are checked[4] before testing and computing their p-value. Figure 4 summarizes our definitions for each test type. For instance, a one-sample t-test is used to compare a

[4]See https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3116565/ for applicability conditions of statistical tests.

S4 Segment Search	Definitions and Coverage
with one-sample test	$allSegments = \{s_E\}$ $Candidates = \{(s_E, pval)\}$ $cover(R, D) = \frac{\Sigma_{s_E \in R} \lvert s_E \rvert}{\lvert D_E \rvert}$
with two-sample test	$allSegments = \{(s_E, s_H)\}$ $Candidates = \{(s_E, s_H, pval)\}$ $cover(R, D) = \Sigma_{(s_E, s_H) \in R}(\frac{\lvert s_E \rvert}{\lvert D_E \rvert} + \frac{\lvert s_H \rvert}{\lvert D_H \rvert})/2$
with multiple-sample test	$allSegments = \{(s_E, \{s_H\})\}$ $Candidates = \{(s_E, \{s_H\}, pval)\}$ $cover(R, D) = \Sigma_{(s_E, \{s_H\}) \in R}(\frac{\lvert s_E \rvert}{\lvert D_E \rvert} + \frac{\lvert \bigcup_{s_H} \rvert}{\lvert D_H \rvert})/2$

TABLE I: Summary of our definitions

mean to a hypothetical value (row #1 of Figure 4) and the p-value is computed for individual segments s_E. For comparing two means, two-sample t-tests are computed for (s_E, s_H) pairs (row #2). For comparing multiple means, an F-test for one way ANOVA is applied to $(s_E, \{s_H\})$ pairs (row #4). In all cases, we adopt a common protocol to compute p-values [15] as described as follows for comparing two means as `AGG` with a two-sample t-test.

P-value Computation Protocol:

1) **Normality check:** Verify that the data distributions of each segment s_E and s_H is normal, normalize it otherwise;
2) **Independence filtering:** Verify that the distributions of s_E ad s_H are independent using χ^2 test; Keep independent pairs;
3) **P-value computation:** Compute the p-value of independent (s_E, s_H) pairs wrt to hypothesis H and add them to the set of candidate hypotheses $Candidates$.

The set *Candidates* along with their p-values (see Table I) is given as input to a subsequent step to control false discoveries.

4) Controlling false discoveries: As the number of candidates increases, the likelihood that spurious hypotheses pass the test increases, causing Type I errors [8]. The significance level of p-values can be adjusted to control the expected proportion of incorrectly rejected null hypotheses. The simplest way to do so is to use the conservative Bonferroni correction [10], a Family-Wise Error Rate (FWER) control method. The Bonferroni correction is preferred when false discoveries are not acceptable (in particular for critical decision-making, e.g., accepting a new medical treatment) or when it is expected that most null hypotheses would be true. A different and more powerful adjustment method is the Benjamini-Yekutieli False Discovery Rate (FDR) procedure [9] that was designed for dependent hypotheses and allows to control the expected proportion of incorrectly rejected null hypotheses. FDR control is preferred in exploratory research, where the number of potential hypotheses is large and false discoveries are not so critical [4].

Both FWER and FDR methods use a significance level α. A value of 0.05 for α indicates a 5% risk of concluding that a difference exists between the two means when there is no actual difference. When the p-value is less than or equal to the significance level α, the null hypothesis is rejected (means are not equal). FWER and FDR control methods readjust the significance threshold α to be more stringent. In this paper, we advocate the use of Benjamini-Yekutieli and we will compare it to the Bonferroni correction in our experiments.

Additionally, we conjecture that re-adjusting the significance level α may return segments that constitute "microphenomena" [4], [11] due to the large number of hypotheses to test. Therefore, we propose two additional mechanisms tested in our experiments. First, we control risk by setting an upper-bound (i.e., a risk capital allocated to the sum of p-values of hypotheses that pass the test), as it was defined for multiple hypothesis testing [12]. Second, we select hypotheses that cover D as much as possible. Table I provides the definition of $cover(R, D)$ for $R \subseteq Candidates$.

C. Problem Statement

We are now ready to formally define the ***S4*** problem. Given

- Q a segment search (H, D, V, S), with hypothesis H to be tested with significance level α on a data set D pivoted and segmented according to V and S respectively,
- κ, a user-defined risk capital, and
- m, the total number of candidates that passed the normalization and independence protocol wrt H, ($m = \lvert Candidates \rvert$),

our problem is to find a set R^* such that:

$$R^* = \underset{R \subseteq Candidates}{\texttt{argmax}} \; cover(R, D)$$

subject to

$$r.pval \leq \frac{\alpha n}{m} \left(\sum_{i=1}^{n} 1/i \right)^{-1} \text{ and } \sum_{r \in R} r.pval \leq \kappa \tag{1}$$

with r is the n^{th} element in R.

This formulation relies on Benjamini Yekutieli's significance adjustment and captures the tests presented in Figure 4. The ***S4*** problem is a variant of the 0-1 Knapsack problem where coverage is akin to the value and the p-values are akin to the weights in the original formulation [13].

IV. Our solution

We present Algorithm 1, our greedy solution to solve the ***S4*** Problem. Initially, the input data D is divided into D_E and D_H subsets based on the specified pivot V (line #1). Then, the algorithm applies a segmentation (function *genSegments()* in line #2) to partition D_E and D_H into all segments, *allSegments*, to be tested wrt the hypothesis H. The next step applies the p-value computation protocol (line #3) and is followed by finding the best results along with their

Algorithm 1: Pseudo-code of the ***S4*** algorithm

Input: H, D, V, S, α, κ
Output: (R, P) Results with their p-values satisfying H for α and κ.

1 $pivot(D, V) := (D_E, D_H)$
2 $allSegments := genSegments(H, D_E, S_H, S)$
3 $Candidates := computePvals(H, allSegments)$
// Applies the p-value computation protocol
4 $m := |Candidates|; R, P := \emptyset; i := 0$
5 **while** $((Candidates <> \emptyset) \wedge (\kappa > 0))$ **do**
6 *best := nextCandidate(Candidates)*
7 *n :=* $|R|$; $i := i + 1$
8 **if** best.pval $\leq \frac{\alpha n}{m \Sigma_{i=1}^{n}(1/i)}$ // According to Eq.(1)
9 **then**
10 *R := R* $\cup$ *best*; *P := P* $\cup$ *best.pval*
11 *Candidates := Candidates* $\setminus$ *best*
12 κ := κ - *best.pval*
13 **else**
14 Break // for p-value scan
15 Continue // for coverage scan
16 **return** (R, P)

p-values (lines #5 to #15). To save space, we refer the reader to an extended version of our paper for more details.

As discussed in Section III-B4, there are many correction methods to adjust the p-value significance threshold. An important aspect when designing an algorithm for multiple hypothesis testing is that its outcome and power largely depend on **the order in which candidates and their p-values are tested** [4]. Bonferroni and Sidak corrections [10] are examples of single-step methods, where the same adjusted significance level is applied to all hypotheses (i.e., pairs of segments). Other methods are step-wise and determine individual significance levels for each hypothesis, depending on the order of p-values and rejection of other hypotheses. In general, single-step methods are considered to be the least powerful, and step-wise methods, where hypotheses are scanned in increasing order of their p-values, are the most powerful [4].

In this paper, we face an additional challenge that stems from the need to stay within risk capital and to cover the input set D. We propose a step-wise method where the next best candidate result is generated and its p-value is tested. Step-wise methods are computationally expensive because they require all hypotheses to be sorted based on their p-values. However, they can be applied in multi-stage procedures (e.g., [5], [6]) that first select constrained sets of candidates which are subsequently tested. To implement that, we use *nextCandidate()* (line #6) that encapsulates two semantics: *p-value scan* that returns candidates in increasing p-values to mimic step-wise methods; *coverage scan* that returns candidates in decreasing coverage to maximize coverage over D. This second approach is effectively solving the problem in Eq.(1) while the one based on p-value scan is the traditional method.

At each iteration, the algorithm uses the Benjamini-Yekutieli test with an adjusted significance level and decides to retain or not the current candidate (line #8). If retained, it removes its p-value from the risk capital κ (line #12). The algorithm iterates until either there is no result left to test in *Candidates*, or when adding the best candidate's p-value exceeds the risk capital.

In the case of p-value scan, the algorithm stops scanning candidates as soon as it encounters a candidate whose p-value does not satisfy the Benjamini Yekutieli condition (line #14). In the case of coverage scan, the algorithm is expected to reach higher values of coverage earlier. These two cases will be examined empirically. The overall worst-case complexity of our algorithm (line #3 onward) is $O(m^2)$ where $m = |Candidates|$. The complexity of segmentation is variable. For window-based, it is $O(|D_E| + |D_H|)$ since the sets are scanned once to create segments. For demographics-based, we use a partition decision tree to favor larger segments. Its complexity is $O(|D_E| \times |A|)$ (resp. $O(|D_H| \times |A|)$) where A is the number of customer data attributes. The complexity of the p-value computation protocol is $O(|allSegments|)$. The worst-case complexity of *nextCandidate()* is $O(mlogm)$ for both semantics since it relies on sorting the candidates either on p-value or on coverage.

V. Experiments

The purpose of the experiments is to: (1) Demonstrate the expressivity of ***S4*** on a variety of data sets and hypotheses (Section V-B); (2) Examine the relationship between significance, risk capital and coverage for different data partitionings (Section V-C); (3) Report the scalability of our algorithm as a function of data size (Section V-E).

A. Experimental Setup

Data sets. We chose real-world data sets that offer different opportunities for data partitioning. Due to space limitations, we have carefully chosen a subset of our results to report in this paper. We refer the reader to our GitHub repository[5] where our code, complete results, and all our segment search examples are made available.

`RETAIL` is a proprietary data set from an industrial partner, containing 250,208 transactions generated by 32,160 unique customers and 7,404 products and spanning a period of 34 months from Feb. 2017 to Dec. 2019. The data includes information on customer location and gender, and a 10-week Prom period (from Dec. 17, 2018 to Feb. 28, 2019) that we used as a pivot. The period preceding the Prom has a total of 90K transactions and the period following it contains 71K transactions.

`TAFENG`[6] is a Kaggle Chinese store transactions data from Nov. 2000 to Feb. 2001. The data contains 10 customer age groups.

[5]https://bit.ly/3ruMEgz
[6]https://www.kaggle.com/chiranjivdas09/ta-feng-grocery-dataset

Data set	Description	Pivots	Segmentations
RETAIL	Period: [2017/2/28-2019/12/30] 250,208 transactions	Promotion: [2018/12/17-2019/2/28] Demographics: location, gender	PWindow: 200, 500, 1K, 2K, 5K, 10K Demographics: location, gender
TAFENG	Period: [2000/11/01-2001/02/28] 817,741 transactions	Promotion: Chinese New Year [2001/1/24] Demographics: age	PWindow: 10K Demographics: age
SALES	Period: [2010/02/05-2012/11/01] 421,570 transactions	Promotion: `isHoliday` weeks	PWindow: 500,2k, 3k, 10k
AMAZON	Period: [2015/01/01-2018/01/01] 747,804 reviews	Promotion: Christmas [2016/11/20-2017/01/01]	PWindow: 500, 1K, 2K, 5K, 10K

TABLE II: Data sets, pivots, and segmentations used in the experiments

SALES[7] is a Kaggle sales data from 45 stores in different departments (anonymized). It contains 421,570 transactions from Feb. 2010 to Nov. 2012. The data contains weekly sales and a special attribute `isHoliday` indicating whether a week is a holiday week. There are 29,661 holiday weeks and 391,909 non-holiday weeks.

AMAZON[8] product data contains 747,804 digital music reviews generated by 425,671 customers for 261,950 tracks.

Pivots, Segmentations, and Tests. To simplify exposition, we focus on hypothesis testing of the form $(H_0, H_A, \texttt{Mean}, =)$ for the one-sample and two-sample questions in Figure 4. Experiments with other aggregations, comparisons, and multiple-sample questions are left for future work. Table II summarizes the pivots and segmentations we used in our examples.

Parameters. To measure performance, we examine the number of segments found, their coverage of the input data set, the rate at which risk capital is consumed, and response time. We vary: (1) the pivot and type of segmentation; (2) the statistical test referred to as $\#\texttt{Q}_\texttt{i}$ in Figure 2 and the control method (i.e., Bonferroni correction [10] or Benjamini-Yekutieli); (3) the significance level α; (4) the size of the segmentation windows (measured as a number of transactions they contain); (5) the risk capital κ; and for the scalability experiment, (6) the size of the input data.

Setup. Experiments were executed on a Linux computer with an Intel® Core™ i7-8650U CPU @ 1.90GHz × 8, and 16GB memory.

B. Segment Search and Significance Adjustment

This first experiment aims to demonstrate the expressivity of ***S4***. Table III summarizes our results on **RETAIL** for the examples in Figure 3. In this experiment, we do not constrain risk capital to examine the whole range of p-values. We omit `#Q2`, `#Q5` and `#Q8` (no segmentation).

Our first observation is that ***S4*** is applicable to a wide range of scenarios involving one- and two-sample tests with various data partitionings and α values (.05 or .01). The second observation is that ***S4*** (with *nextCandidate()* based on p-value scan) is able to identify more hypotheses than Bonferroni [10] for the same α. Recall that the Bonferroni correction is sensitive to hypothesis dependence and is hence prone to errors. The third observation is that ***S4*** with a *nextCandidate()* based on coverage scan has the fewest number of results. The fourth observation is that due to the range of p-values (last two columns), setting a risk capital is a difficult task.

[7]https://www.kaggle.com/manjeetsingh/retaildataset?select=sales+data-set.csv

[8]https://nijianmo.github.io/amazon/index.html

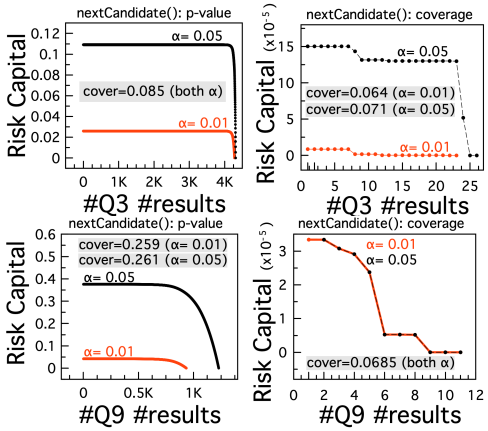


Fig. 5: Risk capital consumption for two α values for `#Q3` and `#Q9` with PWindow[500] on **RETAIL**

We also observe that increasing the number of transactions in point-wise segmentation, as shown in `#Q3` and `#Q9`, generally decreases the number of candidate segments, and *de facto* the number of retrieved segments that satisfy the test. Finally, reducing α from .05 to .01 decreases the number of retrieved segments for `#Q3` PWindow[500] and [1K] and `#Q9` PWindow[500], [1K], and [2K] (the other queries have identical results). These cases show that the type of point-wise segmentation and the value of α should be carefully and jointly chosen depending on the priorities of the application to favor either the number of results or the test significance.

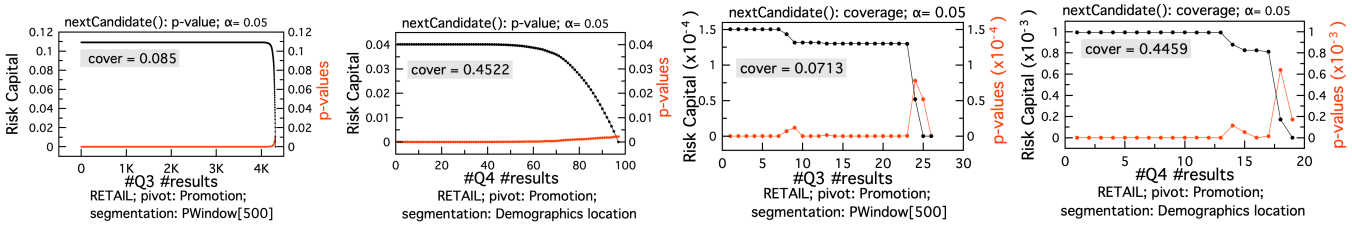
C. Significance, Risk, and Coverage

We now propose a deeper dive to examine (i) the relationship between setting a significance level and risk consumption, and (ii) the impact of various data partitionings on coverage.

Figure 5 shows risk consumption (Y-axis) as a function of the number of results (X-axis) for two values of α, 0.01 and 0.05. We choose `#Q3` and `#Q9` with PWindow[500] on **RETAIL** as they have a high number of results. Not surprisingly, a lower value for α reduces the final number of retrieved segments that satisfy the hypothesis. After a plateau due to very low p-values, risk consumption decreases steeply which means that the cumulated risk increases with the number of tested p-values. One can clearly see that using a significance threshold is not enough for multiple hypothesis

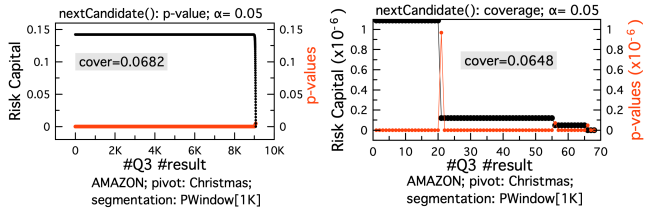
α	$\#Q_i$	**Pivot**	**Segmentation**	**#Candidates**	**#Results** ***S4*** *p-value scan*	**#Results** ***S4*** *coverage scan*	**#Results Bonferroni**	min **p-value**	Σ p-values ***S4*** (p-value)
.05	#Q1	None	PWindow[2K]	74	74	11	10	0	2.14E-276
	#Q3	Prom	PWindow[500]	16 058	4313	26	4122	1.38E-80	0.109
	#Q3	Prom	PWindow[1K]	4033	497	10	493	2.19E-107	0.006
	#Q3	Prom	PWindow[2K]	867	0	0	0	0	0
	#Q4	Prom	Dem[loc]	2 622	97	19	56	1.44E-141	0.040
	#Q4	Prom	Dem[loc&gen]	6461	112	26	60	1.21E-121	0.093
	#Q6	None	Dem[loc]	59	48	48	48	0	4.36E-08
	#Q7	Dem[gen]	Dem[loc&gen]	2 124	16	00	10	4.25E-10	0.002
	#Q7	Dem[gen]	Dem[loc]	2 124	16	8	10	4.25E-10	0.002
	#Q9	Dem[gen]	PWindow[500]	9 360	1229	11	630	2.92E-23	0.37
	#Q9	Dem[gen]	PWindow[1K]	2340	442	10	306	3.84E-29	0.110
	#Q9	Dem[gen]	PWindow[5K]	104	38	4	35	3.65E-27	0.013
	#Q9	Dem[gen]	PWindow[10K]	26	19	2	18	7.85E-39	0.005
.01	#Q3	Prom	PWindow[500]	16 058	4 280	23	4 058	1.38E-80	0.025
	#Q3	Prom	PWindow[1K]	4 033	495	10	492	2.19E-107	0.001
	#Q9	Dem[gen]	PWindow[500]	9 360	935	11	492	2.92E-23	0.041
	#Q9	Dem[gen]	PWindow[1K]	2 340	383	9	270	3.84E-29	0.021
	#Q9	Dem[gen]	PWindow[2K]	585	109	7	78	1.43E-22	0.009

TABLE III: Results on **RETAIL** with two control methods: Bonferroni and Benjamini-Yekutieli (***S4***)



(a) PWindow[500] (b) Demographics (c) PWindow[500] (d) Demographics

Fig. 6: Evolution of risk capital and p-value distribution on **RETAIL** with Prom as pivot and 2 segmentations: Demographics (location), PWindow[500] and *nextCandidate()* based on p-value scan (a-b) and on coverage scan (c-d)



(a) p-value scan (b) coverage scan

Fig. 7: Risk capital consumption and p-value distribution on **AMAZON** with Christmas pivot and 2 segmentations: PWindow[1K] with p-value scan (a) and coverage scan (b); Other PWindow[2K], [5K] or [10K] returned too few results.

testing, and that capping risk is crucial to control significance. An important observation is that **coverage scan returns 2 orders of magnitude fewer results than p-value scan**. Despite that, coverage of the input data set is at least as good. This clearly justifies the benefit of returning fewer hypotheses while maintaining and sometimes improving their coverage of the input data set.

We now examine the dependence of coverage risk consumption on data partitioning and on *nextCandidate()*. We run $\#Q_3$ and $\#Q_4$ on **RETAIL** and $\#Q_3$ on **AMAZON**. Figures 6a and 6b (resp., 6c and 6d) report risk consumption and p-value distribution, for different pivots and segmentation types, with p-value scan (resp., coverage scan). Figures 7a and 7b reports similar results on **AMAZON**. The figures clearly show that different segmentations attain different coverage values, and that **coverage scan returns 1 to 3 orders of magnitude fewer results than p-value scan while maintaining their coverage of the input data.** One can also see that different segmentations consume risk at different rates. Risk is consumed more quickly with coverage scan. That is particularly true for demographics-based segmentations where the size of segments is variable (Figures 6a and 6d).

D. Addressing Information Needs

We describe two scenarios that leverage our data sets to demonstrate how to compose segment search into pipelines to address information needs. We only present results of p-value scan since coverage scan showed similar output.

Retargeting. We start with a two-step example on **RETAIL**. The first step runs **a one-sample search** that identifies highly

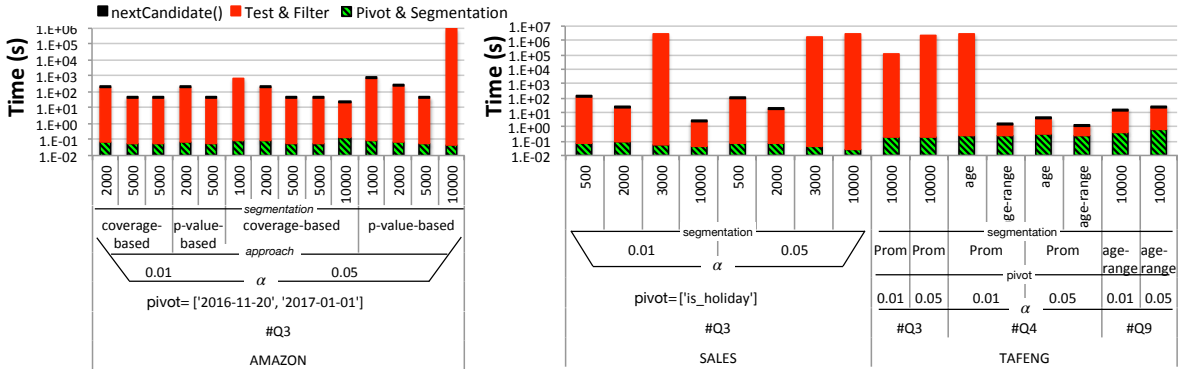


Fig. 8: Breakdown of response times of selected segment searches on our data sets

active customers (H_0 is *"average daily purchases = 10"* with **pivot = NONE** and **segmentation on location**. The alternative hypothesis is *"average daily purchases ¿ 10"*. The number of candidate segments is 61 (locations) that are fed to **a one-sample test** that returns 48 qualifying segments containing 23,392 customers. In the second step, we consider all resulting customers and run **a two-sample test with a Prom pivot and a (location, gender) segmentation**, where H_0 is *"average daily purchases before Prom = after"* and H_a is *"before Prom ¡ after"*. The total number of candidate segments is 2 (gender) $\times$ 48 (qualifying segments from the previous step). This test returns 37 segments that contain 25,216 customers who constitute good candidates for retargeting, a strategy that focuses on advertising to customers who are already familiar with a brand and have recently demonstrated interest.[9]

Risk Aversion. Our second scenario is a two-step example on **TAFENG** that seeks to verify a longstanding hypothesis in Psychology according to which young customers are less prone to risk aversion, i.e., the feeling of missing out if they do not buy a product at a reduced price [16]. Our first step is to find segments whose purchase average remained the same after a **Prom (pivot)** using a **window-based segmentation** (PWindow[5K]). This segmentation returns 14 qualifying segments (containing 8,634 out of 32,266 customers), the highest number among all segmentations. For each output segment, we explore its demographics with **an age-based segmentation and no pivot** using a **one-sample test** to verify which ones have weekly purchases equal to 150. The test returns customers aged [35-39], [45-49], [50-54], and [60-64]. This confirms the original hypothesis.

E. Scalability Analysis

All scalability measures are averaged over 5 runs. Figure 8 reports a breakdown of the time to evaluate #Q3, #Q4, and #Q9 on different data sets, pivots and segmentations (α = .05 and default is p-value scan). Most of the time is spent in the p-value computation protocol. Overall response time is reasonable. For **AMAZON**, p-value scan is more expensive than coverage scan.

[9]https://en.wikipedia.org/wiki/Behavioral_retargeting

We now study response time as a function of data size (#Q3 PWindow[500] on **AMAZON** in Figure 9). The original data was replicated to generate increasing data sizes. The graph shows that coverage scan outperforms p-value scan (with 1M candidates, coverage scan took only 1.84 seconds while p-value scan did not terminate after 90 minutes). This makes coverage scan the method of choice for composing segment searches into advanced pipelines and for interactive multiple hypothesis testing.

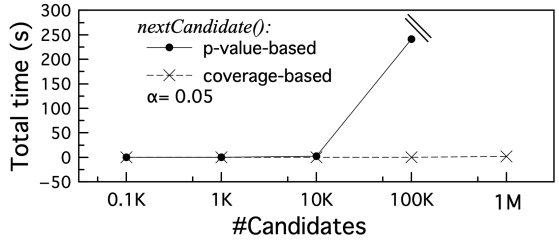


Fig. 9: Scalability on **AMAZON** for #Q3 PWindow[500]

F. Discussion

Due to limited space, and to build a coherent narrative, we intentionally only covered mean, and equality in our experiments. An immediate empirical investigation would be to run all other variants. We now discuss new research questions raised by our work.

On Data Partitioning. While our conclusions are not limited to a single data set, the same segment search may yield contradicting results on different data sets, by rejecting or not the null hypothesis. That could happen even if the data sets are comparable in content and size and the segment search uses the same partitioning. A natural question is how to account for evidence found in different data sets? Existing work on multiple test correction in pattern mining (e.g., LAMP: Limitless Arity Multiple-testing Procedure) [6] could be leveraged with data partitioning and additional pivots.[10]

On Filtering and Composing Hypotheses. To reduce the number of hypotheses, we adopted coverage-based fil-

[10]Promal seasons at https://global-uploads.webflow.com/5fb2b92b11680d2e5ffaa1fa/5fb2b92b11680d59f9faa88f_Seasonal-Promal-Ideas-1024x1024.png

tering [17]. Another promising filtering method would be to revisit segmentation. For instance, demographics-based segmentation can use a partition decision tree where the gain function leverages α to prune unpromising segments [18]. Another possibility is to leverage drift detection [19], and use α to determine window length in window-based segmentation.

As shown in our experiments, segment search can be composed to address complex information needs where a pipeline of hypotheses are tested. This raises new challenges on combining statistical tests in a pipeline. In particular, one may allow the decision of whether to reject a null hypothesis to be reconsidered in light of null hypotheses rejected in subsequent steps. This needs to be considered in conjunction with setting a risk capital.

VI. Conclusion

We developed ***S4***, the first framework that combines expressive data partitioning with powerful statistical testing, to verify common hypotheses in retail and find relevant data segments. To the best of our knowledge, ours is the first extensive experiments of data partitioning and hypothesis testing. Our work also laid the ground for many new research questions discussed in Section V-F.

Our framework is aimed as a first step toward developing a benchmark for combining powerful data partitioning and hypothesis testing. The grid presented in Figure 3 (along with our available code and set of segment search queries) can serve as a basis to define other specific questions that are relevant to a particular domain of interest. Such a benchmark would contribute to building a community around the topic and encouraging experimental repeatability and result reproducibility.

References

[1] M. Jafari and N. Ansari-Pour, "Why, when and how to adjust your p values?" *Cell Journal (Yakhteh)*, vol. 20, no. 4, p. 604, 2019.

[2] R. J. Meijer and J. J. Goeman, "Multiple testing of gene sets from gene ontology: Possibilities and pitfalls," *Briefings Bioinform.*, vol. 17, no. 5, pp. 808–818, 2016.

[3] L. Pellegrina, M. Riondato, and F. Vandin, "Hypothesis testing and statistically-sound pattern mining," in *Proc. of the 25th ACM SIGKDD Intl. Conf. on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, 2019, pp. 3215–3216.

[4] W. Hämäläinen and G. I. Webb, "A tutorial on statistically sound pattern discovery," *Data Min. Knowl. Discov.*, vol. 33, no. 2, pp. 325–377, 2019.

[5] G. I. Webb, "Discovering significant patterns," *Mach. Learn.*, vol. 68, no. 1, pp. 1–33, 2007.

[6] J. Komiyama, M. Ishihata, H. Arimura, T. Nishibayashi, and S. Minato, "Statistical emerging pattern mining with multiple testing correction," in *Proc. of the 23rd ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, 2017, pp. 897–906.

[7] P. Vorburger and A. Bernstein, "Entropy-based concept shift detection," in *Proc. of the 6th IEEE Intl. Conf. on Data Mining (ICDM 2006), 18-22 December 2006, Hong Kong, China*. IEEE Computer Society, 2006, pp. 1113–1118.

[8] E. Roquain, "Type i error rate control for testing many hypotheses: a survey with proofs," 2011.

[9] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Annals of Statistics*, vol. 29, pp. 1165–1188, 2001.

[10] G. Beliakov, S. James, J. Mordelová, T. Rückschlossová, and R. R. Yager, "Generalized bonferroni mean operators in multi-criteria aggregation," *Fuzzy Sets Syst.*, vol. 161, no. 17, pp. 2227–2242, 2010.

[11] S. Greenland, S. J. Senn, K. J. Rothman, J. B. Carlin, C. Poole, S. N. Goodman, and D. G. Altman, "Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations," *European Journal of Epidemiology*, vol. 31, no. 4, pp. 337–350, 2016. [Online]. Available: http://dx.doi.org/10.1007/s10654-016-0149-3

[12] G. I. Webb and F. Petitjean, "A multiple test correction for streams and cascades of statistical hypothesis tests," in *Proc. of the 22nd ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, San Francisco, USA, Aug. 2016*, 2016, pp. 1255–1264.

[13] C. Chekuri and S. Khanna, "A polynomial time approximation scheme for the multiple knapsack problem," *SIAM J. COMPUT*, vol. 38, no. 3, p. 1, 2006.

[14] Y. Benjamini and Y. Hochberg, *Journal of the Royal Statistical Society Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[15] D. Colquhoun, "An investigation of the false discovery rate and the misinterpretation of p-values," *Royal Society Open Science*, vol. 1, no. 3, p. 140216, 2014.

[16] H. Peng, S. Xia, F. Ruan, and B. Pu, "Age differences in consumer decision making under option framing: From the motivation perspective," *Frontiers in Psychology*, vol. 7, p. 1736, 2016.

[17] R. Bourgon, R. Gentleman, and W. Huber, "Independent filtering increases detection power for high-throughput experiments," *Proc. of the National Academy of Sciences*, vol. 107, no. 21, pp. 9546–9551, 2010.

[18] S. Amer-Yahia, S. Kleisarchaki, N. K. Kolloju, L. V. S. Lakshmanan, and R. H. Zamar, "Exploring rated datasets with rating maps," in *Proc. of the 26th Intl. Conf. on World Wide Web, Australia, April 3-7, 2017*, 2017, pp. 1411–1419.

[19] S. Kleisarchaki, S. Amer-Yahia, A. D. Chouakria, and V. Christophides, "Querying temporal drifts at multiple granularities," in *Proc. of the 24th ACM Intl. Conf. on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*, 2015, pp. 1531–1540.