Efficient Optimization of Partition Scan Statistics via the Consecutive Partitions Property

Charles A. Pehlivanian
pehlivaniancharles@gmail.com
JP Morgan Chase & Co.
and
Daniel B. Neill*
daniel.neill@nyu.edu
New York University

Abstract

We generalize the spatial and subset scan statistics from the single to the multiple subset case. The two main approaches to defining the log-likelihood ratio statistic in the single subset case – the population-based and expectation-based scan statistics – are considered, leading to risk partitioning and multiple cluster detection scan statistics, respectively. We show that, for distributions in a separable exponential family, the risk partitioning scan statistic can be expressed as a scaled f-divergence of the normalized count and baseline vectors, and the multiple cluster detection scan statistic as a sum of scaled Bregman divergences. In either case, however, maximization of the scan statistic by exhaustive search over all partitionings of the data requires exponential time.

To make this optimization computationally feasible, we prove sufficient conditions under which the optimal partitioning is guaranteed to be consecutive. This Consecutive Partitions Property generalizes the linear-time subset scanning property from two partitions (the detected subset and the remaining data elements) to the multiple partition case. While the number of consecutive partitionings of n elements into t partitions scales as $O(n^{t-1})$, making it computationally expensive for large t, we present a dynamic programming approach which identifies the optimal consecutive partitioning in $O(n^2 t)$ time, thus allowing for the exact and efficient solution of large-scale risk partitioning and multiple cluster detection problems. Finally, we demonstrate the detection performance and practical utility of partition scan statistics using simulated and real-world data.

Keywords: partitioning, cluster detection, subset scan, dynamic programming

*DBN gratefully acknowledges funding support from the National Science Foundation Program on Fairness in Artificial Intelligence in Collaboration with Amazon, grant IIS-2040898.

1 Introduction

The *spatial scan statistic* is a widely used methodological approach for spatial and space-time cluster detection, first proposed by Kulldorff (Kulldorff, 1997; Kulldorff and Nagarwalla, 1995) and building on prior work in scan statistics by Naus, Glaz, and others (Naus, 1965a,b; Glaz et al., 2001). Spatial scanning has been used for detection of high-risk clusters for cancer and other chronic diseases, emerging infections in human and animal populations, suspicious network activity, areas of increased brain activity from imaging data, and many other applications (Kulldorff, 1997; Neill, 2012). In the usual spatial scan setting, data elements s_i, for $i \in \{1, \ldots, n\}$, represent spatial locations with associated values x_i (representing counts or concurrent measurements) and y_i (representing baselines, expectations, or populations). The goal is to identify a subset of locations with unusual (typically, increased) values of x_i as compared to the expected y_i. This is done by maximizing a log-likelihood ratio statistic $F(S)$ over subsets $S \subseteq \{s_1, \ldots, s_n\}$, assuming a parametric model for the x_i. For example, for Kulldorff's spatial scan statistic (Kulldorff, 1997), we assume $x_i \sim \text{Poisson}(q_i y_i)$ for all s_i. Under the null hypothesis H_0, $q_i = q_{all}$ everywhere for some constant q_{all}, while under the alternative hypothesis $H_1(S)$, $q_i = q_{in}$ for $s_i \in S$ and $q_i = q_{out}$ for $s_i \notin S$, for constants $q_{in} > q_{out}$. The values of q_{in}, q_{out}, and q_{all} are fit by maximum likelihood, leading to the generalized log-likelihood ratio statistic:

$$\begin{aligned} F(S) &= \log \frac{\max_{q_{in} > q_{out}} \Pr(\text{Data} \mid H_1(S))}{\max_{q_{all}} \Pr(\text{Data} \mid H_0)} \\ &= \log \frac{\max_{q_{in} > q_{out}} \prod_{s_i \in S} \Pr(x_i \sim \text{Poisson}(q_{in} y_i)) \prod_{s_i \notin S} \Pr(x_i \sim \text{Poisson}(q_{out} y_i))}{\max_{q_{all}} \prod_{s_i} \Pr(x_i \sim \text{Poisson}(q_{all} y_i))} \\ &= \left(C_{in} \log \frac{C_{in}}{B_{in}} + C_{out} \log \frac{C_{out}}{B_{out}} - C_{all} \log \frac{C_{all}}{B_{all}} \right) \mathbb{1} \left\{ \frac{C_{in}}{B_{in}} > \frac{C_{out}}{B_{out}} \right\}, \end{aligned}$$

where $C_{in} = \sum_{s_i \in S} x_i$, $C_{out} = \sum_{s_i \notin S} x_i$, and $C_{all} = \sum_{s_i} x_i$, and similarly $B_{in} = \sum_{s_i \in S} y_i$, $B_{out} = \sum_{s_i \notin S} y_i$, and $B_{all} = \sum_{s_i} y_i$. Often, baselines are computed such that $B_{all} = C_{all}$, and thus $C_{all} \log \frac{C_{all}}{B_{all}} = 0$; in any case, this term is the same for all S and can be neglected when computing $S^* = \arg\max_S F(S)$. Once the highest-scoring regions are found, a randomization test is used to determine whether they are statistically significant at a given level (Kulldorff, 1997), or in some simple cases a limiting distribution is known. Alternatively, baselines can be updated under a Bayesian framework, and significance can be estimated from the posterior (Neill et al., 2006).

While Kulldorff's original approach focused on maximizing $F(S)$ over the set of circular spatial regions, which can be done exhaustively in $\mathcal{O}(n^2)$ time, Neill (2012) reframed the problem as a (constrained or unconstrained) search over all subsets S, thus allowing detection of irregularly-shaped clusters as well as extensions to non-spatial data. While exhaustive enumeration and evaluation of the 2^n subsets of data elements is computationally infeasible, many log-likelihood ratio scan statistics (including the Kulldorff statistic defined above) were shown to satisfy the *linear-time subset scanning* property, enabling exact optimization over the exponentially many subsets while requiring only a linear number of subsets to be evaluated (Neill, 2012).

The spatial and subset scans can be thought of as partitioning the data elements s_i into two disjoint subsets: a "high-risk" subset S and a "low-risk" subset $\{s_1, \ldots, s_n\} \setminus S$, assuming constant relative risk q within each subset. Real-world datasets, however, may not satisfy this strong assumption, and may be better modeled by identifying multiple disjoint subsets $S_1, S_2, \ldots, S_t$ of varying risk. This finer-grained partitioning of the data is useful for detection of multiple clusters (Li et al., 2011), for risk stratification in disease mapping (Lawson et al., 1999), and for discretization of high-arity attributes such as occupations (Kulldorff et al., 2003).

In this work, we demonstrate that a large class of log-likelihood ratio statistics, including the Poisson scan statistic defined above, can be exactly and efficiently optimized over *partitionings* of the data, requiring worst-case time $\mathcal{O}(n^2 t)$ to divide a set of size n into t partitions. This efficient optimization follows from two sets of results. First, we define *partition score functions* and show that a large class of partition score functions satisfy the *consecutive partitions property*, a generalization of linear-time subset scanning from $t = 2$ to $t \geq 2$ partitions. Second, while naive optimization of the score function over consecutive partitionings would be computationally infeasible for large t, with time complexity $\mathcal{O}(n^{t-1})$, we provide a dynamic programming algorithm which reduces the time cost to quadratic in n and linear in t, while still guaranteeing that the optimal partitioning will be found.

2 Partition scan statistics

Let $n \in \mathbb{N}$ be positive and set $\mathcal{V} = \{1, \ldots, n\}$. Let $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_n\}$ be finite real sequences, with $y_i > 0$ for all i. In many of the applications that follow, each tuple

(x_i, y_i) corresponds to realized and baseline counts or measures of an attribute for the item or spatial location s_i. Denote by $\mathcal{D} = \mathcal{D}_{X,Y}$ the set of tuples $\{(x_i, y_i)\}$ associated with X and Y.

Given the dataset $\mathcal{D}$ and a target number of partitions t, our goal is to maximize the *partition scan statistic* $F(P) = F(\{S_1, \ldots, S_t\})$ over all possible partitionings P of size t, where each subset $S_j \subseteq \{1, \ldots, n\}$ is non-empty, $S_j \cap S_{j'} = \emptyset$ for $j \neq j'$, and $\bigcup_{j=1\ldots t} S_j = \mathcal{V}$. The partition score $F(P)$ is defined in terms of a *score function* $f(x, y)$ as:

$$F(P) = \sum_{j=1\ldots t} F(S_j) = \sum_{j=1\ldots t} f\left(\sum_{i\in S_j} x_i, \sum_{i\in S_j} y_i\right).$$

As a concrete example, the Kulldorff scan statistic defined above is a partition scan statistic for size $t = 2$, with associated score function $f(x, y) = x \log \frac{x}{y}$, and the extension to Poisson partition scan statistics of size $t > 2$ follows naturally from the above:

$$F(P) = \sum_{j=1\ldots t} C_j \log \frac{C_j}{B_j},$$

where $C_j = \sum_{i\in S_j} x_i$ and $B_j = \sum_{i\in S_j} y_i$ are the aggregate count and baseline for partition S_j.

We will also refer to the *rational score functions* $f(x, y) = x^\alpha y^{-\beta}$ for constants $\alpha, \beta > 0$, and their corresponding partition scan statistics $F(P)$. More formally, we define a score function as follows:

Definition 1. *A score function is a continuous function* $f(x, y) \colon \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}$ *with continuous extension to the origin in any wedge* $\mathcal{W}(\mu_1, \mu_2) = \left\{(x, y) : y > 0, \mu_1 \leq \frac{x}{y} \leq \mu_2\right\}$*, for* $-\infty < \mu_1 \leq \mu_2 < \infty$*, with* $\lim_{(x,y)\in\mathcal{W}\to(0,0)} f(x, y) = 0$.

The regularity condition on $\mathcal{W}$ in wedges simply guarantees a continuous extension to the origin on any positive cone in $\mathbf{R}^+$; for rational score functions it is equivalent to the constraint $\alpha > \beta$. Several additional properties have been associated with score functions, namely monotonicity in x or y, quasi-convexity, convexity, or an isotone differences constraint $\frac{\partial^2 f}{\partial x \partial y} \leq 0$ related to submodularity. We do not assume smoothness beyond continuity, nor (quasi-)convexity, etc., unless explicitly stated.

A score function f naturally gives rise to a set function $F \colon 2^N \to \mathbb{R}$, as above, defined by summation over subsets, $F(S) = f(\sum_{i\in S} x_i, \sum_{i\in S} y_i)$, for $S \subseteq \mathcal{V}$. Associating $S \subseteq \mathcal{V}$ with its X,

Y attributes as in $X_S = \sum_{i \in S} x_i, Y_S = \sum_{i \in S} y_i$ allows us to write $F(S) = f(\sum_{x \in X_S} x, \sum_{y \in Y_S} y) = f(X_S, Y_S)$, etc. Likewise, the partition scan statistic $F(P) = F(\{S_1, \ldots, S_t\})$ can be written as $\sum_{j=1 \ldots t} F(S_j)$, as above, and we are interested in identifying the optimal partitioning

$$P^* = \underset{P=\{S_1,\ldots,S_t\}}{\operatorname{argmax}} \sum_{j=1\ldots t} F(S_j) = \underset{P=\{S_1,\ldots,S_t\}}{\operatorname{argmax}} \sum_{j=1\ldots t} f\left(\sum_{i \in S_j} x_i, \sum_{i \in S_j} y_i\right). \tag{1}$$

Common approaches to solving equation (1) rely on properties of the set function F. For F submodular, a class of greedy algorithms can provide $(1 - \frac{1}{e})$-approximate solutions in polynomial time (Filmus and Ward, 2012; Vondrák, 2008), based on Monte Carlo approximation of the multilinear extension $f_m(x_1, \ldots, x_n) = \sum_{S \subseteq \mathcal{V}} F(S) \prod_{i \in S} x_i \prod_{i \in \mathcal{V} \setminus S} (1 - x_i)$, for $x \in [0,1]^n$. Minimization of a submodular F over subsets of $2^{\mathcal{V}}$ admits an exact solution using the Lovasz extension (Grötschel et al., 1981), which is guaranteed to be convex. The essence of these approaches is a "relaxation by expectation" to a continuous problem on the n-dimensional unit hypercube, associating $S \subseteq \mathcal{V}$ with the point $(\mathbb{1}_{\{1 \in S\}}(S), \ldots, \mathbb{1}_{\{n \in S\}}(S))$. In this way every boundary point of the hypercube is associated with a subset of $\mathcal{V}$. In our setting, the ambient extension is already specified as f, and the natural extensions to $\mathbf{R}^n$ (multilinear, Lovasz) do not allow for a convenient restriction of the admissible set of partitionings. Nevertheless, we will demonstrate an alternative approach based on consecutive partitions which enables exact and efficient optimization of equation (1), as described in §3 below.

2.1 Partition scan statistics for the separable exponential family

Let us assume a parametric specification for the generating distribution Ψ for the x_i values, where the null and alternative hypotheses are defined as follows:

$$H_0 \colon x_i \sim \Psi(\cdot; \mu = q_{all} \mu_i) \; \forall s_i,$$
$$H_1(P) \colon x_i \sim \Psi(\cdot; \mu = q_j \mu_i) \; \forall s_i \in S_j,$$

where $P = \{S_1, \ldots, S_t\}$ is a partitioning of $\mathcal{D}$ of size t, and the constants $q_1 \ldots q_t$ (the "relative risk" for each partition $S_j \in P$) and q_{all} are computed by maximum likelihood. The most common choices of distribution are Poisson and Gaussian. For many distributional assumptions, the score function is expressed only in terms of sufficient statistics of the underlying Ψ, and closed-form

solutions for the q_j can be solved for and re-substituted. Let us consider the (generalized) log-likelihood ratio statistic:

$$F(P) = F(\{S_1, \ldots, S_t\}) = \log \frac{\max_{q_1,\ldots,q_t} \prod_{j=1\ldots t} \prod_{s_i \in S_j} \Pr(x_i \mid x_i \sim \psi(\cdot; \mu = q_j \mu_i))}{\max_{q_{all}} \prod_{s_i} \Pr(x_i \mid x_i \sim \psi(\cdot; \mu = q_{all} \mu_i))}. \quad (2)$$

Following Neill (2012), we now assume that Ψ comes from an exponential family, in which case we can write Ψ in terms of its mean μ as $\log \Pr(x|\mu) = T(x)\theta(\mu) - \psi(\theta(\mu)) = T(x)\theta(\mu) - \mu\theta(\mu) + \phi(\mu)$, where $T(x)$ is the sufficient statistic, $\theta(\mu)$ is a function mapping the mean μ to the natural parameter θ, ψ is the log-partition function, and ϕ is the convex conjugate (in the Legendre-Fenchel sense) of ψ. Plugging in the log-likelihood for the exponential family into the expression for $F(P)$ above, we obtain

$$F(P) = \sum_{j=1\ldots t} \sum_{s_i \in S_j} (T(x_i)\theta(q_j\mu_i) - q_j\mu_i\theta(q_j\mu_i) + \phi(q_j\mu_i)) \\ - \sum_{s_i} (T(x_i)\theta(q_{all}\mu_i) - q_{all}\mu_i\theta(q_{all}\mu_i) + \phi(q_{all}\mu_i)).$$

For distributions Ψ in a *separable exponential family* (Neill, 2012), including the Poisson, Gaussian, and exponential distributions, we can further write $\theta(q\mu_i) = z_i\theta_0(q) + v_i$, where the function θ_0 depends only on q, while z_i and v_i can depend on μ_i and σ_i but are independent of q. We also have $\phi(q\mu_i) = \mu_i z_i \phi_0(q) + \mu_i v_i q + K_i$, where $\phi_0(q) = \int \theta_0(q)\,\mathrm{d}q$, and K_i is independent of q. The expression for $F(P)$ can then be simplified to

$$F(P) = \sum_{j=1\ldots t} (C_j\theta_0(q_j) + B_j(\phi_0(q_j) - q_j\theta_0(q_j))) \\ - C_{all}\theta_0(q_{all}) - B_{all}(\phi_0(q_{all}) - q_{all}\theta_0(q_{all})),$$

where C_j and B_j are the sufficient statistics for subset S_j, $C_j = \sum_{s_i \in S_j} T(x_i) z_i$ and $B_j = \sum_{s_i \in S_j} \mu_i z_i$ respectively. Similarly, we have $C_{all} = \sum_{s_i} T(x_i) z_i$ and $B_{all} = \sum_{s_i} \mu_i z_i$. To obtain the maximum likelihood estimate of q_j, we set $\frac{\partial F}{\partial q_j} = 0$, obtaining $q_j = \frac{C_j}{B_j}$, and similarly $q_{all} = \frac{C_{all}}{B_{all}}$. Substituting these values of q_j and q_{all} into the equation for $F(P)$ and simplifying, we find that

$$F(P) = \sum_{j=1\ldots t} B_j \left(\phi_0 \left(\frac{C_j}{B_j} \right) - \phi_0 \left(\frac{C_{all}}{B_{all}} \right) \right) = B_{all} D_f \left(\vec{C} \parallel \vec{B} \right),$$

where $\vec{C}$ and $\vec{B}$ are the normalized count and baseline vectors $\frac{1}{C_{all}}\langle C_1, \ldots, C_t\rangle$ and $\frac{1}{B_{all}}\langle B_1, \ldots, B_t\rangle$ respectively, and D_f is the f-divergence, $D_f(P \,\|\, Q) = \sum_{j=1\ldots t} Q_j f\left(\frac{P_j}{Q_j}\right)$, with $f(q) = \phi_0\left(q\frac{C_{all}}{B_{all}}\right) - \phi_0\left(\frac{C_{all}}{B_{all}}\right)$.[1] $F(P)$ can also be written in terms of the sufficient statistics C_j and B_j, and the score function $f(x, y) = y\phi_0\left(\frac{x}{y}\right)$, as $F(P) = \left(\sum_{j=1\ldots t} f(C_j, B_j)\right) - f(C_{all}, B_{all})$, where the last term is independent of P.

As a concrete example, for the Gaussian distribution we have $\phi_0(q) = \frac{q^2}{2}$, and the corresponding $f(q) = \frac{C_{all}^2}{2B_{all}^2}(q^2 - 1)$. This gives us $F(P) = \left(\sum_{j=1\ldots t} \frac{C_j^2}{2B_j}\right) - \frac{C_{all}^2}{2B_{all}}$, or equivalently, $F(P) = \left(\sum_{j=1\ldots t} f(C_j, B_j)\right) - f(C_{all}, B_{all})$, where $f(x, y) = \frac{x^2}{2y}$. Moreover, we note that $T(x_i) = x_i$ and $z_i = \frac{\mu_i}{\sigma_i^2}$ for the Gaussian, so the sufficient statistics are $C_j = \sum_{s_i \in S_j} \frac{x_i \mu_i}{\sigma_i^2}$ and $B_j = \sum_{s_i \in S_j} \frac{\mu_i^2}{\sigma_i^2}$ respectively.

Similarly, for the Poisson distribution (i.e., the generalization of Kulldorff's spatial scan statistic to $t \geq 2$ partitions), we have $\phi_0(q) = q \log q$, and the corresponding $F(P) = \left(\sum_{j=1\ldots t} f(C_j, B_j)\right) - f(C_{all}, B_{all})$, where $f(x, y) = x \log \frac{x}{y}$. In this case, owing to the form of ϕ_0, the partition scan statistic can also be written as $F(P) = C_{all} D_{KL}(\vec{C}, \vec{B})$, where D_{KL} is the Kullback-Liebler divergence between the normalized count and baseline vectors. Since $T(x_i) = x_i$ and $z_i = 1$ for the Poisson, the sufficient statistics are $C_j = \sum_{s_i \in S_j} x_i$ and $B_j = \sum_{s_i \in S_j} \mu_i$ respectively.

3 The Consecutive Partitions Property

We now return to the general formulation of the partition scan statistic, $F(P) = F(\{S_1, \ldots, S_t\}) = \sum_{j=1\ldots t} f(\sum_{s_i \in S_j} x_i, \sum_{s_i \in S_j} y_i)$, and identify conditions on the score function f that will enable us to efficiently compute the optimal partitioning $P^* = \arg\max_P F(P)$.

We first note that exhaustively searching over all partitionings P is computationally infeasible, even for relatively small values of n and t. More precisely, the number of such partitionings $\left\{{n \atop t}\right\}$ is a Stirling number of the second kind. The Stirling numbers have an asymptotic growth rate which is exponential in n for fixed t, $\left\{{n \atop t}\right\} \sim \frac{t^n}{t!}$. For example, for $n = 30$ and $t = 10$, we already have $\left\{{n \atop t}\right\} \approx 1.73 \times 10^{22}$.

[1] In the common case in which B_{all} is set equal to C_{all} by construction, $F(P)$ simplifies to $B_{all} D_{\phi_0}(\vec{C} \,\|\, \vec{B})$. This is similar to the expectation-based scan statistics for the separable exponential family (Neill, 2012), $F(S) = BD_{\phi_0}(C/B, 1)$, but in the latter case D is the Bregman divergence rather than f-divergence corresponding to ϕ_0.

Moreover, it is well known that even for monotone submodular set functions F, the maximization $\max_{S \in \mathcal{V}} F(S)$ is NP-hard, while the submodular minimization problem can be solved in polynomial time (Bach, 2013; Calinescu et al., 2007). For non-monotone submodular objectives, any polynomial time algorithm for the maximization is only guaranteed to satisfy a lower bound of $(1 - \frac{1}{e})$ from optimality, i.e., $F(S) \geq (1 - \frac{1}{e})F(S^*)$, where S^* is the argmax solution (Freige et al., 2011). For monotone submodular objectives, a greedy algorithm can do better, but only efficient approximations are known. Monotonicity and submodularity are generally not satisfied by even the rational score functions.

On the other hand, the number of candidate partitionings would be greatly reduced by the requirement that each partition S_j be *consecutive*, i.e., of the form $\{s_i, s_{i+1}, \ldots, s_k\}$ for $1 \leq i \leq k \leq n$, assuming some appropriate ordering of the data elements. The set $\mathcal{T}_{n,t}$ of partitionings of $\mathcal{V}$ of size t containing only consecutive elements has size $\binom{n-1}{t-1}$, which grows polynomially in n, as $\mathcal{O}\left(n^{t-1}\right)$. We will show that, under certain sufficient conditions on f, the optimal partitioning is guaranteed to be consecutive when the set $\mathcal{D}$ is ordered under an appropriate priority function g. Formally, we define:

Definition 2. *A priority function is a function $g \colon \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}$ that induces an ordering on the dataset $\mathcal{D}$. We refer to $g(x, y) = \frac{x}{y}$ as the standard priority function.*

Unless otherwise stated, the sets X, Y will be assumed to be already indexed in standard priority order, i.e., $g(x_1, y_1) \leq \cdots \leq g(x_n, y_n)$, where g is the standard priority function.

Given the dataset $\mathcal{D}$ ordered by priority, we can now formally define the *consecutive partitions property.*

Definition 3. *A non-empty subset $S \subseteq \mathcal{D}$ is* consecutive *if it is of the form $\{s_i, s_{i+1}, \ldots, s_k\}$, for $1 \leq i \leq k \leq n$.*

Definition 4. *A partitioning $P = \{S_1, \ldots, S_t\}$ of $\mathcal{D}$ of size t is* consecutive *if each partition S_j is a consecutive subset.*

Definition 5. *A partition scan statistic $F(P)$, with associated score function $f(x, y)$ and priority function $g(x, y)$, satisfies the* consecutive partitions property (CPP) *if, for all datasets*

$\mathcal{D} = \{(x_i, y_i)\}$ *ordered by priority, and all sizes* $1 \leq t \leq |\mathcal{D}|$*, the highest scoring partitioning of* $\mathcal{D}$ *of size* t*,* $P^* = \arg\max_{P=\{S_1,\ldots,S_t\}} \sum_{j=1\ldots t} f\left(\sum_{s_i \in S_j} x_i, \sum_{s_i \in S_j} y_i\right)$*, is consecutive.*

We also define a weaker version of CPP, as follows:

Definition 6. *A partition scan statistic* $F(P)$*, with associated score function* $f(x, y)$ *and priority function* $g(x, y)$*, satisfies the* weak consecutive partitions property (WCPP) *if, for all datasets* $\mathcal{D} = \{(x_i, y_i)\}$ *ordered by priority, and all sizes* $1 \leq t \leq |D|$*, the highest scoring partitioning of* $\mathcal{D}$ *of size* $t' \leq t$*,* $P^* = \arg\max_{P=\{S_1,\ldots,S_{t'}\}, t' \leq t} \sum_{j=1\ldots t'} f\left(\sum_{s_i \in S_j} x_i, \sum_{s_i \in S_j} y_i\right)$*, is consecutive.*

It is clear from these definitions that, if $F(P)$ satisfies CPP, then it also satisfies WCPP. However, the converse is not necessarily true. As a concrete example, Chakravarty et al. (1982) show that, if $f(x, y)$ is convex and $y_i > 0$ for all i, then a consecutive partitioning of size $t' \leq t$ exists that is maximal in the sense of (1) for the standard priority function; in other words, $F(P)$ satisfies WCPP. Let us consider the rational score function $f(x, y) = \frac{x^4}{y}$, which is convex, and consider the dataset $\mathcal{D} = (X, Y)$, where $X = \{8, 2, 9\}$ and $Y = \{8, 1, 3\}$, ordered by priority. We observe that the highest-scoring partitioning $F(P)$ of size $t = 2$ consists of $\{s_1, s_3\}$ and $\{s_2\}$, which is not consecutive, so CPP does not hold. The highest-scoring partitioning of size $t' \leq 2$ is the trivial (size-1) partitioning $P = \{S_1\}$, where $S_1 = \{s_1, s_2, s_3\}$. In such cases, we are unable to say anything about the optimal partitioning for sizes $t > 1$.

Thus we wish to identify sufficient conditions under which CPP and WCPP hold. These are provided by our main results, Theorems 1 and 2:

Theorem 1. *Given a partition scan statistic* $F(P)$ *with associated score function* $f(x, y)$ *and priority function* $g(x, y) = \frac{x}{y}$*. If* f *is convex and subadditive, then* F *satisfies CPP.*

Theorem 2. *Given a partition scan statistic* $F(P)$ *with associated score function* $f(x, y)$ *and priority function* $g(x, y) = \frac{x}{y}$*. If* f *is convex, then* F *satisfies WCPP.*

Note that convexity and subadditivity are both preserved under function addition. The proofs of Theorems 1 and 2 are based on a novel geometric construction, described in detail in the following section. It follows that partition scan statistics for the separable exponential family satisfy CPP:

Corollary 1. *Let $F(P)$ be the partition scan statistic corresponding to the log-likelihood ratio statistic for a distribution Ψ in a separable exponential family, with associated score function $f(x,y) = y\phi_0\left(\frac{x}{y}\right)$, where ϕ_0 is defined as above, and priority function $g(x,y) = \frac{x}{y}$. Then $F(P)$ satisfies CPP.*

Proof. The proof of Corollary 1 follows from the convexity and subadditivity of $f(x,y)$, and from Theorem 1 above. Convexity of f follows from convexity of ϕ_0 and the fact that the perspective of a convex function is also convex. To show subadditivity, we have:

$$\begin{aligned} f(x_1+x_2, y_1+y_2) &= (y_1+y_2)\phi_0\left(\frac{x_1+x_2}{y_1+y_2}\right) \\ &= (y_1+y_2)\phi_0\left(\lambda\left(\frac{x_1}{y_1}\right) + (1-\lambda)\left(\frac{x_2}{y_2}\right)\right) \\ &\leq (y_1+y_2)\left(\lambda\phi_0\left(\frac{x_1}{y_1}\right) + (1-\lambda)\phi_0\left(\frac{x_2}{y_2}\right)\right) \\ &= y_1\phi_0\left(\frac{x_1}{y_1}\right) + y_2\phi_0\left(\frac{x_2}{y_2}\right) \\ &= f(x_1,y_1) + f(x_2,y_2), \end{aligned}$$

where $\lambda = \frac{y_1}{y_1+y_2}$ and the inequality follows from convexity of ϕ_0. □

We can also easily characterize the partition scan statistics corresponding to the rational score functions:

Corollary 2. *Let $F(P)$ be the partition scan statistic with associated score function $f(x,y) = x^\alpha y^{-\beta}$, for constants $\alpha, \beta > 0$, and priority function $g(x,y) = \frac{x}{y}$. Assume $x > 0$. If $\alpha - \beta \geq 1$, then $F(P)$ satisfies WCPP. If $\alpha - \beta = 1$, then $F(P)$ satisfies CPP.*

Proof. We note that the Hessian of f has principal minors

$$\begin{aligned} M_1 &= \alpha(\alpha-1)x^{\alpha-2}y^{-\beta}, \\ M_2 &= \alpha\beta(\alpha-\beta-1)x^{2(\alpha-1)}y^{-2(\beta+1)}, \end{aligned}$$

so that for $x \in \mathbf{R}^+$, f is convex iff $\alpha - \beta \geq 1$, and f is subadditive iff $\alpha - \beta \leq 1$. The result then follows from Theorems 1 and 2 above. □

A similar result holds when x is not restricted to be positive, but in this case α must be an even integer:

Corollary 3. *Let $F(P)$ be the partition scan statistic with associated score function $f(x,y) = x^\alpha y^{-\beta}$, for constants $\alpha, \beta > 0$, and priority function $g(x,y) = \frac{x}{y}$. If $\alpha \in \mathbf{N}$ is even and $\alpha - \beta \geq 1$, then $F(P)$ satisfies WCPP. If $\alpha \in \mathbf{N}$ is even and $\alpha - \beta = 1$, then $F(P)$ satisfies CPP.*

Proof. For $x \in \mathbf{R}$, f is convex iff $\alpha - \beta \geq 1$, and $\alpha \in \mathbf{N}$ is even. f is subadditive iff $\alpha - \beta \leq 1$, and $\alpha, \beta \in \mathbf{N}$. The result then follows from Theorems 1 and 2 above. □

4 A geometric proof of the Consecutive Partitions Property

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ be a dataset, where $x_i \in \mathbf{R}$ and $y_i \in \mathbf{R}^+$, ordered by the standard priority function $g(x,y) = \frac{x}{y}$. Let us consider the subsets $S \subseteq \mathcal{V}$ and their corresponding sums $x_S = \sum_{i \in S} x_i$ and $y_S = \sum_{i \in S} y_i$, associating each subset S with a corresponding point in $\mathbf{R}^2$, $p_S = (x_S, y_S)$, which we term the *partition point* of S. By convention, $p_\emptyset = (0,0)$. Our proofs of Theorems 1 and 2 are based on a geometric characterization of the convex hull of the partition points p_S for $S \subseteq \mathcal{V}$. Denoting the convex hull of a set $S \subseteq \mathbf{R}^m$ as $\hat{S}$, we define:

Definition 7. *For $S \subseteq \mathcal{V}$, let $p_S = (\sum_{i \in S} x_i, \sum_{i \in S} y_i) \in \mathbf{R}^2$. Then the* partition polytope $\mathcal{C}$ *and the* constrained partition polytope $\underline{\mathcal{C}}$ *are defined by the convex hulls $\mathcal{C} = \hat{P}$ and $\underline{\mathcal{C}} = \hat{\underline{P}}$ respectively, where $P = \{p_S : S \subseteq \mathcal{V}\}$ and $\underline{P} = \{p_S : S \subseteq \mathcal{V}, S \neq \emptyset, S \neq \mathcal{V}\}$.*

We now characterize the extreme points of $\mathcal{C}$ and $\underline{\mathcal{C}}$. It will then follow from the convexity of the score function $f(x,y)$ that the optimal size-2 partitioning is $P^* = \{S, \mathcal{V} \setminus S\}$ for some extreme point p_S, and the remainder of the proofs of Theorems 1 and 2 proceed by induction. Let us define:

Definition 8. *A consecutive subset $S \subseteq \mathcal{V}$ is* nonsplitting *if both S and $\mathcal{V} \setminus S$ are consecutive; otherwise, it is* splitting. *A consecutive nonsplitting subset of the form $\{1, \ldots, j\}$ is* ascending, *and a consecutive nonsplitting subset of the form $\{j, \ldots, n\}$ is* descending. *By convention, $\emptyset$ and*

$\mathcal{V}$ *are consecutive nonsplitting, both ascending and descending. Define*

$$\begin{aligned}
\mathcal{T} = \mathcal{T}_n &= \{S \subseteq \mathcal{V} : S \textit{ is consecutive}\},\\
\mathcal{S} = \mathcal{S}_n &= \{S \subseteq \mathcal{V} : S \textit{ is consecutive splitting}\},\\
\mathcal{U} = \mathcal{U}_n &= \{S \subseteq \mathcal{V} : S \textit{ is consecutive nonsplitting}\},\\
\underline{\mathcal{U}} = \underline{\mathcal{U}}_n &= \mathcal{U}_n \setminus \{\emptyset, \mathcal{V}\}\,.
\end{aligned}$$

A special case occurs when S is the singleton $\{j\}$: the pair $S, \mathcal{V}\setminus S$ is called a *singleton splitting* or *nonsplitting* pair, depending on whether $\mathcal{V} \setminus S$ is consecutive. The set of singleton splitting subsets S and their complements in $\mathcal{V}$ is denoted by

$$\underline{\mathcal{S}} = \underline{\mathcal{S}}_n = \{S \subseteq \mathcal{V}\text{: } (S, \mathcal{V} \setminus S) \text{ form a singleton splitting pair}\}.$$

We note that $\mathcal{S} \subset \mathcal{T}$ and $\underline{\mathcal{U}} \subset \mathcal{U} \subseteq \mathcal{T}$, but $\underline{\mathcal{S}} \not\subseteq \mathcal{S}, \mathcal{T}$, since $\underline{\mathcal{S}}$ includes the non-consecutive complement $\mathcal{V} \setminus S$ for each singleton splitting subset S. It is easy to see that $|\mathcal{U}_n| = 2n$, $|\mathcal{S}_n| = \frac{(n-1)(n-2)}{2}$, with $|\mathcal{T}_n| = |\mathcal{U}_n| + |\mathcal{S}_n| = \frac{n(n+1)}{2} + 1$, while $|\underline{\mathcal{S}}_n| = 2n - 4$.

For any partitioning $P = \{S_1, \ldots, S_t\}$, or more generally, a set of subsets of $\mathcal{V}$, denote by $\Pi(P)$ the mapping taking P to the point set $\{\{p_{S_1}\}, \ldots, \{p_{S_t}\}\} \subset \mathbf{R}^2$, so $\Pi(\mathcal{S}) \subset \Pi(\mathcal{T})$, $\Pi(\mathcal{U}) \subseteq \Pi(\mathcal{T})$, etc.

Proposition 1. *Let $\mathcal{D} = \{(x, y)\}$ be a dataset ordered by the standard priority function $g(x, y) = \frac{x}{y}$. Let $\mathcal{E}$ and $\underline{\mathcal{E}}$ be the sets of extreme points of the partition polytope $\mathcal{C}$ and the constrained partition polytope $\underline{\mathcal{C}}$, respectively. Then*

(i) $\mathcal{E} \subseteq \Pi(\mathcal{U})$,

(ii) $\underline{\mathcal{E}} \subseteq \Pi(\underline{\mathcal{U}}) \cup \Pi(\underline{\mathcal{S}})$.

In particular, $\mathcal{E}$ contains only consecutive nonsplitting subsets, while $\underline{\mathcal{E}}$ may also contain consecutive splitting subsets and nonconsecutive subsets, but only singletons and their complements.

Proof. Let $p^* = (x^*, y^*) \in \mathcal{E}$. Then there is an affine $l(x, y) = Ax + By + C$ which separates p^* in $\mathcal{C}$, or equivalently a vector $v = (A, B) \in \mathbf{R}^2$ for which

$$(x^*, y^*) = \operatorname*{argmin}_{p \in \mathcal{C}} v \cdot p = \operatorname*{argmin}_{p_S \in \mathcal{C}: S \subseteq \mathcal{V}} v \cdot p_S. \tag{3}$$

The point p_{S^*}, where $S^* = \{i \in \mathcal{V} \colon v \cdot (x_i, y_i) \leq 0\}$ minimizes the last expression. If $A = 0$ then necessarily $B \neq 0$, in which case $S^* = \{i \in \mathcal{V} \colon y_i \leq 0\} = \emptyset$ or $S^* = \{i \in \mathcal{V} \colon y_i \geq 0\} = \mathcal{V}$, so that S^* and $\mathcal{V} \setminus S^*$ are both consecutive. Otherwise $S^* = \left\{i \in \mathcal{V} \colon \frac{x_i}{y_i} \leq -\frac{B}{A}\right\}$ or $S^* = \left\{i \in \mathcal{V} \colon \frac{x_i}{y_i} \geq -\frac{B}{A}\right\}$, depending on the sign of A, and the same conclusion holds. So $\mathcal{E} \subseteq \Pi(\mathcal{U})$.

We now prove that $\underline{\mathcal{E}} \subseteq \Pi(\underline{\mathcal{U}}) \cup \Pi(\underline{\mathcal{S}})$. For $\underline{p} = \left(\underline{x}, \underline{y}\right) \in \underline{\mathcal{E}}$, we can again find a separating affine function $l(x, y) = Ax + By + C$ and corresponding $v = (A, B)$ satisfying

$$\underline{p} = \operatorname*{argmin}_{p \in \underline{\mathcal{C}}} v \cdot p = \operatorname*{argmin}_{\substack{p_S \in \mathcal{C} : S \subseteq \mathcal{V}, \\ S \neq \emptyset, \\ S \neq \mathcal{V}}} v \cdot p_S. \tag{4}$$

Let $p = \operatorname{argmin}_{p_S \in \mathcal{C} : S \subseteq \mathcal{V}} v \cdot p_S$, which by the previous argument is of the form p_S for some consecutive nonsplitting subset S. If $0 < |S| < n$, then it also minimizes the expression in (4), and $p \in \Pi(\underline{\mathcal{U}})$. Otherwise there are two cases.

Case 1: $S = \emptyset$. Since the minimizing subset S can be written $S = \{i \in \mathcal{V} \colon v \cdot (x_i, y_i) \leq 0\}$, we must have $Ax + By > 0$ for all $(x, y) \in \mathcal{C}$. So expression (4) is minimized by selecting $p = p_S$, for $S = \{i\}$, where $i^* = \operatorname{argmin}_{i \in \mathcal{V}}(Ax_i + By_i)$, a singleton. It follows that p lies in either $\Pi(\underline{\mathcal{U}})$ or $\Pi(\underline{\mathcal{S}})$, depending on whether $\{i\}$ is splitting.

Case 2: $S = \mathcal{V}$. By similar reasoning, $Ax + By < 0$ for all $(x, y) \in \mathcal{C}$. Thus expression (4) is minimized by removing a single element from $\mathcal{V}$; take $S = \mathcal{V} \setminus \{i^*\}$, where $i^* = \operatorname{argmax}_{i \in \mathcal{V}}(Ax_i + By_i)$, and set $p = p_S$. Again p lies in either $\Pi(\underline{\mathcal{U}})$ or $\Pi(\underline{\mathcal{S}})$.

Thus $\underline{\mathcal{E}} \subseteq \Pi(\underline{\mathcal{U}}) \cup \Pi(\underline{\mathcal{S}})$. □

We now define a symmetric score function $\overline{f} \colon \mathcal{C} \subseteq \mathbf{R} \times \mathbf{R}^+ \to \mathbf{R}$ by

$$\overline{f}(x, y) = \begin{cases} f(x, y) + f\left(C_x^n - x, C_y^n - y\right), & (x, y) \in \mathcal{C} \setminus \left\{(0, 0), \left(C_x^n, C_y^n\right)\right\} \\ f\left(C_x^n, C_y^n\right), & (x, y) \in \left\{(0, 0), \left(C_x^n, C_y^n\right)\right\} \end{cases}$$

where $C_x^n = \sum_{i=1 \ldots n} x_i$, $C_y^n = \sum_{i=1 \ldots n} y_i$. We note that $\overline{f}$ is continuous on $\mathcal{C}$ for any score function f, and is convex on $\mathcal{C}$ for any convex score function f.

By Proposition 1, a convex $\overline{f}$ will take its maximum on $\underline{\mathcal{C}}$ at either a splitting or nonsplitting consecutive partition point. We show that the former case represents a kind of degenerate behavior, associated with a collapse of the unconstrained case to the trivial partitioning:

Proposition 2. *Let $f(x,y)$ be a convex score function and define the symmetric score function $\overline{f}(x,y)$ as above. If the set $\operatorname{argmax}_{(x,y)\in\underline{\mathcal{C}}}\overline{f}\,(x,y)$ contains only partition points associated with consecutive splitting subsets, then $p_{\mathcal{V}} = \operatorname{argmax}_{(x,y)\in\mathcal{C}}\overline{f}(x,y)$.*

Proof. Let $\underline{p} = \operatorname{argmax}_{(x,y)\in\underline{\mathcal{C}}}\overline{f}(x,y)$, with $\underline{p}$ consecutive splitting in $\mathcal{V}$. Let $p = \operatorname{argmax}_{(x,y)\in\mathcal{C}}\overline{f}(x,y)$. Convexity of $\overline{f}$ on $\mathcal{C}$ follows from convexity of f, so $p \in \mathcal{E}$. Since $\overline{f}(p) \geq \overline{f}(\underline{p})$, and $\underline{p} \notin \mathcal{E}$ by Proposition 1, we have $\overline{f}(p) > \overline{f}(\underline{p})$. So p cannot lie in $\underline{\mathcal{E}}$, and thus $p \in \mathcal{E} \setminus \underline{\mathcal{E}} = \{p_\emptyset, p_{\mathcal{V}}\}$. □

Given Propositions 1 and 2, we can now prove the main theorems:

Theorem 1. *Given a partition scan statistic $F(P)$ with associated score function $f(x,y)$ and priority function $g(x,y) = \frac{x}{y}$. If f is convex and subadditive, then F satisfies CPP.*

Proof. The proof proceeds by induction on t. Let $t = 2$, and note that $\operatorname{argmax}_P F(P) = \{S^*, \mathcal{V} \setminus S^*\}$, where $S^* = \operatorname{argmax}_S \overline{f}(p_S)$. By convexity of $\overline{f}$ and Proposition 1, the maximum of $\overline{f}$ on $\underline{\mathcal{C}}$ occurs at a point $p_S \in \Pi(\underline{\mathcal{U}}) \cup \Pi(\underline{\mathcal{S}})$. Since f is subadditive, the maximum of $\overline{f}$ on $\mathcal{C}$ cannot occur at $p_{\mathcal{V}}$, hence $p_S \in \Pi(\underline{\mathcal{U}})$ by Proposition 2. So the maximal partitioning $\{S, \mathcal{V} \setminus S\}$ is consecutive.

Now let $t \geq 3$. The arguments here are motivated by Chakravarty et al. (1982). Define, for any subset $S \subseteq \mathcal{V}$, $M_S = \max i : i \in S$, $m_S = \min i : i \in S$, and $d(S) = M_S - m_S$. Let $P = \{S_1, \ldots, S_t\}$ be a size-t partitioning that maximizes $F(P)$ and also minimizes $\sum_{j=1\ldots t} d(S_j)$ among all optimal size-t partitionings. If not all S_j are consecutive, then we can find two partitions $S_i, S_j \in P$ and an element $k \in S_j$ with $m_{S_i} < k < M_{S_i}$. Note that $\min(M_{S_i}, M_{S_j}) \geq k$, and $\max(m_{S_i}, m_{S_j}) \leq k$, so that $\min(M_{S_i}, M_{S_j}) - \max(m_{S_i}, m_{S_j}) \geq 0$. By the induction hypothesis, we can find an optimal partitioning of $S_i \cup S_j$ into nonempty consecutive sets S_i', S_j'. Then $d\left(S_i'\right) + d\left(S_j'\right) \leq \max\left(M_{S_i}, M_{S_j}\right) - \min\left(m_{S_i}, m_{S_j}\right) - 1$, so

$$\begin{aligned} d\left(S_i\right) + d\left(S_j\right) &= M_{S_i} - m_{S_i} + M_{S_j} - m_{S_j} \\ &= \max\left(M_{S_i}, M_{S_j}\right) + \min\left(M_{S_i}, M_{S_j}\right) \\ &\quad - \min\left(m_{S_i}, m_{S_j}\right) - \max\left(m_{S_i}, m_{S_j}\right) \\ &\geq d\left(S_i'\right) + d\left(S_j'\right) + 1, \end{aligned}$$

a contradiction. So all of the S_j are consecutive and nonempty. □

Theorem 2. *Given a partition scan statistic $F(P)$ with associated score function $f(x, y)$ and priority function $g(x, y) = \frac{x}{y}$. If f is convex, then F satisfies WCPP.*

Proof. The proof proceeds by induction as in the convex, subadditive case, but we allow the partitioning $P = \{\mathcal{V}, \emptyset\}$ at the initial step, and thereafter for the base assumption in the induction step. Call this the trivial size-2 partitioning. In this way we will create a partitioning of size $t' \leq t$ consisting of consecutive elements and possibly repetitions of instances of the empty set.

To this end, let $t = 2$ and consider $p^* = \text{argmax}_{p \in \mathcal{C}} \overline{f}(p)$. By Proposition 1, $p^* \in \Pi(\mathcal{U})$, and we have a (possibly trivial) maximal partitioning of $\mathcal{V}$ with size $t' \leq 2$.

The inductive step proceeds as before, and assuming S_i, S_j nonconsecutive, we produce the consecutive partitions S_i' and S_j' with $S_i' \cup S_j' = S_i \cup S_j$, with possibly one of S_i', S_j' empty. If, without loss of generality, $S_j' = \emptyset$, then $d(S_i') + d(S_j') \leq \max(M_{S_i}, M_{S_j}) - \min(m_{S_i}, m_{S_j}) - 1$ as before, and we arrive at a contradiction. Thus we obtain, for any t, an optimal consecutive partitioning P of $\mathcal{V}$, possibly containing empty sets and thus having size $t' \leq t$. Therefore, WCPP holds for F. □

4.1 Further characterization of the partition polytopes

Proposition 1 demonstrates that the extreme points of the unconstrained partition polytope $\mathcal{C}$ consist only of consecutive nonsplitting subsets, while the extreme points of the constrained partition polytope $\underline{\mathcal{C}}$ may include consecutive nonsplitting subsets (excluding $\emptyset$ and $\mathcal{V}$), singleton splitting subsets, and their (non-consecutive) complements. We can further characterize these polytopes by showing that (i) all consecutive nonsplitting subsets $p_S \in \Pi(\mathcal{U})$ are on the boundary of the unconstrained partition polytope, and (ii) all consecutive nonsplitting subsets $p_S \in \Pi(\underline{\mathcal{U}})$ are on the boundary of the constrained partition polytope.

Given the dataset $\mathcal{D} = \{(x_i, y_i)\}$, we define the summations $C_x^j = \sum_{i=1\ldots j} x_i$, $C_y^j = \sum_{i=1\ldots j} y_i$, $C_x^{-j} = \sum_{i=j\ldots n} x_i$, and $C_y^{-j} = \sum_{i=j\ldots n} y_i$, for $1 \leq j \leq n$, with the conventions that $C_x^0 = C_y^0 = C_x^{-(n+1)} = C_y^{-(n+1)} = 0$ and $x_0 = y_0 = 0$. Thus $p_{C^j} = (C_x^j, C_y^j)$ is the partition point associated with the ascending nonsplitting subset $C^j = \{1, \ldots, j\}$, while $p_{C^{-j}} = (C_x^{-j}, C_y^{-j})$ is the partition point associated with the descending nonsplitting subset $C^{-j} = \{j, \ldots, n\}$.

We can now construct the convex hull of the points p_{C^j} (for $j = 0 \ldots n$) and $p_{C^{-j}}$ (for $j = 2 \ldots n$)

using a Graham scan (Graham, 1972). By repeated application of the inequality $\frac{x_j}{y_j} \leq \frac{x_j+x_k}{y_j+y_k} \leq \frac{x_k}{y_k}$ for $j < k$, it follows that $\frac{C_x^1}{C_y^1} \leq \cdots \leq \frac{C_x^n}{C_y^n} \leq \frac{C_x^{-2}}{C_y^{-2}} \leq \cdots \leq \frac{C_x^{-n}}{C_y^{-n}}$.

So a ray sweeping out an angle clockwise from $-\pi$ to π radians, anchored at the origin p_{C^0}, will meet the extreme points in the order $p_{C^1}, \ldots, p_{C^n}, p_{C^{-2}}, \ldots, p_{C^{-n}}$.

We note that p_{C^0} and p_{C^n} are extreme points since they have the lowest and highest y-coordinates respectively. For other points, we can use the Graham scan's "right turn rule" to check whether they are on the boundary of the convex hull.

For p_{C^j} (where $1 \leq j \leq n-1$), we have:

$$\begin{aligned}(C_x^j - C_x^{j-1})(C_y^{j+1} - C_y^{j-1}) - (C_y^j - C_y^{j-1})(C_x^{j+1} - C_x^{j-1}) &= x_j(y_j + y_{j+1}) - y_j(x_j + x_{j+1}) \\ &= x_j y_{j+1} - y_j x_{j+1} \\ &= y_j y_{j+1}\left(\frac{x_j}{y_j} - \frac{x_{j+1}}{y_{j+1}}\right) \\ &\leq 0.\end{aligned}$$

Thus p_{C^j} is on the boundary of the unconstrained partition polytope, and is an extreme point if $\frac{x_j}{y_j}$ is strictly less than $\frac{x_{j+1}}{y_{j+1}}$. Similarly, for $p_{C^{-j}}$ (where $2 \leq j \leq n$), we have:

$$\begin{aligned}(C_x^{-j} - C_x^{-(j-1)})(C_y^{-(j+1)} - C_y^{-(j-1)}) - (C_y^{-j} - C_y^{-(j-1)})(C_x^{-(j+1)} - C_x^{-(j-1)}) \\ &= x_{j-1}(y_{j-1} + y_j) \quad y_{j-1}(x_{j-1} + x_j) \\ &= x_{j-1}y_j - y_{j-1}x_j \\ &= y_j y_{j-1}\left(\frac{x_{j-1}}{y_{j-1}} - \frac{x_j}{y_j}\right) \\ &\leq 0.\end{aligned}$$

Thus $p_{C^{-j}}$ is on the boundary of the unconstrained partition polytope, and is an extreme point if $\frac{x_{j-1}}{y_{j-1}}$ is strictly less than $\frac{x_j}{y_j}$.

From this argument, we can see that all of the consecutive nonsplitting subsets $S \in \mathcal{U}$ correspond to points p_S on the boundary of the unconstrained partition polytope $\mathcal{C}$, and to extreme points if the priority function $g(x_j, y_j) = \frac{x_j}{y_j}$ is strictly increasing with j. Again, from Proposition 1, no other extreme points exist.

For the constrained partition polytope $\underline{\mathcal{C}}$, we remove $p_\emptyset = p_{C^0}$ and $p_\mathcal{V} = p_{C^n}$, and consider the convex hull of the remaining points. All of the points p_S corresponding to the consecutive

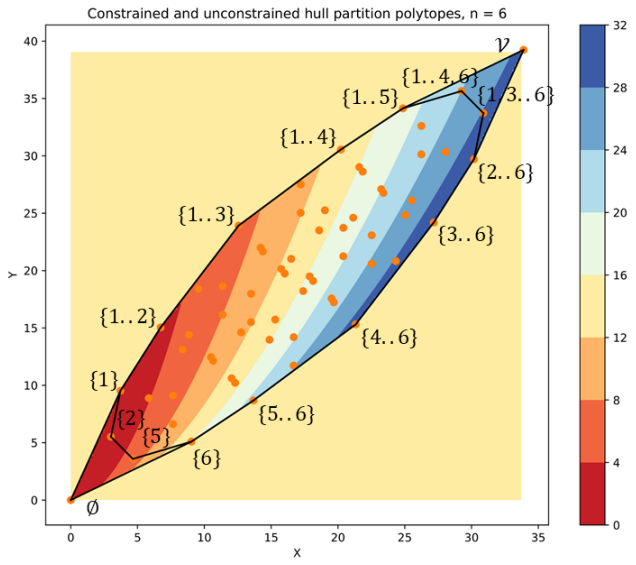


Figure 1: Example of partition polytope $\mathcal{C}$ and constrained partition polytope $\underline{\mathcal{C}}$ for $n = 6$ points, with level sets for $f(x, y) = \frac{x^2}{y}$. When points $\emptyset$ and $\mathcal{V}$ are removed from $\mathcal{C}$, the singleton splitting subsets $\{2\}$ and $\{5\}$, and their non-consecutive complements, become extreme points of $\underline{\mathcal{C}}$.

nonsplitting subsets $S \in \underline{\mathcal{U}}$ remain on the boundary of the convex hull when these two points are removed. Additionally, from Proposition 1, the singleton splitting pairs must be considered, as some or all of these may be vertices. See Figure 1 for an example of unconstrained and constrained partition polytopes, where removal of $p_\emptyset$ and $p_\mathcal{V}$ adds a singleton splitting pair to the boundary of the convex hull.

If any singleton is a vertex of $\underline{\mathcal{C}}$, it will lie between $p_{C^{-n}}$ and p_{C^1} when the vertices of $\mathcal{C}$ are traversed in a clockwise direction along the partition polytope. Since the singleton splitting subsets $p_{\{i\}}$ are indexed by $1 < i < n$, they can be enumerated and tested successively. Thus we can perform a Graham scan to identify the lower convex hull of the points $p_{C^{-n}}$, $p_{\{n-1\}}$, $p_{\{n-2\}}$, $\ldots$, $p_{\{2\}}$, p_{C^1} (in clockwise order). For any points $p = (x, y)$, $p' = (x', y')$, and $p'' = (x'', y'')$ in the given order, if $(x' - x)(y'' - y) - (y' - y)(x'' - x) > 0$, then p' is in the interior of $\underline{\mathcal{C}}$ and can be removed, and this process can be repeated until all remaining points are on the boundary of $\underline{\mathcal{C}}$.

Finally, the extreme point classification allows us to specify sufficient conditions on the dataset $\mathcal{D} = \{(x, y)\}$ under which the solution to expression (1) occurs for a consecutive partitioning of size t, for f convex but not necessarily subadditive:

Theorem 3. *Given a partition scan statistic $F(P)$ with associated score function $f(x, y)$ and priority function $g(x, y) = \frac{x}{y}$, and a dataset $\mathcal{D} = \{(x_i, y_i)\}$ (for $i = 1 \ldots n$) ordered by priority. If f is convex, and $(x_i - x_n)(y_1 - y_n) - (y_i - y_n)(x_1 - x_n) \geq 0$ for $i = 2 \ldots n-1$, then the solution to*

$$\operatorname*{argmax}_{P=\{S_1,\ldots,S_t\}} \sum_{j=1\ldots t} f(\sum_{i \in S_j} x_i, \sum_{i \in S_j} y_i)$$

is a consecutive partitioning of size t.

Proof. By the right turn rule, if $(x_i - x_n)(y_1 - y_n) - (y_i - y_n)(x_1 - x_n) \geq 0$, then the singleton splitting subset $p_{\{i\}}$ lies above the segment joining the vertices $p_{C^{-n}}$ and p_{C^1}, and is thus in the interior of $\underline{\mathcal{C}}$, not an extreme point. If all singleton splitting subsets are in the interior, then by Proposition 1, the only extreme points of $\underline{\mathcal{C}}$ correspond to consecutive nonsplitting subsets, and the symmetric score function $\overline{f}(x, y)$ is maximized at one of these subsets. Finally, following Theorem 1, the partition scan statistic is maximized for a consecutive partitioning. □

For example, for the sets $X = \{7.0, 9.4, 8.9, 7.5, 6.9, 7.2, 7.5, 6.3\}$ and $Y = \{8.4, 9.3, 6.7, 4.7, 1.8, 1.4, 0.6, 0.1\}$, we know that any partition scan statistic with a convex score function f satisfies CPP for this particular choice of X and Y.

5 Efficient optimization over consecutive partitionings

As noted in §3 above, an exhaustive search over all size-t partitionings P, in order to find the optimal partitioning $P^* = \operatorname{argmax}_{P=\{S_1 \ldots S_t\}} F(P)$, is computationally infeasible even for relatively small datasets, requiring time exponential in the number of data records n. However, if a partition scan statistic is known to satisfy CPP or WCPP, then the corresponding optimization can be accelerated substantially by evaluating only the much smaller set of consecutive partitionings.

A constrained optimization over the set of all size-t consecutive partitionings has cost that grows as $\binom{n-1}{t-1}$, i.e., as a polynomial of degree $t-1$ in n. For very small t, exhaustive enumeration of the consecutive partitionings is a low-degree polynomial in n, and thus computationally feasible, but this approach does not scale well for larger t.

Thus we propose an alternative approach based on dynamic programming (Algorithm 1), which runs in quadratic $\mathcal{O}(n^2 t)$ time and requires linear $\mathcal{O}(nt)$ space. Algorithm 1 finds the

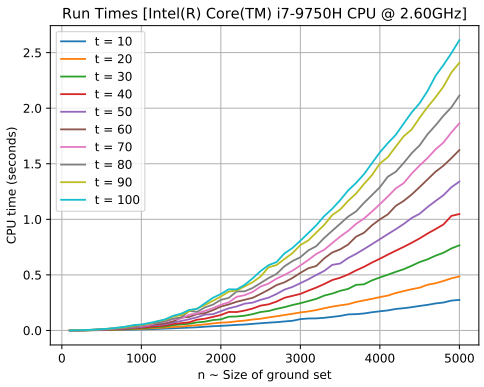


Figure 2: Run times of Algorithm 1 as a function of n, with score function $f(x, y) = x^2 y^{-1}$

highest-scoring consecutive partitioning, $\arg\max_{P=\{S_1 \ldots S_{t'}\}} F(P)$, for each size $t' \in \{1, \ldots, t\}$. For a partition scan statistic $F(P)$ satisfying CPP, we are guaranteed that the highest-scoring consecutive partitioning of size t is the optimal size-t partitioning. For a partition scan statistic satisfying WCPP but not CPP, the highest-scoring consecutive partitioning of size t is only guaranteed to be the optimal size-t partitioning if no other consecutive partitioning of size $t' < t$ has higher score.

For each size $t' \in \{1, \ldots, t\}$ and each element $j \in \{1, \ldots, (n + 1 - t')\}$, Algorithm 1 computes two quantities: (1) the maximum score $F^*(j, t')$ for dividing elements $\{j, \ldots, n\}$ into t' consecutive partitions, and (2) the starting element $\rho(j, t')$ of the second partition of the highest-scoring consecutive size-t' partitioning of $\{j, \ldots, n\}$. In the final stage of the algorithm (lines 25-36), the highest-scoring consecutive partitioning of each size $t' \in \{1, \ldots, t\}$ is recovered by repeatedly applying ρ, starting at $j = 1$, and its corresponding score is $F^*(1, t')$.

The key step of the algorithm (lines 16-20) is that $F^*(j, t') = \max_k f(C_x^{j,k}, C_y^{j,k}) + F^*(k+1, t'-1)$, where $k \in \{j, \ldots, (n + 1 - t')\}$, $f(C_x^{j,k}, C_y^{j,k})$ is the score of the partition formed by elements $\{j, \ldots, k\}$, and $F^*(k+1, t'-1)$ is the maximum score for dividing elements $\{k+1, \ldots, n\}$ into $t'-1$ consecutive partitions. We note that the sufficient statistics $(C_x^{j,k}, C_y^{j,k})$ are aggregated iteratively (lines 4 and 15), giving constant-time computation for each combination of $t \in \{1, \ldots, t'\}$, $j \in \{1, \ldots, (n + 1 - t')\}$, and $k \in \{j, \ldots, (n + 1 - t')\}$, for a total run time of $\mathcal{O}(n^2 t)$.

A display of run times of Algorithm 1 as a function of n, for various values of t, is shown in Figure 2. We observe the quadratic dependence of run time on n and linear dependence on t. Average run time was under 2.7 seconds for all $n \leq 5000$ and $t \leq 100$.

Algorithm 1 Dynamic Programming Algorithm for Optimizing over Consecutive Partitionings

```
# Base case (t' = 1)
(C_x, C_y) = (0, 0)
for j ∈ {n, n − 1, ..., 1} do
    (C_x, C_y) = (C_x, C_y) + (x_j, y_j)
    F*(j, 1) = f(C_x, C_y)
    ρ(j, 1) = n + 1
end for

# Iterative step
for t' ∈ {2, ..., t} do
    for j ∈ {1, ..., (n + 1 − t')} do
        F*(j, t') = −∞
        (C_x, C_y) = (0, 0)
        for k ∈ {j, ..., (n + 1 − t')} do
            (C_x, C_y) = (C_x, C_y) + (x_k, y_k)
            F^{j,k} = f(C_x, C_y) + F*(k + 1, t' − 1)
            if F^{j,k} > F*(j, t') then
                F*(j, t') = F^{j,k}
                ρ(j, t') = k + 1
            end if
        end for
    end for
end for

# Recover the highest-scoring consecutive partitioning for each t'
for t' ∈ {1, ..., t} do
    scores[t'] = F*(1, t').
    partitions[t'] = []
    j = 1;
    for t'' ∈ {t', (t' − 1), ..., 1} do
        j_next = ρ(j, t'')
        Append [j, j_next − 1] to partitions[t']
        j = j_next
    end for
end for
return scores, partitions
```

6 Partition scan statistics for multiple cluster detection

The partition scan statistics presented in §2.1 generalize Kulldorff's spatial scan statistic (Kulldorff, 1997) from $t = 2$ to $t > 2$ partitions and from the Poisson distribution to other distributions in a separable exponential family. As such, they are best applied to *risk partitioning*, where the goal is to optimally divide the dataset into t partitions of differing risks $q_1, \ldots, q_t$. We now consider the application of partition scan statistics to the related problem of *multiple cluster detection*, where we place additional constraints on the values of $q_1, \ldots, q_t$ (and q_{all}) in order to distinguish clusters of elevated risk from the remaining background data. More precisely, consider the null and alternative hypotheses in §2.1 above,

$$H_0 \colon x_i \sim \Psi(\cdot; \mu = q_{all}\mu_i) \; \forall s_i,$$
$$H_1(P) \colon x_i \sim \Psi(\cdot; \mu = q_j\mu_i) \; \forall s_i \in S_j,$$

with constraints $q_j > 1$ for $j \in \{2, \ldots, t\}$, and $q_1 = q_{all} = 1$. Then equation (2) simplifies to:

$$F_{clust}(P) = F_{clust}(\{S_1, \ldots, S_t\}) = \log \max_{q_2,\ldots,q_t>1} \prod_{j=2\ldots t} \prod_{s_i \in S_j} \frac{\Pr(x_i \mid x_i \sim \psi(\cdot; \mu = q_j\mu_i))}{\Pr(x_i \mid x_i \sim \psi(\cdot; \mu = \mu_i))}.$$

This generalizes the expectation-based scan statistic (Neill, 2012) from a single cluster ($t = 2$) to multiple clusters ($t > 2$). For consistency with Neill (2012), we allow partitions to be empty, so that the detected clusters (non-empty partitions with $q_j > 1$) could include some, all, or none of the data elements.

For expectation-based scan statistics in a separable exponential family, following Neill (2012):

$$F_{clust}(P) = F_{clust}(\{S_1, \ldots, S_t\}) = \sum_{j=2\ldots t} f(C_j, B_j),$$

where C_j and B_j are the sufficient statistics for partition S_j, $C_j = \sum_{s_i \in S_j} T(x_i)z_i$ and $B_j = \sum_{s_i \in S_j} \mu_i z_i$ respectively, and the score function $f(x, y) = yD_{\phi_0}(\frac{x}{y}, 1)$ for $x > y$, 0 otherwise, where D_{ϕ_0} is a Bregman divergence. We can now show:

Theorem 4. *Let* $F_{clust}(P) = F_{clust}(\{S_1, \ldots, S_t\}) = \sum_{j=2\ldots t} f(\sum_{s_i \in S_j} x_i, \sum_{s_i \in S_j} y_i)$ *be the multiple cluster scan statistic corresponding to the expectation-based scan statistic for a distribution* Ψ *in a separable exponential family, with associated score function* $f(x, y) = yD_{\phi_0}(\frac{x}{y}, 1)\, \mathbb{1}\{x > y\}$, *where*

D_{ϕ_0} is a Bregman divergence, and priority function $g(x, y) = \frac{x}{y}$. Let dataset $\mathcal{D} = \{(x_i, y_i)\}$, $i = 1 \dots n$, be ordered by priority. Then the solution to $P^* = \arg\max_{P=\{S_1 \dots S_t\}} F_{clust}(P)$, allowing empty partitions, is a consecutive partitioning of $\mathcal{D}$, of size $t' \leq t$, with S_1 ascending consecutive.

Proof. We first note that the score function $f(x, y)$ is convex, since Bregman divergences are convex in their first argument and the perspective of a convex function is convex. We now proceed by induction. For the case $t = 2$, the convex score function $f(x, y)$ is maximized at an extreme point of the partition polytope $\mathcal{C}$, which is consecutive by Proposition 1. By the reasoning in §4.1, the order of clockwise traversal of the vertices of $\mathcal{C}$, starting from the origin $p_{C^0} = p_{C^{-(n+1)}}$, is $p_{C^1}, \dots p_{C^n}, p_{C^{-2}}, \dots, p_{C^{-n}}$. Since the y-coordinates are increasing on $p_{C^1}, \dots p_{C^n}$, it follows that for each point p_{C^j} for $j \in \{1, \dots, n-1\}$, there is a $\delta > 0$ such that $(C_x^j + \delta, C_y^j)$ lies in $\mathcal{C}$. Since f is nondecreasing in x, the maximizing point must be descending consecutive of the form $p_{C^{-k}}$, for $1 \leq k \leq (n+1)$. Thus we have S_2 descending consecutive and S_1 ascending consecutive, allowing for the possibility that either partition could be empty.

For $t > 2$, consider any pair of subsets S_i and S_j, and assume without loss of generality that $j \neq 1$. There are two cases. First, if $i = 1$, then by the above reasoning, the optimal subset S'_j of $S_i \cup S_j$ is descending consecutive, and the corresponding S'_1 is ascending consecutive. Second, if $i \neq 1$, we proceed as in Theorem 2. Since f is convex, the symmetric score function $\overline{f}$ is also convex, and is maximized at an extreme point of $\mathcal{C}$.[2] By identical reasoning to Theorem 2, there exist optimal consecutive partitions S'_i and S'_j whose union is $S_i \cup S_j$. □

This generalizes the linear-time subset scanning property for expectation-based scan statistics in the separable exponential family (Neill, 2012) from a single detected cluster ($t = 2$) to multiple detected clusters ($t > 2$). Randomization testing can be used to assess the statistical significance of the detected clusters, as in Kulldorff (1997) and Neill (2012). For $t = 2$, the fast subset scan approach of Neill (2012) can be used to identify the highest-scoring subset in $O(n \log n)$ time while evaluating only $O(n)$ of the 2^n subsets. For $t > 2$, a slight modification of the dynamic programming approach described in §5 above can identify the highest-scoring partitioning, and

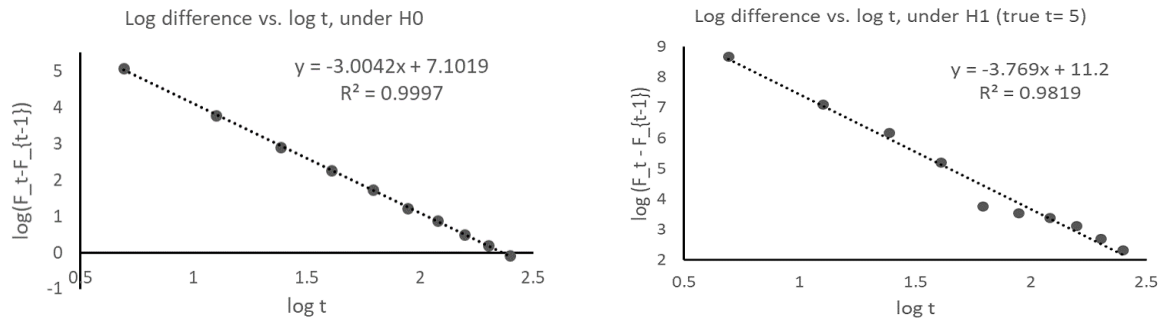
[2] We note that quasi-convexity of f is sufficient for the $t = 2$ case (as proved in Neill (2012)) but not for $t > 2$, since the symmetric score function $\overline{f}$ corresponding to a quasi-convex score function f is not necessarily quasi-convex. Subadditivity is not necessary here, since we allow empty partitions.

thus detect multiple clusters, in $O(n^2)$ time. To do so, we compute $F^*(j, t')$ as above, and $F^*_{clust} = \max_{t'} \max_j F^*(j, t')$, for $t' \in \{1, \ldots, t-1\}$ and $j \in \{1 \ldots (n+1-t')\}$.

Several alternative approaches have been proposed for extending the spatial scan statistic to multiple cluster detection (Kulldorff, 1997; Zhang et al., 2010; Li et al., 2011; Takahashi and Shimadzu, 2020). Most of these approaches work sequentially, first detecting the primary cluster and removing it from the data before searching for significant secondary clusters. Kulldorff's original spatial scan (Kulldorff, 1997) compares the scores of the secondary clusters to the same detection threshold, obtained by randomization testing, as the primary cluster. Zhang et al. (2010) note that this approach is conservative and adjust the threshold for more sensitive detection of secondary clusters, while Takahashi and Shimadzu (2020) propose a generalized linear model and information-theoretic criterion to select the number of clusters, starting with candidate clusters identified by the sequential method. Repeated single cluster detection is computationally efficient, scaling linearly with the number of detected clusters, but does not directly optimize the log-likelihood ratio statistic for multiple clusters. In contrast, we exactly and efficiently find the highest scoring partitioning in the unconstrained sense, which can be thought of as maximizing the joint LLR score corresponding to multiple detected clusters. This is most similar to Li et al. (2011), who explicitly scan over tuples of clusters and compute joint scores, but their approach requires a exhaustive search that scales exponentially with the number of clusters to be detected (and thus is typically limited to two or three clusters), while our approach can easily scale to very large n and t. Note that the goal of these previously proposed methods is different from ours: they focus entirely on spatial cluster detection (typically scanning over circular regions in space), while our approach identifies an optimal partitioning of the data without an explicit spatial constraint.

7 Choosing the number of partitions t

We now consider how to choose the number of partitions t, in cases when the desired value of t is not known *a priori*. We propose a novel, computationally efficient heuristic for finding t, and show in our experiments below that this approach can obtain the correct t value in practice. For a given dataset $\mathcal{D}$, let F^*_t denote the maximum score $\max_{P=\{S_1,\ldots,S_t\}} F(P)$ for separating $\mathcal{D}$ into t partitions, and assume that we wish to choose $t \in \{1, 2, \ldots, t_{max}\}$ for some constant t_{max}. Our



(a) Data generated under H_0.

(b) Data generated under H_1.

Figure 3: Plot of $\log(F_t^* - F_{t-1}^*)$ vs. $\log t$, for data generated under the null hypothesis H_0 and under the alternative hypothesis H_1 (with five true partitions and $\epsilon = .75$) respectively. Values of $t \in \{2, 3, \ldots, 11\}$ are shown. Note the large negative residual at $t = 6$ in plot (b).

heuristic is based on several observations. First, as shown in Figure 3a, we identify a power law relationship between the average score difference $\Delta_t = F_t^* - F_{t-1}^*$ and the value of t, for data generated under the null hypothesis H_0, such that $\log \Delta_t$ is linearly related to $\log t$. Second, as shown in Figure 3b, for data generated under the alternative hypothesis $H_1(P)$ with t_{true} distinct partitions, the relationship between $\log \Delta_t$ and $\log t$ deviates from linearity. In this case, we observe that the point $(\log t, \log \Delta_t)$ corresponding to $t = t_{true} + 1$ partitions has a large negative residual, indicating that the increase in score from adding partition $t_{true} + 1$ is less than expected given the overall trend. Intuitively, when partitions are sufficiently distinct, the Δ_t from separating two true partitions, when $t \leq t_{true}$, will be large compared to the Δ_t from splitting apart a single true partition, when $t > t_{true}$. Finally, we note that all scores $F_1^*, \ldots, F_{t_{max}+1}^*$ can be obtained efficiently from a single run of Algorithm 1, as opposed to requiring a separate run for each t.

This suggests the following algorithmic approach:

1. Compute $F_{t_{max}+1}^*$ by dynamic programming and also store the intermediate results $F_1^*, \ldots, F_{t_{max}}^*$.
2. Compute $\log \Delta_t = \log(F_t^* - F_{t-1}^*)$ for each $t \in \{2, 3, \ldots, t_{max} + 1\}$.
3. Fit an OLS linear regression $\log \Delta_t = w_1(\log t) + w_0$, and compute residuals for each t.
4. Identify the t with most negative residual $\log \Delta_t - w_1(\log t) - w_0$, and return $t - 1$.

The time complexity of this algorithm is $O(n^2 t_{max}) + O(t_{max}) = O(n^2 t_{max})$, where the first term for the dynamic programming dominates the second term for the linear regression.

8 Simulation experiments

We now perform two sets of simulation experiments to compare the partition scan to the previously proposed Kulldorff and expectation-based Poisson scan statistics, in the risk partitioning and multiple cluster detection settings respectively.

8.1 Risk partitioning simulations

For the first set of simulations, we evaluate the partition scan statistic with Poisson score function, $F(P) = \sum_{j=1...t} f(C_j, B_j) - f(C_{all}, B_{all})$, where $f(x, y) = x \log(\frac{x}{y})$. We compare $t = 2$, $t = 3$, and $t = 10$ partitions, where the $t = 2$ case corresponds to the original Poisson scan (Kulldorff, 1997).

We consider three different simulated scenarios (two true partitions, three true partitions, and ten true partitions). In each case, we generate simulated datasets each consisting of $n = 2500$ data records, where each data record has baseline $y_i \sim \text{Poisson}(100)$ and count $x_i \sim \text{Poisson}(q_i y_i)$, where q_i is the *relative risk* for that data record.

Each of the three scenarios is indexed by the "signal strength" ϵ, where we consider values of ϵ ranging from 0 (no signal) to 0.5 (strongest signal). For "two true partitions", the relative risks q_i are $1 - \epsilon$ and $1 + \epsilon$ for data records in the first and second partition respectively. For "three true partitions", we use relative risks of $1 - \epsilon$, 1, and $1 + \epsilon$ for the three partitions. For "ten true partitions", the relative risks are $1 - \epsilon$, $1 - 0.8\epsilon$, $1 - 0.6\epsilon$, $1 - 0.4\epsilon$, $1 - 0.2\epsilon$, $1 + 0.2\epsilon$, $1 + 0.4\epsilon$, $1 + 0.6\epsilon$ $1 + 0.8\epsilon$, and $1 + \epsilon$. For all three scenarios, the partitions are equal in size. For each scenario and each signal strength ϵ, we generated 500 simulated datasets. For each dataset and for each parameter setting ($t = 2$, $t = 3$, and $t = 10$), we compute the optimal partitioning $P^* = \arg\max_P F(P)$ and its score $F^* = F(P^*)$.

The methods were evaluated based on *ranking quality*, which measures how well the partitions created by the method correspond to the true partitions used to generate the data. To compute ranking quality, we consider the true partitions $S_1^{true}, \ldots, S_m^{true}$, ordered from smallest to largest

relative risk, and the detected partitions $S_1, \ldots, S_t$, likewise ordered from smallest to largest relative risk. We then compute the probability that a pair of data records (i, j) drawn uniformly at random with replacement, with corresponding true partitions S_i^{true} and S_j^{true} and corresponding detected partitions S_i and S_j, are ordered correctly by the method (i.e., $S_i > S_j$ if $S_i^{true} > S_j^{true}$, $S_i < S_j$ if $S_i^{true} < S_j^{true}$, and $S_i = S_j$ if $S_i^{true} = S_j^{true}$). Ranking quality can be calculated as:

$$RQ = \frac{\sum_{i=1}^{m}\sum_{j=1}^{t}\sum_{i'=1}^{m}\sum_{j'=1}^{t} N_{ij}N_{i'j'}(\mathbb{1}\{i < i'\}\mathbb{1}\{j < j'\} + \mathbb{1}\{i > i'\}\mathbb{1}\{j > j'\} + \mathbb{1}\{i = i'\}\mathbb{1}\{j = j'\})}{\sum_{i=1}^{m}\sum_{j=1}^{t}\sum_{i'=1}^{m}\sum_{j'=1}^{t} N_{ij}N_{i'j'}},$$

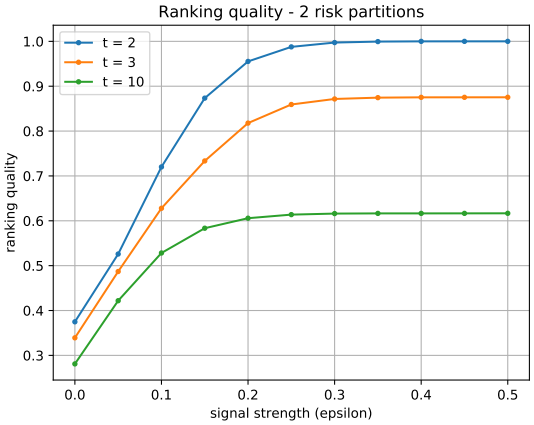
where N_{ij} is the number of data records assigned to $S_i^{true} \cap S_j$.

8.1.1 Risk partitioning simulation results

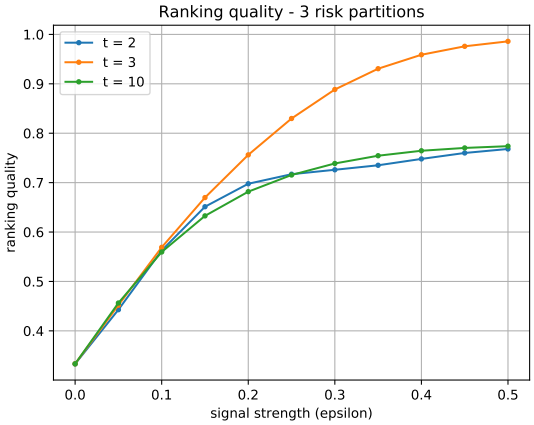
We compare the ranking quality for partition scans with $t = 2$, $t = 3$, and $t = 10$, for varying numbers of true partitions, in Figure 4. For two true partitions (Figure 4a), we observe that the partition scan with $t = 2$ converges to the two correct partitions, while the ranking quality of the $t = 3$ and $t = 10$ partition scans is reduced because a single true partition is split between multiple detected partitions. For three true partitions (Figure 4b), we observe that the ranking quality of the three methods is very similar for low ϵ values. For higher ϵ, the correct number of partitions ($t = 3$) outperforms $t = 2$ and $t = 10$, with near-perfect partitions for the highest ϵ values considered. For ten true partitions (Figure 4c), we again observe substantial improvement in ranking quality for the correct number of partitions ($t = 10$) as compared to $t = 3$ and $t = 2$.

Across these three scenarios, we observe that, when the number of partitions t is correctly specified, the partition scan converges to the correct partitioning of the data as signal strength ϵ increases. However, when the number of partitions t is misspecified, ranking quality is reduced either because a single true partition is split across multiple partitions, or because multiple true partitions are combined into a single partition. This is illustrated in Figure 5, which shows sample confusion matrices with $\epsilon = 0.5$ for the $t = 10$ case, when two true partitions exist, and the $t = 2$ case, when ten true partitions exist.

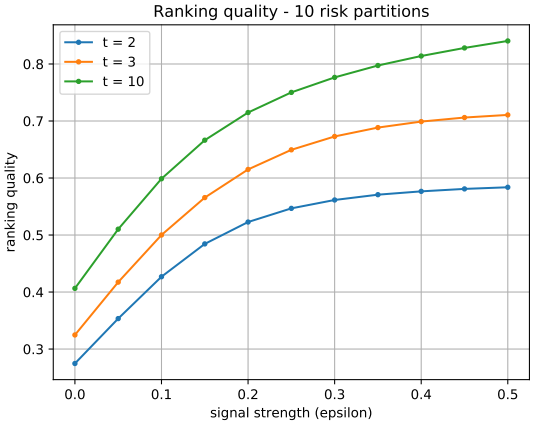
These examples demonstrate the importance of choosing the correct value of the number of partitions t. Thus we evaluate our efficient heuristic for choosing t, described in §7 above. We compute the mean and 95% confidence interval of the chosen t (across the 500 simulated datasets)



(a) Two true partitions

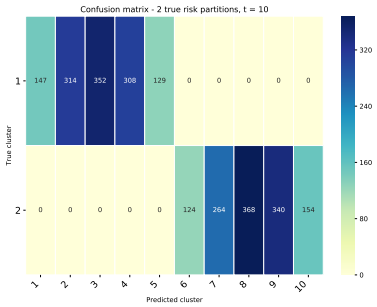


(b) Three true partitions

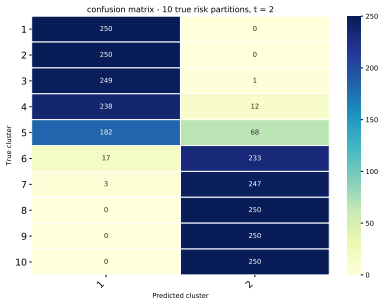


(c) Ten true partitions

Figure 4: Ranking quality for two, three, and ten true partitions, comparing partition scans with $t = 2$ (blue line), $t = 3$ (orange line), and $t = 10$ (green line) partitions.



(a) Two true partitions, $t = 10$



(b) Ten true partitions, $t = 2$

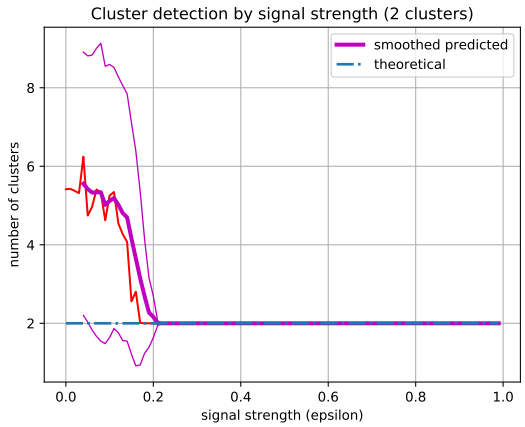
Figure 5: Sample confusion matrices illustrating under- and over-specification of the true number of partitions, with $\epsilon = 0.5$.

for the cases of two, three, five, and seven true partitions, as shown in Figure 6. We observe that, as the signal strength ϵ increases, our heuristic is able to reliably choose the correct number of partitions t. However, the signal strength required for convergence increases with the number of true partitions, from $\epsilon \approx 0.2$ for two true partitions, to $\epsilon \approx 0.9$ for seven true partitions.

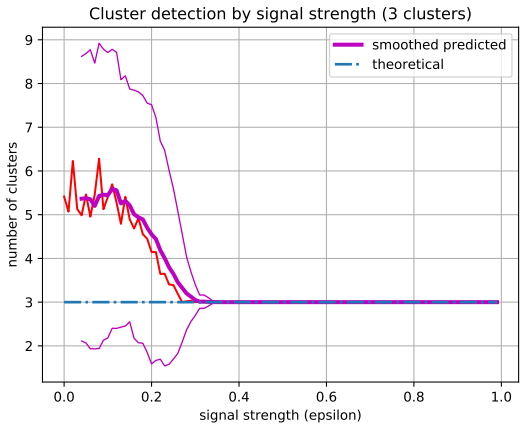
8.2 Multiple cluster detection simulations

For the second set of simulations, we evaluate the multiple cluster detection (MCD) scan statistic with Poisson score function, $F(P) = \sum_{j=2\ldots t} f(C_j, B_j)$, where $f(x, y) = x \log(\frac{x}{y}) + y - x$ if $x > y$ and 0 otherwise. We compare $t = 2$ and $t = 3$ partitions, where the $t = 2$ case corresponds to the expectation-based Poisson (EBP) scan statistic (Neill, 2012). We denote the $t = 2$ and $t = 3$ cases as EBP and MCD respectively. We also compare performance to Kulldorff's Poisson scan (i.e., partition scan with $t = 2$ partitions) and the partition scan with $t = 3$ partitions, which we denote as KULL and PART respectively.

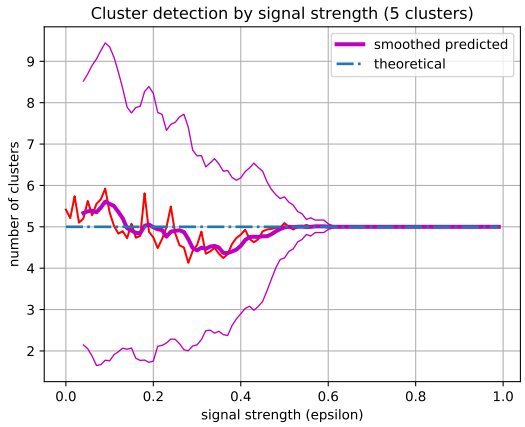
As in the risk partitioning experiments, we generate datasets consisting of $n = 2500$ data records, where each data record has baseline $y_i \sim \text{Poisson}(100)$ and count $x_i \sim \text{Poisson}(q_i y_i)$. We focus on the case where there are two true clusters to be detected, a primary cluster with relative risk q_1 and a secondary cluster with relative risk q_2, where $q_1 > q_2 > 1$. Each cluster consists of 10% of the data records, while the remaining 80% of records are generated with $q_i = 1$. As in the risk partitioning experiments, we consider a range of signal strengths ϵ, where the primary cluster



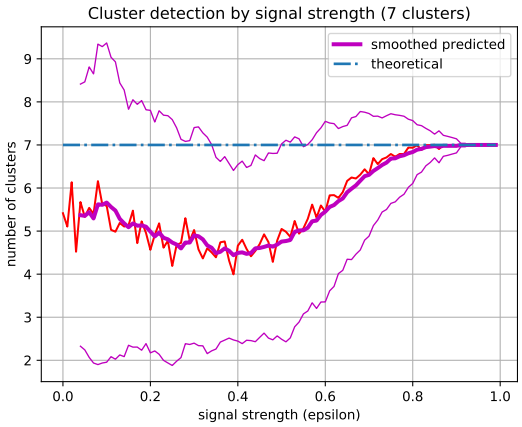
(a) Two true partitions



(b) Three true partitions



(c) Five true partitions



(d) Seven true partitions

Figure 6: Mean and 95% confidence interval of chosen number of partitions t for two, three, five, and seven true partitions, as a function of signal strength ϵ.

has relative risk $q_1 = 1 + \epsilon$. We consider three different scenarios: in the "low" scenario, the secondary cluster has relative risk $1 + \frac{\epsilon}{4}$, in the "medium" scenario, the secondary cluster has relative risk $1 + \frac{\epsilon}{2}$, and in the "high" scenario, the secondary cluster has relative risk $1 + \frac{3\epsilon}{4}$. For each scenario and each signal strength ϵ, we generated 500 simulated datasets. An additional 10000 datasets were generated under the null hypothesis H_0 ($q_i = 1$ for all data records), to be used for computing detection power. For each dataset and for each method (MCD, PART, EBP, and KULL), we compute the optimal partitioning $P^* = \arg\max_P F(P)$ and its score $F^* = F(P^*)$.

The methods were evaluated based on two criteria, *detection power*, which measures a method's ability to distinguish datasets generated under the alternative hypothesis from those generated under the null hypothesis of no clusters, and *detection accuracy*, which measures how well the clusters identified by the method correspond to the true clusters used to generate the data. To compute detection power, we first compute the threshold score needed for a method to detect at a fixed false positive rate of $\alpha = .05$. To do so, we compute the maximum scores F^* for 10000 datasets under the null hypothesis and use the 95th percentile of these scores as the detection threshold. We then compute the proportion of the 500 datasets (for a given scenario and a given signal strength ϵ) for which F^* exceeds the threshold.[3] Thus detection power is the statistical power (1 − Type II error) at a fixed false positive rate (Type I error) of 5%.

We compute three different pairwise measures of detection accuracy which together give a more complete picture of each method's ability to identify the primary and secondary clusters. Consider the detected partitions $S_1, \ldots, S_t$, ordered from smallest to largest relative risk. Then *primary cluster detection accuracy* is defined as the probability that, for a data record i from the primary cluster and an unaffected data record j, each selected uniformly at random, with corresponding partitions S_i and S_j, that $S_i > S_j$. Similarly, *secondary cluster detection accuracy* is defined as the probability that $S_i > S_j$ for a data record i from the secondary cluster and an unaffected data record j. Finally, *cluster differentiation accuracy* is defined as the probability that $S_i > S_j$ for data records i from the primary cluster and j from the secondary cluster. For $i = 1 \ldots t$, let $N_i^{(p)}$, $N_i^{(s)}$, and $N_i^{(u)}$ equal the number of records in the primary cluster, secondary cluster, and unaffected records respectively that are assigned to partition i. Then primary cluster detection

[3] For more accurate evaluation, we also average these detection proportions over 1000 bootstrapped samples of the detection threshold.

accuracy is defined as:

$$PCDA = \frac{\sum_{i=2\ldots t} \sum_{i'=1\ldots(i-1)} N_i^{(p)} N_{i'}^{(u)}}{\sum_{i=1\ldots t} \sum_{i'=1\ldots t} N_i^{(p)} N_{i'}^{(u)}}.$$

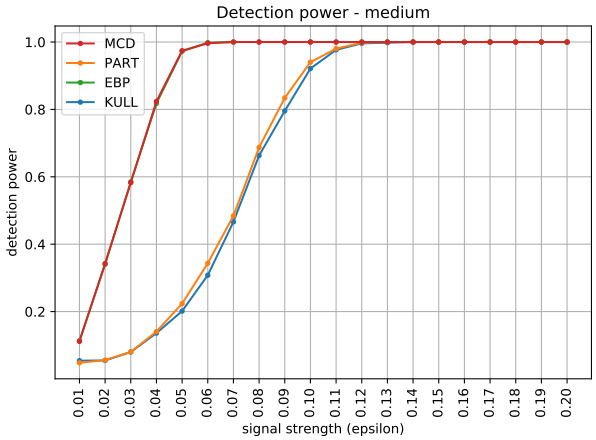
Secondary cluster detection accuracy and cluster differentiation accuracy are defined analogously. For $t = 3$ partitions, perfect detection—where all unaffected records are assigned to partition S_1, all records in the secondary cluster are assigned to partition S_2, and all records in the primary cluster are assigned to partition S_3—would score 1 for each of these three accuracy measures.

8.2.1 Multiple cluster detection simulation results

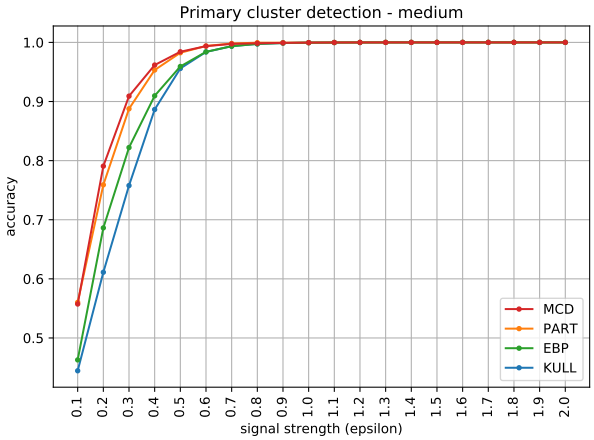
We first consider the "medium" scenario, for which the primary and secondary clusters have relative risks $1 + \epsilon$ and $1 + \frac{\epsilon}{2}$ respectively. Detection power, primary cluster detection accuracy, secondary cluster detection accuracy, and cluster differentiation accuracy for the MCD, PART, EBP, and KULL methods are shown in Figure 7.[4] For detection power, we observe that the cluster detection methods (MCD and EBP) have substantially higher detection power than the risk partitioning methods (PART and KULL). PART slightly outperforms KULL, while MCD and EBP have very similar detection power. For primary cluster detection accuracy, the methods with $t = 3$ partitions (MCD and PART) outperform the methods with $t = 2$ (EBP and KULL), and the cluster detection methods slightly outperform the risk partitioning methods. For secondary cluster detection accuracy, MCD and PART again outperform EBP and KULL. MCD outperforms PART across most of the range of ϵ values, while KULL slightly outperforms EBP. Finally, for cluster differentiation accuracy, we see a clear ordering of methods with MCD > PART > EBP > KULL. As the signal strength ϵ increases, MCD and PART correctly distinguish the primary cluster, the secondary cluster, and the remaining records, while EBP and KULL detect the primary and secondary cluster together as a single cluster, and thus, fail to distinguish primary from secondary.

Next we consider the "low" scenario, for which the primary and secondary clusters have relative risks $1 + \epsilon$ and $1 + \frac{\epsilon}{4}$ respectively. Detection power, primary cluster detection accuracy, secondary cluster detection accuracy, and cluster differentiation accuracy for the MCD, PART, EBP, and KULL methods are shown in Figure 8. We see the same general trends in detection power, primary

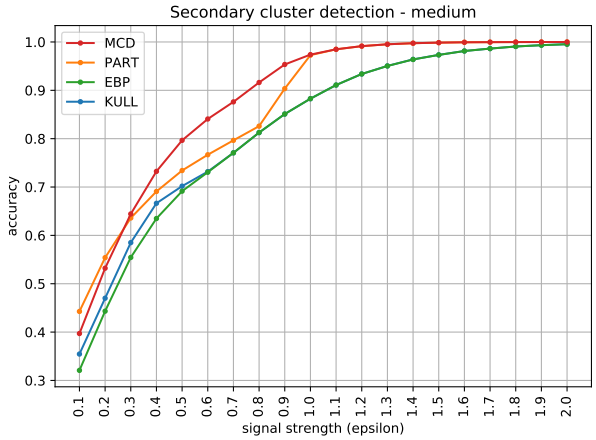
[4]Note the smaller x-axis ranges for Figures 7a, 8a, and 9a as compared to the rest of Figures 7-9. All methods had perfect detection power for $\epsilon > 0.2$, so we focus on the lower signal strengths for the detection power graphs.



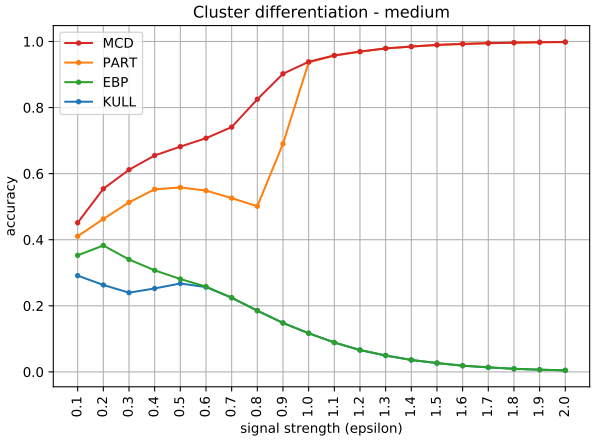
(a) Detection power



(b) Primary cluster detection accuracy

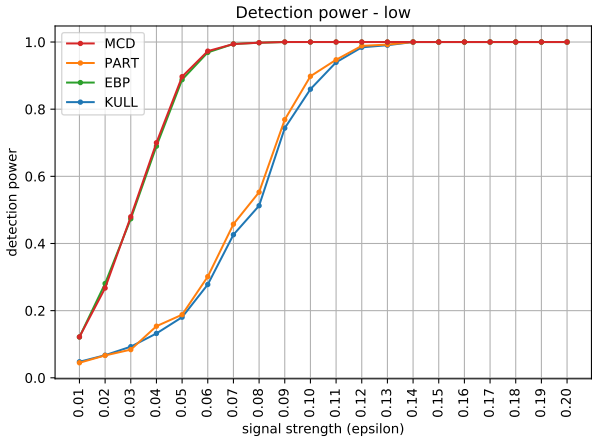


(c) Secondary cluster detection accuracy

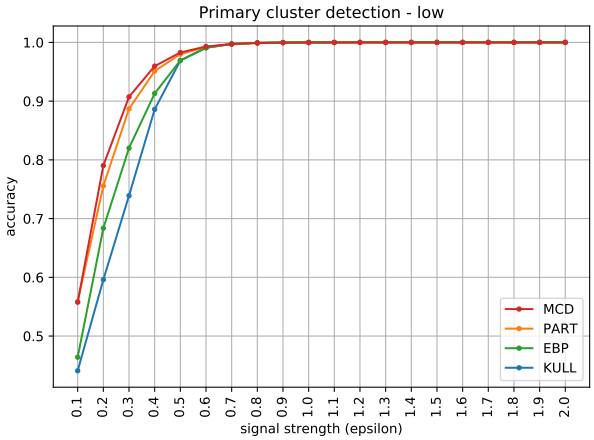


(d) Cluster differentiation accuracy

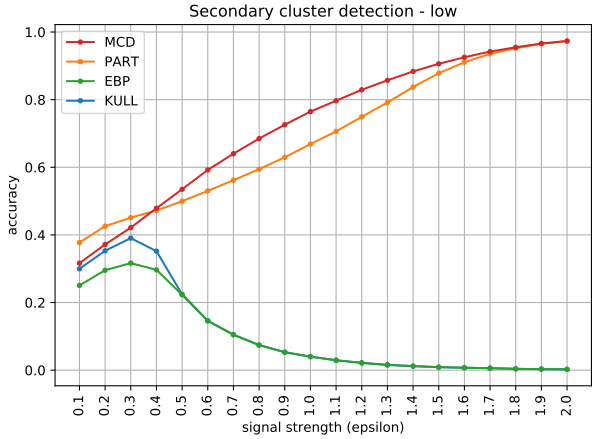
Figure 7: Multiple cluster detection with medium-strength secondary cluster. MCD (red line) is multiple cluster detection with $t = 3$ partitions, PART (orange line) is risk partitioning with $t = 3$, EBP (green line) is cluster detection with $t = 2$, and KULL (blue line) is risk partitioning with $t = 2$.



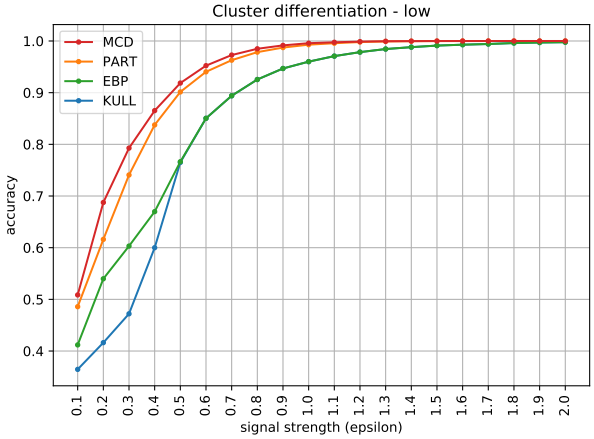
(a) Detection power



(b) Primary cluster detection accuracy

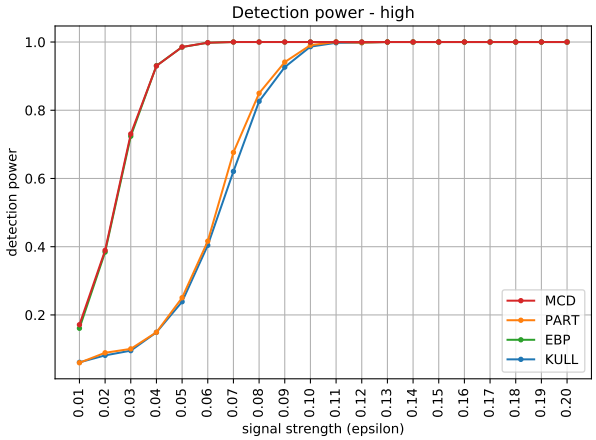


(c) Secondary cluster detection accuracy

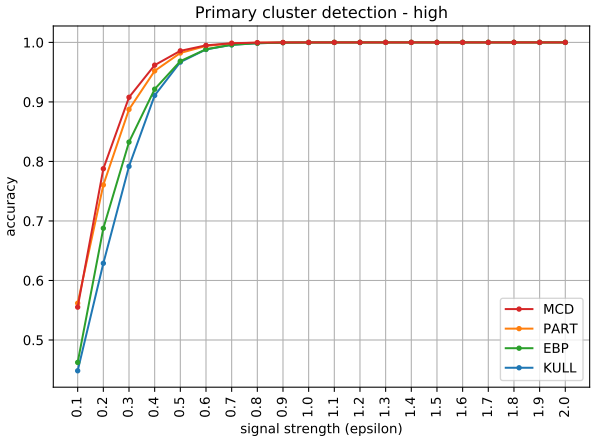


(d) Cluster differentiation accuracy

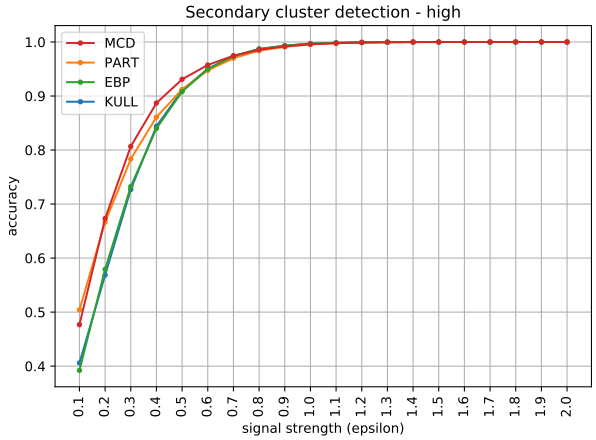
Figure 8: Multiple cluster detection with low-strength secondary cluster. MCD (red line) is multiple cluster detection with $t = 3$ partitions, PART (orange line) is risk partitioning with $t = 3$, EBP (green line) is cluster detection with $t = 2$, and KULL (blue line) is risk partitioning with $t = 2$.



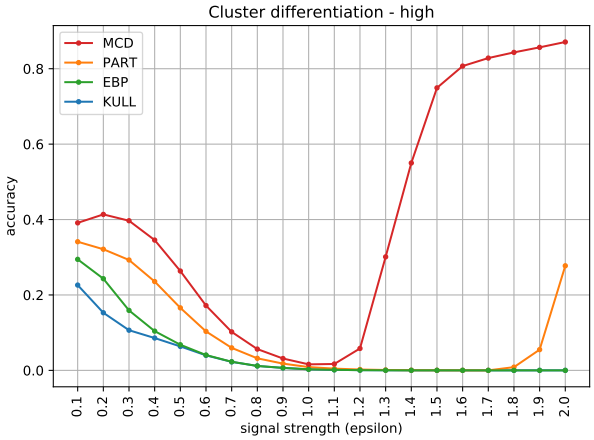
(a) Detection power



(b) Primary cluster detection accuracy



(c) Secondary cluster detection accuracy



(d) Cluster differentiation accuracy

Figure 9: Multiple cluster detection with high-strength secondary cluster. MCD (red line) is multiple cluster detection with $t = 3$ partitions, PART (orange line) is risk partitioning with $t = 3$, EBP (green line) is cluster detection with $t = 2$, and KULL (blue line) is risk partitioning with $t = 2$.

cluster detection accuracy, and cluster differentiation accuracy as in the medium case. One notable difference is in secondary cluster detection accuracy: as signal strength increases, the methods with $t = 2$ partitions (EBP and KULL) only detect the primary cluster and fail to detect the secondary cluster, while the methods with $t = 3$ (MCD and PART) correctly distinguish the primary cluster, the secondary cluster, and the remaining records.

Finally, we consider the "high" scenario, for which the primary and secondary clusters have relative risks $1 + \epsilon$ and $1 + \frac{3\epsilon}{4}$ respectively. Detection power, primary cluster detection accuracy, secondary cluster detection accuracy, and cluster differentiation accuracy for the MCD, PART, EBP, and KULL methods are shown in Figure 9. We see the same trends in detection power and cluster detection accuracy as in the medium case. One notable difference is in cluster differentiation accuracy. As in the medium case, the methods with $t = 2$ partitions (KULL and EBP) detect the primary and secondary cluster together as a single cluster, and thus, fail to distinguish primary from secondary. However, the methods with $t = 3$ (MCD and PART) also have difficulty distinguishing primary from secondary: for risk partitioning with $t = 3$, even for high signal strengths, the primary and secondary clusters are detected as a single cluster in the third (highest-risk) partition, while a number of unaffected records are incorrectly detected as a "false positive" cluster in the second (medium-risk) partition. For multiple cluster detection with $t = 3$, this problem occurs for medium signal strengths ($\epsilon \approx 1$), while for high signal strengths ($\epsilon \approx 2$), MCD is able to correctly distinguish the primary cluster, secondary cluster, and unaffected records.

Thus, across these three scenarios, we observe that multiple cluster detection with $t = 3$ (MCD) outperforms risk partitioning with $t = 3$ (PART) as well as the methods with $t = 2$ (EBP and KULL). Moving from single to multiple cluster detection (i.e., moving from $t = 2$ to $t = 3$) consistently improves detection accuracy and ability to differentiate between the primary and secondary clusters, while the single cluster detection methods will either fail to detect the secondary cluster (if its signal is much weaker than the primary cluster) or combine the primary and secondary clusters into a single detected cluster. Additionally, while the partition scan performs adequately for the multiple cluster detection task, the multiple cluster detection scan consistently outperforms PART in terms of both detection power and detection accuracy.

To confirm these results, we perform an additional set of simulation runs, based on the

"medium-strength secondary cluster" scenario. Instead of fixing each cluster size at 10% of the data records, and allowing the signal strength ϵ to vary, we fix $\epsilon = 0.5$ and allow cluster sizes to vary between 1% and 20% of the data records, as shown in Figure 10. As expected, smaller cluster size, like weaker signal strength, reduces the performance of all methods, while keeping the relative performance of methods unchanged. We observe that MCD outperforms PART, EBP, and KULL in terms of both detection power and accuracy. PART has higher primary cluster detection accuracy, secondary cluster detection accuracy, and cluster differentiation accuracy than EBP and KULL, while EBP has higher detection power than PART for small cluster sizes.

9 Empirical study - cancer incidence

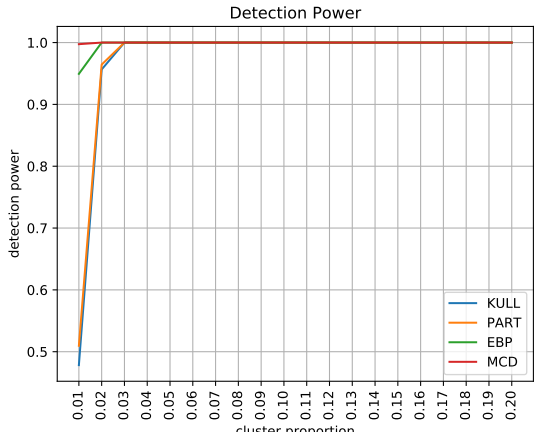
We apply the multiple cluster detection scan approach to identify high-risk cancer clusters in New York City, using cancer incidence data from 2013-2017, grouped by census tract. Data were obtained from the New York State Cancer Registry (New York State Dept. of Health, 2021), including the observed count (number of cases) and the expected count for the three most common cancer types (prostate, lung, and breast) for each tract during the five-year time period. Expected counts were population- and age-adjusted, using the statewide cancer rate as a baseline. In this way we identify regions of high cancer risk and possibly correlate them with a common feature: demographic, socio-economic, geographic, etc. We use the approach described in §7 to choose the number of partitions t between $t = 1$ and $t = 20$, obtaining $t = 3$, $t = 4$, and $t = 5$ (i.e., 2, 3, and 4 detected clusters) for lung, breast, and prostate cancers respectively. The results are shown in Figures 11a, 11b, and 11c respectively, where darker shades correspond to partitions with higher relative risks q. We observe that all three cancer types cluster spatially, and that much of this clustering can be explained by neighborhood demographic characteristics. For example, spatial clustering is most evident for prostate cancer, which is known to afflict African-American males at a much higher rate than white males (New York State Dept. of Health, 2021). We also compare the results to single cluster detection ($t = 2$) for each cancer type; we show prostate cancer as an example in Figure 11d. We observe that the higher numbers of partitions t provide a finer-grained (and thus, more informative) gradation of risk as compared to the single cluster detection ($t = 2$) case, distinguishing between clusters with a higher and lower degree of elevated

risk. For example, for prostate cancer, cluster relative risks varied from $q = 1.21$ to $q = 2.43$, as compared to $q = 1.70$ for the single cluster detection approach. Additionally, as expected, the chosen t values more closely fit the observed data than $t = 2$, as measured by total log-likelihood: LL = -5095 vs -5205 for lung cancer, -5322 vs. -5441 for breast cancer, and -5580 vs. -5786 for prostate cancer.

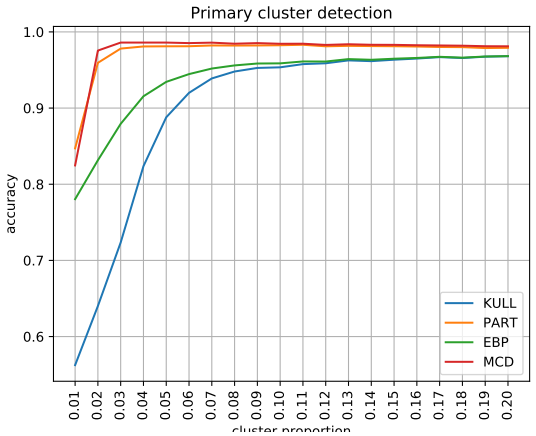
10 Conclusions

The empirical results shown above demonstrate several advantages of generalizing the spatial and subset scan statistics to multiple partitions of the data: increased detection power and accuracy for multiple clusters, as well as the ability to accurately capture finer-grained gradations in risk. The practical utility of our risk partitioning and multiple cluster detection scan statistics was enabled by three novel methodological contributions: (i) formulating these partition scan statistics as generalizations of the population-based and expectation-based scan statistics respectively; (ii) identifying sufficient conditions under which the optimal partitioning is guaranteed to be consecutive; and (iii) developing a new dynamic programming algorithm for identification of the optimal consecutive partitioning in quadratic time. Together, these contributions enable exact and efficient solutions to large-scale risk partitioning and multiple cluster detection problems for any partition scan statistic satisfying the Consecutive Partitions Property. Further, since CPP generalizes the linear-time subset scanning property from $t = 2$ to $t \geq 2$ partitions of the data, our novel, geometric proof of CPP also provides a new approach to proving LTSS.

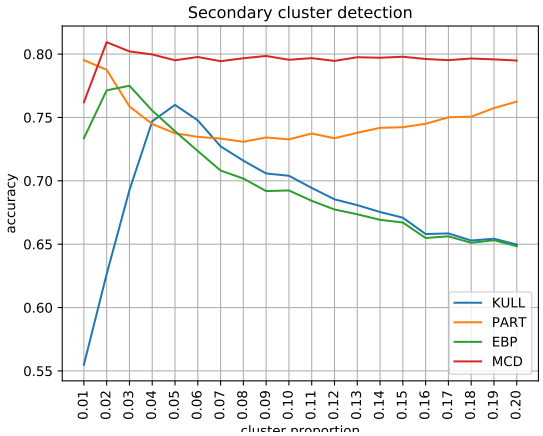
One important limitation of the present approach, which we will believe will provide a fertile ground for future work, is that it only solves the "best unconstrained partitioning" problem, while for real-world problems it is desirable to include constraints, such as spatial proximity or connectivity, on the detected subsets or partitions of the data. We believe that, just as the LTSS property has proved to be a useful building block for solving constrained pattern detection problems (Neill, 2012; Speakman et al., 2015), CPP will enable scalable solutions to a variety of constrained partitioning problems. However, significant work remains in defining constraints or penalties which encode desirable features of a partitioning while still preserving the efficient and exact optimization guaranteed by CPP.



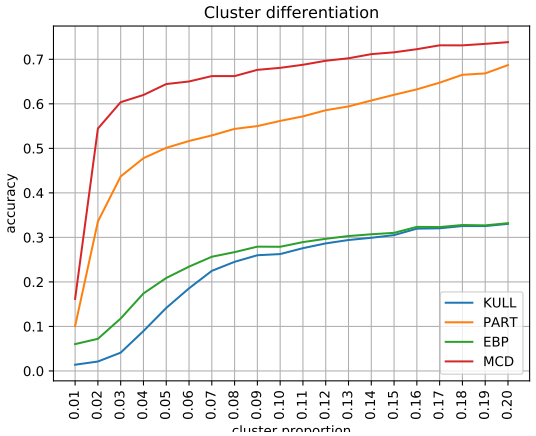
(a) Detection power



(b) Primary cluster detection accuracy

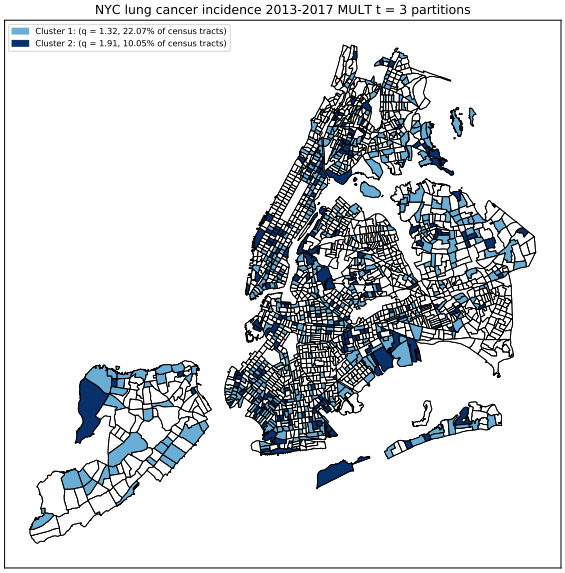


(c) Secondary cluster detection accuracy

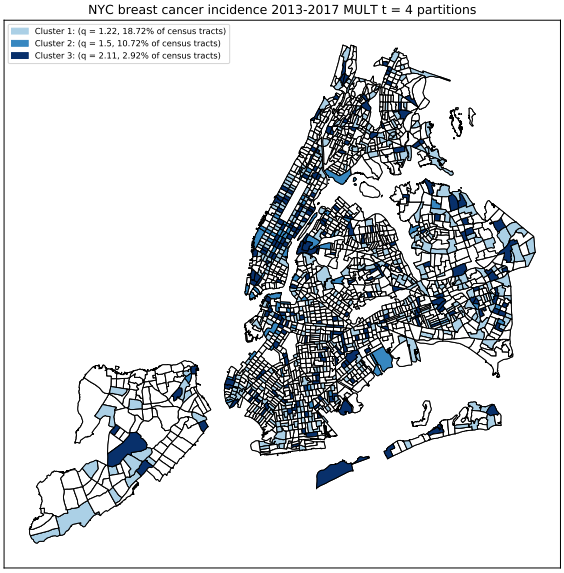


(d) Cluster differentiation accuracy

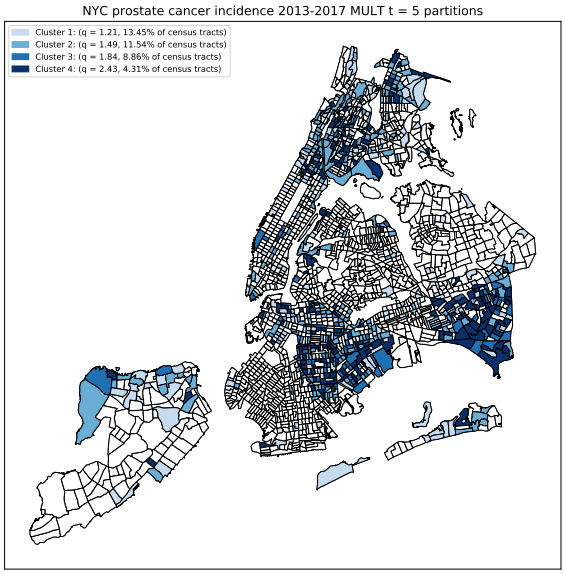
Figure 10: Multiple cluster detection with medium-strength secondary cluster, fixed signal strength of $\epsilon = 0.5$, and varying cluster size. MCD (red line) is multiple cluster detection with $t = 3$ partitions, PART (orange line) is risk partitioning with $t = 3$, EBP (green line) is cluster detection with $t = 2$, and KULL (blue line) is risk partitioning with $t = 2$.



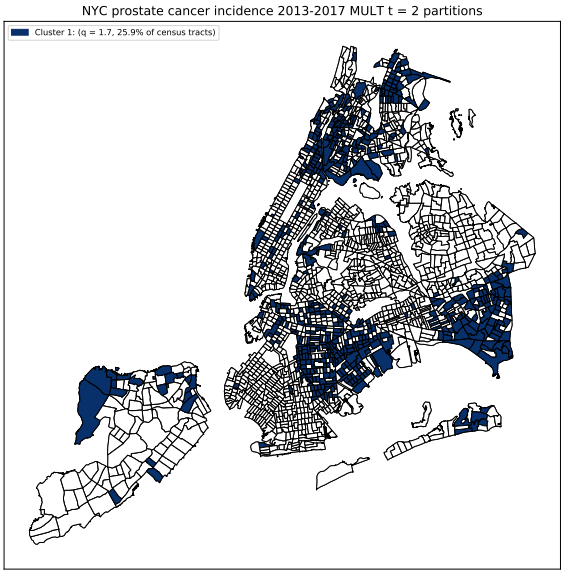
(a) Lung cancer, $t = 3$ partitions



(b) Breast cancer, $t = 4$ partitions



(c) Prostate cancer, $t = 5$ partitions



(d) Prostate cancer, $t = 2$ partitions

Figure 11: New York City cancer incidence clusters by census tract, 2013-2017, using the multiple cluster detection scan with chosen numbers of partitions ((a)-(c)) and $t = 2$ partitions ((d)).

References

Francis Bach. *Learning with Submodular Functions: A Convex Optimization Perspective*. Foundations and Trends in Machine Learning. Now Publishers, 2013.

Gruia Calinescu, Chandra Chekuri, Martin Pál, and Jan Vondrák. Maximizing a submodular set function subject to a matroid constraint. In *Integer Programming and Combinatorial Optimization, Proceedings of IPCO 2007*, pages 182–196. 2007.

Amiya K. Chakravarty, James B. Orlin, and Uriel G. Rothblum. A partitioning problem with additive objective with an application to optimal inventory groupings for joint replenishment. *Operations Research*, 30(5):1018–1022, 1982.

Yuval Filmus and Justin Ward. A tight combinatorial algorithm for submodular maximization subject to a matroid constraint. In *Proc. 53rd Annual IEEE Symposium on Foundations of Computer Science*, pages 659–668. 2012.

Uriel Freige, Vahab S. Mirrokni, and Jan Vondrak. Maximizing non-monotonic submodular functions. *SIAM Journal on Computing*, 40(4):1133–1153, 2011.

Joseph Glaz, Joseph Naus, and Sylvan Wallenstein. *Scan Statistics*. Springer, New York, 2001.

R. L. Graham. An efficient algorithm for determining the convex hull of a finite planar set. *Information Processing Letters*, 1(4):132–133, 1972.

M. Grötschel, L. Lovasz, and A. Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1(2):169–197, 1981.

Martin Kulldorff. A spatial scan statistic. *Communications in Statistics: Theory and Methods*, 26 (6):1481–1496, 1997.

Martin Kulldorff and Neville Nagarwalla. Spatial disease clusters: detection and inference. *Statistics in Medicine*, 14:799–810, 1995.

Martin Kulldorff, Zixing Fang, and Stephen J. Walsh. A tree-based scan statistic for database disease surveillance. *Biometrics*, 59(2):323–331, 2003.

A. Lawson, A. Biggeri, D. Bohning, E. Lesare, J. F. Viel, and R. Bertollini, editors. *Disease Mapping and Risk Assessment for Public Health*. Wiley, Chichester, 1999.

X. Z. Li, J. F. Wang, W. Z. Yang, Z. J. Li, and S. J. Lai. A spatial scan statistic for multiple clusters. *Mathematical Biosciences*, 233:135–142, 2011.

Joseph Naus. The distribution of the size of maximum cluster of points on a line. *Journal of the American Statistical Association*, 60:523–538, 1965a.

Joseph Naus. Clustering of random points in two dimensions. *Biometrika*, 52:263–267, 1965b.

D. B. Neill, A. W. Moore, and G. F. Cooper. A Bayesian spatial scan statistic. In *Advances in Neural Information Processing Systems 18*, pages 1003–1010. 2006.

Daniel B. Neill. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society B*, 74(2):337–360, 2012.

New York State Dept. of Health. New York State Cancer Registry and Cancer Statistics, 2021. URL `http://www.health.ny.gov/statistics/cancer/registry/`.

S. Speakman, E. McFowland III, and D. B. Neill. Scalable detection of anomalous patterns with connectivity constraints. *Journal of Computational and Graphical Statistics*, 24(4):1014–1033, 2015. doi: 10.1080/10618600.2014.960926.

K. Takahashi and H. Shimadzu. Detecting multiple spatial disease clusters: information criterion and scan statistic approach. *International Journal of Health Geographics*, 19:33, 2020. doi: 10.1186/s12942-020-00228-y.

Jan Vondrák. Optimal approximation for the submodular welfare problem in the value oracle model. In *Proc. 40th ACM Symposium on Theory of Computing*, pages 67–74. 2008.

Zhenkui Zhang, Renato Assuncao, and Martin Kulldorff. Spatial scan statistics adjusted for multiple clusters. *Journal of Probability and Statistics*, 10, 2010.