Identifying Consistent Statements about Numerical Data with Dispersion-Corrected Subgroup Discovery

Mario Boley∗†
mboley@mmci.uni-saarland.de

Bryan R. Goldsmith†
goldsmith@fhi-berlin.mpg.de

Luca M. Ghiringhelli‡
ghiringhelli@fhi-berlin.mpg.de

Jilles Vreeken∗
jilles@mpi-inf.mpg.de

Abstract

Existing algorithms for subgroup discovery with numerical targets do not optimize the error or target variable dispersion of the groups they find. This often leads to unreliable or inconsistent statements about the data, rendering practical applications, especially in scientific domains, futile. Therefore, we here extend the optimistic estimator framework for optimal subgroup discovery to a new class of objective functions: we show how tight estimators can be computed efficiently for all functions that are determined by subgroup size (non-decreasing dependence), the subgroup median value, and a dispersion measure around the median (non-increasing dependence). In the important special case when dispersion is measured using the average absolute deviation from the median, this novel approach yields a linear time algorithm. Empirical evaluation on a wide range of datasets shows that, when used within branch-and-bound search, this approach is highly efficient and indeed discovers subgroups with much smaller errors.

1. Introduction

Subgroup discovery is a well-established KDD technique (Klösgen (1996); Friedman and Fisher (1999); Bay and Pazzani (2001); see Atzmueller (2015) for a recent survey) with applications, e.g., in Medicine (Schmidt et al 2010), Social Science (Grosskreutz et al 2010).
Figure 1. To gain an understanding of the contribution of long-range van der Waals interactions (y-axis; above) to the total energy (x-axis; above) of gas-phase gold nanoclusters, subgroup discovery is used to analyze a dataset of such clusters simulated ab initio by density functional theory (Goldsmith et al, 2017; in press); available features describe nanocluster geometry and contain, e.g., number of atoms $a$, fraction of atoms with $i$ bonds $c_i$, and radius of gyration $r$. Here, similar to other scientific scenarios, a subgroup constitutes a useful piece of knowledge if it conveys a statement about a remarkable amount of van der Waals energy (captured by the group’s central tendency) with high consistency (captured by the group’s dispersion/error); optimal selector $\sigma_0$ with standard objective has high error and contains a large fraction of gold nanoclusters with a target value below the global median (0.13) (a); this is not the case for selector $\sigma_1$ discovered through dispersion-corrected objective (b), which therefore can be more consistently stated to describe gold nanoclusters with high van der Waals energy.

2010), and Materials Science (Goldsmith et al 2017; in press). In contrast to global modeling, which is concerned with the complete characterization of some variable defined for a given population, subgroup discovery aims to detect intuitive descriptions or selectors of subpopulations in which, locally, the target variable takes on a useful distribution. In scientific domains, like the ones mentioned above, such local patterns are typically considered useful if they are not too specific (in terms of subpopulation size) and indicate insightful facts about the underlying physical process that governs the target variable. Such facts could for instance be: ‘patients of specific demographics experience a low response to some treatment’ or ‘materials with specific atomic composition exhibit a high thermal conductivity’. For numeric (metric) variables, subgroups need to satisfy two criteria to truthfully represent such statements: the local distribution of the target variable must have a shifted central tendency (effect), and group members must be described well by that shift (consistency). The second requirement is captured by the group’s dispersion, which determines the average error of associating group members with the central tendency value (see also Song et al 2016).

Despite all three parameters—size, central tendency, and dispersion—being impor-
tant, the only known approach for the efficient discovery of globally optimal subgroups, branch-and-bound search [Webb 1995, Wrobel 1997], is restricted to objective functions that only take into account size and central tendency. That is, if we denote by \( Q \) some subpopulation of our global population \( P \) then the objective functions \( f \) currently available to branch-and-bound can be written as

\[
f(Q) = g(|Q|, c(Q))
\]  

(1)

where \( c \) is some measure of central tendency (usually average or median) and \( g \) is a function that is monotonically increasing in the subpopulation size \(|Q|\). A problem with all such functions is that they inherently favor larger groups with scattered target values over smaller more focused groups with the same central tendency. That is, they favor the discovery of inconsistent statements over consistent ones—surprisingly often identifying groups with a local error that is almost as high or even higher than the global error (see Fig. 1 for an illustration of this problem that abounded from the authors’ research in Materials Science). Although dispersion-corrected objective functions that counter-balance size by dispersion have been proposed (e.g., ‘t-score’ by Klösgen, 2002 or ‘mmad’ by Pieters et al, 2010), it remained unclear how to employ such functions outside of heuristic optimization frameworks such as greedy beam search (Lavrač et al, 2004) or selector sampling (Boley et al, 2012, Li and Zaki, 2016). Despite often finding interesting groups, such frameworks do not guarantee the detection of optimal results, which can not only be problematic for missing important discoveries but also because they therefore can never guarantee the absence of high quality groups—which often is an insight equally important as the presence of a strong pattern. For instance, in our example in Fig. 1 it would be remarkable to establish that long-range interactions are to a large degree independent of nanocluster geometry.

Therefore, in this paper (Sec. 3), we extend branch-and-bound search to objective functions of the form

\[
f(Q) = g(|Q|, \text{med}(Q), d(Q))
\]  

(2)

where \( g \) is monotonically increasing in the subpopulation size, monotonically decreasing in any dispersion measure \( d \) around the median, and, besides that, depends only (but in arbitrary form) on the subpopulation median. This involves developing an efficient algorithm for computing the tight optimistic estimator given by the optimal value of the objective function among all possible subsets of target values:

\[
\hat{f}(Q) = \max \{f(R): R \subseteq Q \},
\]  

(3)

which has been shown to be a crucial ingredient for the practical applicability of branch-and-bound (Grosskreutz et al, 2008, Lemmerich et al, 2016). So far, the most general approach to this problem (first codified in Lemmerich et al (2016); generalized here in Sec. 3.1) is to maintain a sorted list of target values throughout the search process and then to compute Eq. (3) as the maximum of all subsets \( R_i \subseteq Q \) that contain all target values of \( Q \) down to target value \( i \)—an algorithm that does not generalize to objective functions depending on dispersion. This paper presents an alternative idea (Sec. 3.2).
where we do not fix the size of subset \( R_i \) as in the previous approach but instead fix its median to target value \( i \). It turns out that this suffices to efficiently compute the tight optimistic estimator for all objective functions of the form of Eq. (2). Moreover, we end up with a linear time algorithm (Sec. 3.3) in the important special case where the dependence on size and dispersion is determined by the dispersion-corrected coverage defined by

\[
dcc(Q) = \frac{|Q|}{|P|} \max \left\{ 1 - \frac{\text{amd}(Q)}{\text{amd}(P)}, 0 \right\}
\]

where \( \text{amd} \) denotes the average absolute deviation from the median. This is the same computational complexity as the objective function itself. Consequently, this new approach can discover subgroups according to a more refined selection criterion without increasing the worst-case computational cost. Additionally, as demonstrated by empirical results on a wide range of datasets (Sec. 4), it is also highly efficient and successfully reduces the error of result subgroups in practice.

2. Subgroup Discovery

Before developing the novel approach to tight optimistic estimator computation, we recall in this section the necessary basics of optimal subgroup discovery with numeric target attributes. We focus on concepts that are essential from the optimization point of view (see, e.g., Duivesteijn and Knobbe 2011 and references therein for statistical considerations). As notional convention, we are using the symbol \([m]\) for a positive integer \( m \) to denote the set of integers \( \{1, \ldots, m\} \). Also, for a real-valued expression \( x \) we write \((x)_+\) to denote \( \max\{x, 0\}\).

2.1. Description languages, objective functions, and closed selectors

Let \( P \) denote our given global population of entities, for each of which we know the value of a real target variable \( y : P \rightarrow \mathbb{R} \) and additional descriptive information that is captured in some abstract description language \( L \) of subgroup selectors \( \sigma : P \rightarrow \{\text{true}, \text{false}\} \). Each of these selectors describes a subpopulation \( \text{ext}(\sigma) \subseteq P \) defined by

\[
\text{ext}(\sigma) = \{ p \in P : \sigma(p) = \text{true} \}
\]

that is referred to as the extension of \( \sigma \). Subgroup discovery is concerned with finding descriptions \( \sigma \in L \) that have a useful (or interesting) distribution of target values in their extension \( y_\sigma = \{ y(p) : p \in \text{ext}(\sigma) \} \). This notion of usefulness is given by an objective function \( f : L \rightarrow \mathbb{R} \). That is, the formal goal is to find elements \( \sigma \in L \) with maximal \( f(\sigma) \). Since we assume \( f \) to be a function of the multiset of \( y \)-values, let us define \( f(\sigma) = f(\text{ext}(\sigma)) = f(y_\sigma) \) to be used interchangeably for convenience. One example of a commonly used objective function is the impact measure \( \text{ipa} \) (see Webb 2001 here a scaled but order-equivalent version is given) defined by

\[
\text{ipa}(Q) = \text{cov}(Q) \left( \frac{\text{avg}(Q) - \text{avg}(P)}{\text{max}(P) - \text{avg}(P)} \right)_+.
\]

(4)
where \( \text{cov}(Q) = \frac{|Q|}{|P|} \) denotes the **coverage** or relative size of \( Q \) (here—and wherever else convenient—we identify a subpopulation \( Q \subseteq P \) with the multiset of its target values).

The standard description language in the subgroup discovery literature\(^1\) is the language \( \mathcal{L}_{cnj} \) consisting of **logical conjunctions** of a number of base propositions (or predicates). That is, \( \sigma \in \mathcal{L}_{cnj} \) are of the form

\[
\sigma(\cdot) \equiv \pi_{i_1}(\cdot) \land \cdots \land \pi_{i_l}(\cdot)
\]

where the \( \pi_{i_j} \) are taken from a pool of **base propositions** \( \Pi = \{\pi_1, \ldots, \pi_k\} \). These propositions usually correspond to equality or inequality constraints with respect to one variable \( x \) out of a set of description variables \( \{x_1, \ldots, x_n\} \) that are observed for all population members (e.g., \( \pi(p) \equiv x(p) \geq v \)). However, for the scope of this paper it is sufficient to simply regard them as abstract Boolean functions \( \pi : P \rightarrow \{ \text{true}, \text{false} \} \).

In this paper, we focus in particular on the refined language of **closed conjunctions** \( \mathcal{C}_{cnj} \subseteq \mathcal{L}_{cnj} \) (Pasquier et al, 1999), which is defined as \( \mathcal{C}_{cnj} = \{ \sigma \in \mathcal{L}_{cnj} : c(\sigma) = \sigma \} \) by the fixpoints of the **closure operation** \( c : \mathcal{L}_{cnj} \rightarrow \mathcal{L}_{cnj} \) given by

\[
c(\sigma) = \bigwedge \{ \pi \in \Pi : \text{ext}(\pi) \supseteq \text{ext}(\sigma) \} . \tag{5}
\]

These are selectors to which no further proposition can be added without reducing their extension, and it can be shown that \( \mathcal{C}_{cnj} \) contains at most one selector for each possible extension. While this can reduce the search space for finding optimal subgroups by several orders of magnitude, closed conjunctions are the longest (and most redundant) description for their extension and thus do not constitute intuitive descriptions by themselves. Hence, for reporting concrete selectors (as in Fig. 1), closed conjunctions have to be simplified to selectors of approximately minimum length that describe the same extension (Boley and Grosskreutz, 2009).

### 2.2. Branch-and-bound and optimistic estimators

The standard algorithmic approach for finding optimal subgroups with respect to a given objective function is branch-and-bound search—a versatile algorithmic puzzle solving framework with several forms and flavors (see, e.g., Mehlhorn and Sanders, 2008, Chap. 12.4). At its core, all of its variants assume the availability and efficient computability of two ingredients:

1. A **refinement operator** \( r : \mathcal{L} \rightarrow 2^\mathcal{L} \) that is monotone, i.e., for \( \sigma, \varphi \in \mathcal{L} \) with \( \varphi \in r(\sigma) \) it holds that \( \text{ext}(\varphi) \subseteq \text{ext}(\sigma) \), and that non-redundantly generates \( \mathcal{L} \).
   That is, there is a root selector \( \bot \in \mathcal{L} \) such that for every \( \sigma \in \mathcal{L} \) there is a unique sequence of selectors \( \bot = \sigma_0, \sigma_1, \ldots, \sigma_l = \sigma \) with \( \sigma_i \in r(\sigma_{i-1}) \). In other words, the refinement operator implicitly represents a directed tree (arborescence) on the description language \( \mathcal{L} \) rooted in \( \bot \).

---

\(^1\)In this article we remain with this basic setting for the sake of simplicity. It is, however, useful to note that several generalizations of this concept have been proposed (e.g., Parthasarathy et al, 1999; Huan et al, 2003), to which the contributions of this paper remain applicable.
Bst-BB(\mathcal{F}, \sigma): \quad \text{// } \mathcal{F} \text{ max. priority queue w.r.t. } \hat{f}, \sigma \text{ current } f \text{-maximizer} \\
\text{begin} \\
\text{if } \mathcal{F} = \emptyset \text{ or } \hat{f}(\text{top}(\mathcal{F}))/f(\sigma) \leq a \text{ then} \\
\quad \text{return } \sigma \\
\text{else} \\
\quad \mathcal{R} = r(\text{top}(\mathcal{F})) \quad \text{// refinement of } \hat{f} \text{-maximizer in queue} \\
\quad \sigma' = \arg\max(f(\varphi) : \varphi \in \{\sigma \} \cup \mathcal{R}) \\
\quad \mathcal{F}' = (\mathcal{F} \setminus \{\text{top}(\mathcal{F})\}) \cup \{\varphi \in \mathcal{R} : \hat{f}(\varphi)/f(\sigma') \geq a\} \\
\quad \text{return } \text{Bst-BB}(\mathcal{F}', \sigma') \\
\text{end} \\
\sigma^* = \text{Bst-BB}(\{\bot\}, \bot) \quad \text{// call with root element to find global solution} \\
\text{end}

Algorithm 1: Best-first branch-and-bound that finds $a$-approximation to objective function $f$ based on refinement operator $r$ and optimistic estimator $\hat{f}$; depth-limit and multiple solutions (top-$k$) parameters omitted; top denotes the find max operation for priority queue.

2. An optimistic estimator (or bounding function) $\hat{f}: \mathcal{L} \to \mathbb{R}$ that bounds from above the attainable subgroup value of a selector among all more specific selectors, i.e., it holds that $\hat{f}(\sigma) \geq f(\varphi)$ for all $\varphi \in \mathcal{L}$ with $\text{ext}(\varphi) \subseteq \text{ext}(\sigma)$.

Based on these ingredients, a branch-and-bound algorithm simply enumerates all elements of $\mathcal{L}$ starting from $\bot$ using $r$ (branch), but—based on $\hat{f}$—avoids expanding descriptions that cannot yield an improvement over the best subgroups found so far (bound). Depending on the order in which language elements are expanded, one distinguishes between depth-first, breadth-first, breadth-first iterating deepening, and best-first search. In the last variant, the optimistic estimator is not only used for pruning the search space, but also to select the next element to be expanded, which is particularly appealing for informed, i.e., tight, optimistic estimators. An important feature of branch-and-bound is that it effortlessly allows to speed-up the search in a sound way by relaxing the result requirement from being $f$-optimal to just being an $a$-approximation. That is, the found solution $\sigma$ satisfies for all $\sigma' \in \mathcal{L}$ that $f(\sigma)/f(\sigma') \geq a$ for some approximation factor $a \in (0, 1]$. The pseudo-code given in Alg. 1 summarizes all of the above ideas. Note that, for the sake of clarity, we omitted here some other common parameters such as a depth-limit and multiple solutions (top-$k$), which are straightforward to incorporate (see Lemmerich et al, 2016).

An efficiently computable refinement operator has to be constructed specifically for the desired description language. For example for the language of conjunctions $\mathcal{L}_{\text{conj}}$, one can define $r_{\text{conj}}: \mathcal{L}_{\text{conj}} \to \mathcal{L}_{\text{conj}}$ by

\[ r_{\text{conj}}(\sigma) = \{\sigma \land \pi_i : \max\{j : \pi_j \in \sigma\} < i \leq k\} \]
where we identify a conjunction with the set of base propositions it contains. For the closed conjunctions \( cnj \), let us define the lexicographical prefix of a conjunction \( \sigma \in L^{cnj} \) and a base proposition index \( i \in [k] \) as \( \sigma|_i = \sigma \cap \{\pi_1, \ldots, \pi_i\} \). Moreover, let us denote with \( i(\sigma) \) the minimal index such that the \( i \)-prefix of \( \sigma \) is extension-preserving, i.e.,

\[
    i(\sigma) = \min\{i : \text{ext}(\sigma|_i) = \text{ext}(\sigma)\}.
\]

That is, a selector \( \phi \) is among the refinements of \( \sigma \) if \( \phi \) can be generated by an application of the closure operator given in Eq. (5) that is prefix-preserving.

How to obtain an optimistic estimator for an objective function of interest depends on the definition of that objective. For instance, the coverage function \( \text{cov} \) is a valid optimistic estimator for the impact function \( \text{ipa} \) as defined in Eq. (4), because the second factor of the impact function is upper bounded by 1. In fact there are many different optimistic estimators for a given objective function. Clearly, the smaller the value of the bounding function for a candidate subpopulation, the higher is the potential for pruning the corresponding branch from the enumeration tree. Ideally, one would like to use

\[
    \hat{f}(\sigma) = \max\{f(\phi) : \text{ext}(\phi) \subseteq \text{ext}(\sigma)\},
\]

which is the most strict function that still is a valid optimistic estimator. In general, however, computing this function is as hard as the original subgroup optimization problem we started with. Therefore, as a next best option, one can disregard selectability and consider the (selection-unaware) tight optimistic estimator given by

\[
    \hat{f}(\sigma) = \max\{f(R) : R \subseteq \text{ext}(\sigma)\}.
\]

This leaves us with a new combinatorial optimization problem: given a subpopulation \( Q \subseteq P \), find a sub-selection of \( Q \) that maximizes \( f \). In the following section we will discuss strategies for solving this optimization problem efficiently for different classes of objective functions—including dispersion-corrected objectives.

### 3. Efficiently Computable Tight Optimistic Estimators

Throughout this section we will identify a given subpopulation \( Q \subseteq P \) with the multiset of its target values \( \{y_1, \ldots, y_m\} \) and assume that the target values are indexed in ascending order, i.e., \( y_i \leq y_j \) for \( i \leq j \). Also, it is helpful to define the following partial order defined on finite multisets. Let \( Y = \{y_1, \ldots, y_m\} \) and \( Z = \{z_1, \ldots, z_{m'}\} \) be two multisets such that their elements are indexed in ascending order. We say that \( Y \) is element-wise less or equal to \( Z \) and write \( Y \leq_e Z \) if \( y_i \leq z_i \) for all \( i \in [\min\{m, m'\}] \).

#### 3.1. The standard case: monotone functions of a central tendency measure

The most general previous approach for computing the tight optimistic estimator for subgroup discovery with a metric target variable is described by Lemmerich et al (2016) where it is referred to as estimation by ordering. Here, we review this approach and give...
a uniform and generalized version of that paper’s results. For this, we define the general notion of a measure of central tendency as follows.

**Definition 1.** We call a mapping $c : \mathbb{N}^\mathbb{R} \rightarrow \mathbb{R}$ a (monotone) measure of central tendency if for all multisets $Y, Z \in \mathbb{N}^\mathbb{R}$ with $Y \leq c Z$ it holds that $c(Y) \leq c(Z)$.

One can check that this definition applies to the standard measures of central tendency, i.e., the arithmetic and geometric average as well as the median $\text{med}(Q) = y_{[m/2]}$, and also to weighted variants of them (note, however, that it does not apply to the mode).

With this we can define the class of objective functions for which the tight optimistic estimator can be computed efficiently by the standard approach as follows. We call

$$f : 2^P \rightarrow \mathbb{R}$$

a monotone level 1 objective function if it can be written as

$$f(Q) = g(|Q|, c(Q))$$

where $c$ is some measure of central tendency and $g$ is a function that is non-decreasing in both of its arguments. One can check that the impact measure $\text{ipa}$ falls under this category of functions as do many of its variants.

The central observation for computing the tight optimistic estimator for monotone level 1 functions is that the optimum value must be attained on a sub-multiset that contains a consecutive segment of elements of $Q$ from the top element w.r.t. $y$ down to some cut-off element. Formally, let us define the top sequence of sub-multisets of $Q$ as

$$T_i = \{y_{m-i+1}, \ldots, y_m\} \text{ for } i \in [m]$$

and note the following observation:

**Proposition 1.** Let $f$ be a monotone level 1 objective function. Then the tight optimistic estimator of $f$ can be computed as the maximum value on the top sequence, i.e.,

$$\hat{f}(Q) = \max\{f(T_i) : i \in [m]\}.$$  

Proof. Let $R \subseteq Q$ be of size $k$ with $R = \{y_{i_1}, \ldots, y_{i_k}\}$. Since $y_{i_j} \leq y_{m-j+1}$, we have for the top sequence element $T_k$ that $R \leq c T_k$ and, hence, $c(R) \leq c(T_k)$ implying

$$f(R) = g(k, c(R)) \leq g(k, c(T_k)) = f(T_k).$$

It follows that for each sub-multiset of $Q$ there is a top sequence element of at least equal objective value.\[\square\]

From this insight it is easy to derive an $O(m)$ algorithm for computing the tight optimistic estimator under the additional assumption that we can compute $g$ and the “incremental central tendency problem” $(i, Q, (c(T_1), \ldots, c(T_{i-1})) \mapsto c(T_i))$ in constant time. Note that computing the incremental problem in constant time implies to only access a constant number of target values and of the previously computed central tendency values. This can for instance be done for $c = \text{avg}$ via the incremental formula $\text{avg}(T_i) = ((i-1) \text{avg}(T_{i-1}) + y_{m-i+1})/i$ or for $c = \text{med}$ through direct index access of

\[\text{med}(Q) = y_{[m/2]}\]  

In this paper, we are using the simple definition of the median as the 0.5-quantile (as opposed to defining it as $(y_{m/2} + y_{i+1+m/2})/2$ for even $m$), which simplifies many of the definitions below and additionally is well-defined in settings where averaging of target values is undesired.
either of the two values $y_{m-\lfloor (i-1)/2 \rfloor}$ or $y_{m-\lceil (i-1)/2 \rceil}$. Since, according to Prop. \[1\], we have to evaluate $f$ only for the $m$ candidates $T_i$ to find $\hat{f}(Q)$ we can do so in time $O(m)$ by solving the problem incrementally for $i = 1, \ldots, m$. The same overall approach can be readily generalized for objective functions that are monotonically decreasing in the central tendency or those that can be written as the maximum of one monotonically increasing and one monotonically decreasing level 1 function. However, it breaks down for objective functions that depend on more than just size and central tendency— which inherently is the case when we want to incorporate dispersion-control.

3.2. Dispersion-corrected objective functions based on the median

We will now extend the previous recipe for computing the tight optimistic estimator to objective functions that depend not only on subpopulation size and central tendency but also on the target value dispersion in the subgroup. Specifically, we focus on the median as measure of central tendency and consider functions that are both monotonically increasing in the described subpopulation size and monotonically decreasing in some dispersion measure around the median. To precisely describe this class of functions, we first have to formalize the notion of dispersion measure around the median. For our purpose the following definition suffices. Let us denote by $Y^\text{med}$ the multiset of absolute differences to the median of a multiset $Y \in \mathbb{N}^R$, i.e., $Y^\text{med} = \{|y_1 - \text{med}(Y)|, \ldots, |y_m - \text{med}(Y)|\}$.

**Definition 2.** We call a mapping $d : \mathbb{N}^R \to \mathbb{R}$ a *dispersion measure around the median* if $d(Y)$ is monotone with respect to the multiset of absolute differences to its median $Y^\text{med}$, i.e., if $Y^\text{med} \leq Z^\text{med}$ then $d(Y) \leq d(Z)$.

One can check that this definition contains the measures median absolute deviation around the median $\text{mmd}(Y) = \text{med}(Y^\text{med})$, the root average of squared deviations around the median $\text{rsm}(Y) = \text{avg}(\{x^2 : x \in Y^\text{med}\})^{1/2}$, as well as the average absolute deviation around the median $\text{amd}(Y) = \text{avg}(Y^\text{med})$. Based on Def. \[2\] we can specify the class of objective functions that we aim to tackle as follows: we call a function $f : 2^P \to \mathbb{R}$ a dispersion-corrected or level 2 objective function (based on the median) if it can be written as

$$f(Q) = g(|Q|, \text{med}(Q), d(Q))$$  \hspace{1cm} (6)

where $d$ is some dispersion measure around the median and $g : \mathbb{R}^3 \to \mathbb{R}$ is a real function that is non-decreasing in its first argument and non-increasing in its third argument.

Our recipe for optimizing these functions is then to consider only subpopulations $R \subseteq Q$ that can be formed by selecting all individuals with a target value in some interval. Formally, for a fixed index $z \in \{1, \ldots, m\}$ define $m_z \leq m$ as the maximal cardinality of a sub-multiset of the target values that has median index $z$, i.e.,

$$m_z = \min\{2z, 2(m-z) + 1\}.$$  \hspace{1cm} (7)

Now, for $k \in [m_z]$, let us define $Q^k_z$ as the set with $k$ consecutive elements around index $z$. That is

$$Q^k_z = \left\{y_z - \lfloor \frac{k-1}{2} \rfloor, \ldots, y_z, \ldots, y_z + \lceil \frac{k-1}{2} \rceil \right\}.$$  \hspace{1cm} (8)
With this we can define the elements of the median sequence $Q_z$ as those subsets of the form of Eq. (8) that maximize $f$ for some fixed index $z \in [m]$. That is, $Q_z = Q_z^{k^*_z}$ where $k^*_z \in [m_z]$ is minimal with

$$f(Q_z^{k^*_z}) = g(k^*_z, y_z, d(Q_z^{k^*_z})) = \max\{f(Q_z^k): k \in [m_z]\}.$$  

Thus, the number $k^*_z$ is the smallest cardinality that maximizes the trade-off of size and dispersion encoded by $g$ (given the fixed median $y_z = \text{med}(Q_z^{k^*_z})$ for all $k$). In the following proposition we note that, as desired, searching the median sequence is sufficient for finding optimal subsets of $Q$.

**Proposition 2.** Let $f$ be a dispersion-corrected objective function based on the median. Then the tight optimistic estimator of $f$ can be computed as the maximum value on the median sequence, i.e., $\hat{f}(Q) = \max\{f(Q_z): z \in [m]\}$.

**Proof.** For a sub-multiset $R \subseteq Q$ let us define the gap count $\gamma(R)$ as

$$\gamma(R) = |\{y \in Q \setminus R: \min R < y < \max R\}|.$$  

Let $O \subseteq Q$ be an $f$-maximizer with minimal gap count, i.e., $f(R) < f(O)$ for all $R$ with $\gamma(R) < \gamma(O)$. Assume that $\gamma(O) > 0$. That means there is a $y \in Q \setminus O$ such that $\min O < y < \max O$. Define

$$S = \begin{cases} (O \setminus \{\min O\}) \cup \{y\}, & \text{if } y \leq \text{med}(O) \\ (O \setminus \{\max O\}) \cup \{y\}, & \text{otherwise} \end{cases}.$$  

Per definition we have $|S| = |O|$ and $\text{med}(S) = \text{med}(O)$. Additionally, we can check that $S^\text{med}_\Delta \leq O^\text{med}_\Delta$, and, hence, $d(S) \leq d(Q)$. This implies that

$$f(S) = g(|S|, \text{med}(S), d(S)) \geq g(|O|, \text{med}(O), d(O)) = f(O).$$  

However, per definition of $S$ it also holds that $\gamma(S) < \gamma(O)$, which contradicts that $O$ is an $f$-optimizer with minimal gap count. Hence, any $f$-maximizer $O$ must have a gap count of zero. In other words, $O$ is of the form $O = Q_z^{k_t}$ as in Eq. (8) for some median $z \in [m]$ and some cardinality $k \in [m_z]$ and per definition we have $f(Q_z) \geq f(O)$ as required. 

Consequently, we can compute the tight optimistic estimator for any dispersion-corrected objective function based on the median in time $O(m^2)$ for subpopulations of size $m$—again, given a suitable incremental formula for $d$. While this is not generally a practical algorithm in itself, it is a useful departure point for designing one. In the next section we show how it can be brought down to linear time when we introduce some additional constraints on the objective function.
Figure 2. Dispersion-corrected coverage of the sets $Q^k_z$ as defined in Eq. (8) for median indices $z \in \{9, 10, 11, 12\}$ with 21 random target values; median sequence elements $Q_z$ can be found in incremental constant time since optimal size $k^*_z$ is within a constant range of $k^*_{z+1}$ (Thm. 3).

3.3. Reaching linear time—objectives based on dispersion-corrected coverage

Equipped with the general concept of the median sequence, we can now address the special case of dispersion-corrected objective functions where the trade-off between the subpopulation size and target value dispersion is captured by a linear function of size and the sum of absolute differences from the median. For concreteness we consider objective functions of the form

$$f(Q) = g(dcc(Q), \text{med}(Q))$$

where $g$ is non-decreasing in its first argument and $dcc$ denotes the dispersion-corrected coverage defined by

$$dcc(Q) = \frac{|Q|}{|P|} \left( 1 - \frac{\text{amd}(Q)}{\text{amd}(P)} \right)_+.$$

Let us note, however, that we could replace the $dcc$ function by any linear function that depends positively on size and negatively on the sum of absolute deviations from the median $\text{smd}(Q) = \sum_{y \in Q} |y - \text{med}(Q)|$.

The key to computing the tight optimistic estimator $\hat{f}$ in linear time is that the members of the median sequence $Q_z$ can be computed incrementally in constant time. Indeed, we can prove the following theorem, which states that the optimal size for a multiset around median index $z$ is within 3 of the optimal size for a multiset around median index $z + 1$. 

11
Theorem 3. Let \( f \) be of the form of Eq. (9). For \( z \in [m-1] \) it holds for the size \( k^*_z \) of the \( f \)-optimal multiset with median \( z \) that
\[
\begin{align*}
  k^*_z &\in \{ \text{max}(0,k^*_z-3),\ldots,\text{min}(m_z,k^*_z+3) \} \tag{10}.
\end{align*}
\]

One idea to prove this theorem is to show that a) the gain in \( f \) for increasing the multiset around a median index \( z \) is alternating between two discrete concave functions and b) that the gains for growing multisets between two consecutive median indices are bounding each other. The details of this proof, which is rather long and partially technical, can be found in Appendix A.

It follows that, after computing the objective value of \( Q_m \) trivially as \( f(Q_m) = g(1/|P|,y_m) \), we can obtain \( f(Q_{z-1}) \) for \( z = m,\ldots,2 \) by checking the at most 7 candidate set sizes given by Eq. (10) as
\[
\hat{f}(Q) = \max \{ f(Q_z) : z \in [m] \}
\]

Algorithm 2: Linear time algorithm for computing tight optimistic estimator \( \hat{f}(Q) \) of objective \( f(Q) = g(\text{acc}(Q),\text{med}(Q)) \) as in Eq. (9).

**Theorem 3.** Let \( f \) be of the form of Eq. (9). For \( z \in [m-1] \) it holds for the size \( k^*_z \) of the \( f \)-optimal multiset with median \( z \) that
\[
\begin{align*}
  k^*_z &\in \{ \text{max}(0,k^*_z-3),\ldots,\text{min}(m_z,k^*_z+3) \} \tag{10}.
\end{align*}
\]

One idea to prove this theorem is to show that a) the gain in \( f \) for increasing the multiset around a median index \( z \) is alternating between two discrete concave functions and b) that the gains for growing multisets between two consecutive median indices are bounding each other. The details of this proof, which is rather long and partially technical, can be found in Appendix A.

It follows that, after computing the objective value of \( Q_m \) trivially as \( f(Q_m) = g(1/|P|,y_m) \), we can obtain \( f(Q_{z-1}) \) for \( z = m,\ldots,2 \) by checking the at most 7 candidate set sizes given by Eq. (10) as
\[
\hat{f}(Q) = \max \{ f(Q_z) : z \in [m] \}
\]

Algorithm 2: Linear time algorithm for computing tight optimistic estimator \( \hat{f}(Q) \) of objective \( f(Q) = g(\text{acc}(Q),\text{med}(Q)) \) as in Eq. (9).

**Theorem 3.** Let \( f \) be of the form of Eq. (9). For \( z \in [m-1] \) it holds for the size \( k^*_z \) of the \( f \)-optimal multiset with median \( z \) that
\[
\begin{align*}
  k^*_z &\in \{ \text{max}(0,k^*_z-3),\ldots,\text{min}(m_z,k^*_z+3) \} \tag{10}.
\end{align*}
\]

One idea to prove this theorem is to show that a) the gain in \( f \) for increasing the multiset around a median index \( z \) is alternating between two discrete concave functions and b) that the gains for growing multisets between two consecutive median indices are bounding each other. The details of this proof, which is rather long and partially technical, can be found in Appendix A.

It follows that, after computing the objective value of \( Q_m \) trivially as \( f(Q_m) = g(1/|P|,y_m) \), we can obtain \( f(Q_{z-1}) \) for \( z = m,\ldots,2 \) by checking the at most 7 candidate set sizes given by Eq. (10) as
\[
\hat{f}(Q) = \max \{ f(Q_z) : z \in [m] \}
\]
and \( e_l(1) = e_r(m) = 0 \). Subsequently, we can compute sums of deviations from center points of arbitrary subpopulations in constant time, as the following statement shows (see Appendix B).

**Proposition 4.** Let \( Q = \{y_1, \ldots, y_a, \ldots, y_z, \ldots, y_b, \ldots, y_m\} \) be a multiset with \( 1 \leq a < z < b \leq m \) and \( y_i \leq y_j \) for \( i \leq j \). Then the sum of absolute deviations to \( y_i \) of all elements of the submultiset \( \{y_a, \ldots, y_z, \ldots, y_b\} \) can be expressed as

\[
\sum_{i=a}^{b} |y_z - y_i| = e_l(z) - e_l(a) - (a - 1)(y_z - y_a) \\
+ e_r(z) - e_r(b) - (m - b)(y_b - y_z) .
\]

With this we can compute \( k \mapsto f(Q^k) \) in constant time (assuming \( g \) can be computed in constant time). Together with Prop. 2 and Thm. 3 this results in a linear time algorithm for computing \( Q \mapsto \hat{f}(Q) \) (see Alg. 2 for a pseudo-code that summarizes all ideas).
| Dataset | Selection Bias | Efficiency |
|---------|----------------|------------|
| Name | \(|P|\) | \(|II|\) | \(\text{med}(P)\) | \(\text{amd}(P)\) | \(\text{cov}(Q_0)\) | \(\text{cov}(Q_1)\) | \(\text{med}(Q_0)\) | \(\text{med}(Q_1)\) | \(\text{amd}(Q_0)\) | \(\text{amd}(Q_1)\) | \(\alpha_{\text{eff}}\) | \(\|\mathcal{E}_0\|\) | \(\|\mathcal{E}_1\|\) | \(t_0\) | \(t_1\) |
| abalone | 4,177 | 69 | 9 | 2.359 | 0.544 | 0.191 | 11 | 11 | 2.257 | 1.662 | 1 | 848,258,609,177 | 304 | 339 |
| ailerons | 13,750 | 357 | -0.0008 | 0.000383 | 0.906 | 0.59 | -0.0007 | -0.0006 | 0.000288 | 0.000198 | 0.3 | 1,069,456 | 54,103 | 6,542 | 460 |
| auto MPG8 | 392 | 24 | 22.5 | 6.524 | 0.497 | 0.497 | 29 | 29 | 4.791 | 4.791 | 1 | 97 | 67 | 0.11 | 0.09 |
| baseball | 337 | 24 | 74.0 | 954.386 | 0.362 | 0.003 | 1550 | 2500 | 1245.092 | 0 | 1 | 117 | 117 | 0.22 | 0.21 |
| california | 20,640 | 72 | 179.700 | 88.354 | 0.385 | 0.019 | 262.500 | 500.001 | 9261 | 294.00 | 0.4 | 1,368,662 | 65,707 | 2,676 | 368 |
| compact | 8,192 | 202 | 89 | 9.661 | 0.464 | 0.603 | 94 | 93 | 7.8 | 3.472 | 0.5 | 2,458,105 | 59,053 | 5,161 | 208 |
| concrete | 1,030 | 70 | 34.4 | 13.427 | 0.284 | 0.1291 | 48.97 | 50.7 | 12.744 | 9.512 | 1 | 512,195,221 | 322 | 43.9 | 35.8 |
| dee | 365 | 60 | 2.787 | 0.831 | 0.523 | 0.381 | 3.815 | 4.008 | 0.721 | 0.434 | 1 | 18,663,263 | 205 | 1.29 |
| delta_aill | 7,129 | 66 | -0.0001 | 0.000231 | 0.902 | 0.392 | 0.0001 | 0.000026 | 0.000119 | 0.3 | 104,985,1,415 | 8.9 | 4.01 |
| delta elv | 9,517 | 66 | 0.001 | 0.00198 | 0.384 | 0.369 | 0.002 | 0.0012 | 0.00108 | 0.05 | 6,356,456 | 526,114 | 13,712 | 2,891 |
| elevators | 16,599 | 155 | 0.02 | 0.00411 | 0.113 | 0.283 | 0.03 | 0.021 | 0.00813 | 0.00373 | 0.05 | 6,356,456 | 526,114 | 13,712 | 2,891 |
| forest fires | 517 | 70 | 0.52 | 12.832 | 0.01 | 0.002 | 86.45 | 278.53 | 56.027 | 0 | 1 | 340,426,264 | 207 | 23 | 23.7 |
| friedman | 1,200 | 48 | 14.651 | 4.234 | 0.387 | 0.294 | 18.934 | 19.277 | 3.065 | 2.73 | 1 | 19,209,248 | 323 | 1.56 |
| house | 22,784 | 160 | 33.200 | 28.456 | 0.56 | 0.723 | 45.200 | 34.000 | 40.576 | 27.214 | 0.002 | 1,221,696,114,566 | 7.937 | 1,308 |
| laser | 993 | 42 | 46 | 35.561 | 0.32 | 0.093 | 109 | 1.5 | 30.134 | 15.662 | 2 | 5,908,85 | 0.96 | 0.83 |
| mortgage | 1,409 | 128 | 6.71 | 2.373 | 0.256 | 0.097 | 11.61 | 14.41 | 2.081 | 0.98 | 1 | 40,753,1,270 | 11.6 | 1.59 |
| nj | 40,768 | 79 | -0.02086 | 8.599 | 0.497 | 0.349 | 0.076 | 0.193 | 8.54 | 2.032 | 1 | 6,513,1,017 | 31.9 | 13.2 |
| pole | 1,998 | 260 | 0 | 28.949 | 0.40 | 0.24 | 100 | 100 | 35.905 | 16.692 | 0.2 | 1,014,146 | 2,966 | 2,638 | 15 |
| puma32h | 8,192 | 318 | 0.00028 | 0.023 | 0.299 | 0.244 | 0.026 | 0.031 | 0.018 | 0.017 | 0.4 | 3,141,046 | 5,782 | 2,648 | 15.5 |
| stock | 950 | 80 | 46.625 | 5.47 | 0.471 | 0.337 | 52.5 | 54.375 | 3.741 | 2.515 | 1 | 85,602,1,822 | 12.5 | 1.56 |
| treasury | 1,049 | 128 | 6.61 | 2.473 | 0.182 | 0.339 | 13.16 | 8.65 | 2.591 | 0.863 | 1 | 49,197,9,247 | 14.8 | 5.91 |
| wankara | 1,609 | 87 | 47.7 | 12.753 | 0.545 | 0.296 | 60.6 | 67.6 | 8.873 | 4.752 | 1 | 191,053,4,081 | 11.9 | 1.24 |
| wizard | 1,461 | 82 | 60 | 12.622 | 0.6 | 0.349 | 72.9 | 78.5 | 8.527 | 3.899 | 1 | 177,768,1,409 | 38.5 | 1.48 |

Table 1. Datasets with corresponding population size \(|P|\), number of base propositions \(|II|\), global median \((\text{med}(P))\) and average median deviation \((\text{amd}(P))\) followed by coverage \((\text{cov}(Q_0), \text{cov}(Q_1))\), median \((\text{med}(Q_0), \text{med}(Q_1))\), and average absolute median deviation \((\text{amd}(Q_0), \text{amd}(Q_1))\) for best subgroup w.r.t. non-dispersion corrected function \(f_0\) and dispersion-corrected function \(f_1\), respectively; boldface indicates higher coverage and median and lower dispersion, underlines indicate higher dispersion than in global population; final column segment contains accuracy parameter used in the efficiency study \((\alpha_{\text{eff}})\) as well as number of expanded nodes \((\|\mathcal{E}_0\|, \|\mathcal{E}_1\|)\) and computation time in seconds \((t_0, t_1)\) for optimistic estimator based on top sequence \(f_0\) and tight optimistic estimator \(f_1\), respectively—in both cases when optimizing \(f_1\); depth-limit of 10 is used for all datasets with \(a < 1\), no depth-limit otherwise.
4. Dispersion-corrected Subgroup Discovery in Practice

The essential qualitative effects of dispersion-correction and of efficient tight optimistic value estimation are obvious: the first reduces the target value dispersion of result subgroups compared to uncorrected objectives; the second reduces the part of the description language that has to be enumerated in order to find (approximately) optimal patterns compared to weaker estimators. What remains to be seen is the practical quantitative effect of both. For that purpose, this section presents empirical results gathered from applying the introduced techniques to a range of publicly available real-world datasets.  

To assess the quantitative effect of dispersion correction, let us consider the objective function $f_1(Q) = \text{dcc}(Q)\text{mds}_+(Q)$ based on dispersion-corrected coverage and the positive relative median shift

\[ \text{mds}_+(Q) = \left( \frac{\text{med}(Q) - \text{med}(P)}{\max(P) - \text{med}(P)} \right)_{+}. \]

This function obeys Eq. (9); thus, its tight optimistic estimator can be computed using the linear time algorithm from Sec. 3.3. To contrast its selection bias let us also consider the non-dispersion corrected variant $f_0(Q) = \text{cov}(Q)\text{mds}_+(Q)$ where we simply replace the dispersion-corrected coverage by the ordinary coverage. This function is a monotone level 1 function, hence, its tight optimistic estimator $\hat{f}_0$ can be computed in linear time using the top sequence approach. The language of closed conjunctions $C_{cnj}$ is used for all datasets as description language.

Fig. 3 shows the characteristics of the optimal subgroups that are discovered with respect to both of these objective functions (see also Tab. 1 for exact numerical values). The first observation is that—as enforced by design—for all dataset the average absolute deviation from the median is lower for the dispersion-corrected variant (except in one case where both functions yield the same optimal subgroup). On average the dispersion for $f_1$ is 49 percent of the global dispersion, whereas it is 113 percent for $f_0$, i.e., when not optimizing the dispersion it is on average higher in the subgroups than in the global population. When it comes to the other subgroup characteristics, coverage and median target value, the global picture is that $f_1$ discovers somewhat more specific groups (average coverage 0.3 versus 0.44 for $f_0$) with higher median shift (on average 0.73 normalized median deviations higher). However, in contrast to dispersion, the behavior for median shift and coverage varies across the datasets. In Fig. 3 the datasets are ordered according to the difference in subgroup medians between the optimal subgroups w.r.t. $f_0$ and those w.r.t. $f_1$. This ordering reveals the following categorization of outcomes:

- On the left we have five datasets (11 to 6) for which $f_0$ yields subgroups with a higher median than those resulting from $f_1$ (on average 1.07 normalized median deviations higher). At the same time, all of these subgroups have a lower coverage

---

Datasets contain all regression datasets from the KEEL repository (Alcalá et al., 2010) with at least 5 attributes and two materials datasets from the Nomad Repository nomad-coe.eu/; see Tab. 1. Implementation available in open source Java library realKD bitbucket.org/realKD/; Computation times determined on MacBook Pro 3.1 GHz Intel Core i7.
than their dispersion-corrected counterparts (avg. 0.47) and, in four out of five cases, their average absolute deviation from the median is larger than in the global population, which renders them incoherent (avg. normalized median deviation in this group 1.32 for $f_0$ and 0.6 for $f_1$).

- In contrast, the five datasets on the other side of the spectrum (15 to 12) are characterized by very small groups with a very high median for $f_1$ and very high dispersion subpopulations for $f_0$ (average coverage 0.04 for $f_1$ versus 0.27 for $f_0$). In two cases, $f_1$ yields groups containing only a single member with maximum target value. However, in all datasets of this group the results for $f_0$ have an extremely large dispersion (in 4 cases larger than in the global population) of a factor of 1.75 of the global dispersion (in contrast 0.24 for $f_1$).

- Finally, there is the main group of 15 datasets (18 to 22) where the medians range from approximately equal to somewhat higher for $f_1$ (by on average 0.2 normalized median deviations). The subgroup selectors for both functions are fairly general, but those resulting from the dispersion-corrected objective are somewhat more specific (average coverages of 0.53 for $f_0$ and 0.33 for $f_1$). The average normalized dispersion for these datasets is 0.85 for $f_0$ and 0.54 for $f_1$.

In summary, when our description language is not able to reduce the error of subgroups with very high median value, $f_1$ settles for more coherent groups with a less extreme but still outstanding central tendency. On the other end of the scale, when no coherent groups with moderate size and median shift can be identified, the dispersion-corrected object resorts to the selection of very small groups with the most extreme target values. The majority of datasets obey the global trend of dispersion-correction leading to somewhat more specific subgroups with higher median that are, as intended, more coherent.

To study the effect of the tight optimistic estimator, let us compare its performance to that of a baseline estimator that can be computed with the standard top sequence approach. Since $f_1$ is upper bounded by $f_0$, $\hat{f}_0$ is a valid, albeit non-tight, optimistic estimator for $f_1$ and can thus be used for this purpose. The exact speed-up factor is determined by the ratio of enumerated nodes for both variants as well as the ratio of computation times for an individual optimistic estimator computation. While both factors determine the practically relevant outcome, the number of nodes evaluated is a much more stable quantity, which indicates the full underlying speed-up potential independent of implementation details. Similarly, “number of nodes evaluated” is also an insightful unit of time for measuring optimization progress. Therefore, in addition to the computation time in seconds $t_0$ and $t_1$, let us denote by $E_0, E_1 \subseteq L$ the set of nodes enumerated by branch-and-bound using $\hat{f}_0$ and $\hat{f}_1$, respectively—but in both cases for optimizing the dispersion-corrected objective $f_1$. Moreover, when running branch-and-bound with optimistic estimator $\hat{f}_i$, let us denote by $\sigma_i^*(n)$ and $\sigma_i^+(n)$ the best selector found and the top element of the priority queue (w.r.t. $\hat{f}_i$), respectively, after $n$ nodes have been enumerated.
Figure 4. Effect of tight optimistic estimator: (right) optimistic estimation ($\hat{f}_1(\sigma_1^+), \hat{f}_0(\sigma_0^+)$) of remaining search space and value of current best solution ($f_1(\sigma_1^+), f_1(\sigma_0^+)$) resulting from tight estimator and top sequence estimator, respectively—per processed nodes for dataset *binaries*; (left) speedup factor ($t_0/t_1$) in order for all datasets plus potential speed-up factor ($|E_0|/|E_1|$).

The plot on the left side of Fig. 4 shows the speed-up factor $t_1/t_0$ on a logarithmic axis for all datasets in increasing order along with the potential speed-up factors $|E_0|/|E_1|$ (see Tab. 1 for numerical values). There are seven datasets for which the speed-up turns out to be minor followed by four datasets with a modest speed-up factor of 2. For the remaining 14 datasets, however, we have solid speed-up factors between 4 and 20 and in four cases immense values between 100 and 4,000. This demonstrates the decisive potential effect of tight value estimation even when compared to another non-trivial estimator like $\hat{f}_0$ (which itself improves over simpler options by orders of magnitude; see Lemmerich et al. 2016). In almost all cases the potential speed-up given by the ratio of enumerated nodes is considerably higher than the actual speed-up, which shows that, despite the same asymptotic time complexity, an individual computation of the tight optimistic estimator is slower than the simpler top sequence based estimator—but also indicates that there is room for improvements in the implementation. When zooming in on the optimization progress over time for the *binaries* dataset, which exhibits the most extreme speed-up (right plot in Fig. 4), we can see that not only does the tight optimistic estimator close the gap between best current selector and current highest potential selector much faster—thus creating the huge speed-up factor—but also that it causes better solutions to be found earlier. This is an important property when we want to use the algorithm as an *anytime algorithm*, i.e., when allowing the user to terminate computation preemptively, which is important in interactive data analysis systems. This is an advantage enabled specifically by using the tight optimistic estimators in conjunction with the best-first node expansion strategy.
5. Concluding Remarks

This paper presented for the first time an effective algorithm for simultaneously optimizing size, central tendency, and dispersion in subgroup discovery with a numerical target. In particular, the ability to discover optimal subgroups w.r.t. the proposed optimization criteria allows to also draw conclusions from the absence of patterns from datasets.

Since both requirements originally abounded from the authors’ own work in Materials Science (Goldsmith et al., 2017; in press), they hope that these contributions also prove to be useful in related disciplines and therefore to generally raise the impact of subgroup discovery in scientific applications.

The specific measures of central tendency and dispersion used in the most efficient variant of the presented algorithm are the median and the average deviation from the median. This is a sound combination because the median of a set of values minimizes the sum of absolute deviations. On the other hand, it is natural to ask whether the proposed techniques can be extended to other measures—including mean and standard deviation, which come with a wealth of statistical tools (e.g., Chebyshev’s inequality and similar) to assess the statistical significance of findings. While this question is open for future research, let us point out that the given approach not only serves already as a good optimization proxy for those specific measures, it also provides other degrees of freedom, which might be interesting to exploit in their own right: The non-monotone dependence on the median allows to incorporate a much more sophisticated influence of the central tendency value than simple monotone mean shifts. For instance, given a suitable statistical model for the global distribution, the effect of the median could be a function of the probability \( \Pr[\text{med}(Q)] \). Furthermore, the feasible dispersion measures allow for interesting weighting schemes, which include possibilities of asymmetric effects of the error (e.g., for only punishing one-sided deviation from the median).

Another valuable direction for future research is the extension of consistency and error optimization to the case of multidimensional target variables where subgroup parameters can represent complex statistical models (known as exceptional model mining [Duivesteijn et al., 2016]). While this setting is algorithmically more challenging than the univariate case covered here, the underlying motivation remains: balancing group size and exceptionality, i.e., distance of local to global model parameters, with consistency, i.e., local model fit, should lead to the discovery of more meaningful statements about the underlying data generating process.

References

Alcalá J, Fernández A, Luengo J, Derrac J, García S, Sánchez L, Herrera F (2010) Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. Journal of Multiple-Valued Logic and Soft Computing 17(2-3):255–287

Atzmueller M (2015) Subgroup discovery. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 5(1):35–49
Bay SD, Pazzani MJ (2001) Detecting group differences: Mining contrast sets. Data Mining and Knowledge Discovery 5(3):213–246

Boley M, Grosskreutz H (2009) Non-redundant subgroup discovery using a closure system. In: Joint European Conf. on Machine Learning and Knowledge Discovery in Databases, Springer, pp 179–194

Boley M, Moens S, Gärtner T (2012) Linear space direct pattern sampling using coupling from the past. In: Proc. of the 18th ACM SIGKDD int. conf. on Knowledge discovery and data mining, ACM, pp 69–77

Duivesteijn W, Knobbe A (2011) Exploiting false discoveries–statistical validation of patterns and quality measures in subgroup discovery. In: IEEE 11th Int. Conf. on Data Mining, IEEE, pp 151–160

Duivesteijn W, Feelders AJ, Knobbe A (2016) Exceptional model mining. Data Mining and Knowledge Discovery 30(1):47–98

Friedman JH, Fisher NI (1999) Bump hunting in high-dimensional data. Statistics and Computing 9(2):123–143

Goldsmit BR, Boley M, Vreeken J, Scheffler M, Ghiringhelli LM (2017; in press) Uncovering structure-property relationships of materials by subgroup discovery. New Journal of Physics

Grosskreutz H, Rüping S, Wrobel S (2008) Tight optimistic estimates for fast subgroup discovery. In: Joint European Conf. on Machine Learning and Knowledge Discovery in Databases, Springer, pp 440–456

Grosskreutz H, Boley M, Krause-Traudes M (2010) Subgroup discovery for election analysis: a case study in descriptive data mining. In: Int. Conf. on Discovery Science, Springer, pp 57–71

Huan J, Wang W, Prins J (2003) Efficient mining of frequent subgraphs in the presence of isomorphism. In: 3rd IEEE Int. Conf. on Data Mining, IEEE, pp 549–552

Klösgen W (1996) Explora: A multipattern and multistrategy discovery assistant. In: Advances in knowledge discovery and data mining, American Association for Artificial Intelligence, pp 249–271

Klösgen W (2002) Data mining tasks and methods: Subgroup discovery: deviation analysis. In: Handbook of data mining and knowledge discovery, Oxford University Press, Inc., pp 354–361

Lavrač N, Kavšek B, Flach P, Todorovski L (2004) Subgroup discovery with cn2-sd. Journal of Machine Learning Research 5(Feb):153–188

Lemmerich F, Atzmueller M, Puppe F (2016) Fast exhaustive subgroup discovery with numerical target concepts. Data Mining and Knowledge Discovery 30(3):711–762
A. Proof of Theorem 3

In order to prove Thm. 3, let us start by noting that for functions of the form of Eq. (9), finding the set size \( k^*_z \) corresponds to maximizing the dispersion-corrected coverage among all multisets with consecutive elements around median \( y_z \) (as defined in Eq. 8). In order to analyze this problem, let us write

\[
h_z(k) = \text{dcc}(Q^k_z) = \frac{|Q^k_z|}{|P|} - \frac{\text{smd}(Q^k_z)}{\text{smd}(P)}
\]
for the dispersion-corrected coverage of the multiset $Q_z^k$. Let $\Delta h_z : [m_z] \to \mathbb{R}$ denote the difference or gain function of $h_z$, i.e., $\Delta h_z(k) = h_z(k) - h_z(k - 1)$ where we consider $Q_z^0 = \emptyset$ and, hence, $h_z(0) = 0$. With this definition we can show that $h_z$ is alternating between two concave functions, i.e., considering either only the even or only the odd subset of its domain, the gains are monotonically decreasing. More precisely:

**Lemma 5.** For all $k \in [m_z] \setminus \{1, 2\}$ we have that $\Delta h_z(k) \leq \Delta h_z(k - 2)$.

**Proof.** For $k \in [m_z]$, let us denote by $q_z^k$ the additional $y$-value that $Q_z^k$ contains compared to $Q_z^{k-1}$ (considering $Q_z^0 = \emptyset$), i.e., $Q_z^k \setminus Q_z^{k-1} = \{q_z^k\}$. We can check that

$$q_z^k = \begin{cases} q_z - \lfloor \frac{k-1}{2} \rfloor, & k \text{ odd} \\ q_z + \lfloor \frac{k-1}{2} \rfloor, & k \text{ even} \end{cases}.$$

With this and using the shorthands $n = |P|$ and $d = \text{smd}(P)$ we can write

$$\Delta g_z(k) - \Delta g_z(k - 2) = g_z(k) - g_z(k - 1) - (g_z(k - 2) - g_z(k - 3))$$

$$= \frac{k}{n} \cdot \frac{\text{smd}(Q_z^k)}{d} - \frac{k - 1}{n} + \frac{\text{smd}(Q_z^{k-1})}{d} - \frac{k - 2}{n} + \frac{\text{smd}(Q_z^{k-2})}{d} + \frac{k - 3}{n} - \frac{\text{smd}(Q_z^{k-3})}{d}$$

$$= \frac{1}{n} \sum_{\Delta = 0} \left( |q_z^k - y_z| - |q_z^{k-1} - y_z| + |q_z^{k-2} - |q_z^{k-2} - y_z| \right)$$

$$= \frac{1}{d} \left( |q_z^k - y_z| - |q_z^{k-1} - y_z| + |q_z^{k-2} - y_z| \right)$$

**case** $k$ odd

$$= \frac{1}{d} \left( y_z - y_{z - \lfloor \frac{k-1}{2} \rfloor} + (y_z - y_{z - \lceil \frac{k-2}{2} \rceil}) \right) = y_z - \lfloor \frac{k-1}{2} \rfloor - y_{z - \lceil \frac{k-3}{2} \rceil} \leq 0$$

**case** $k$ even

$$= \frac{1}{d} \left( y_{z + \lceil \frac{k-1}{2} \rceil} - y_z + (y_{z + \lfloor \frac{k-2}{2} \rfloor} - y_z) \right) = y_{z + \lfloor \frac{k-3}{2} \rfloor} - y_{z + \lfloor \frac{k-1}{2} \rfloor} \leq 0$$

\[ \square \]

One important consequence of this fact is that the operation of growing a set around median $z$ by two elements—one to the left and one to the right—has monotonically decreasing gains. In other words, the smoothed function $h_z(k) = h_z(k) + h_z(k - 1)$ is concave or formally

$$\Delta h_z(k) + \Delta h_z(k - 1) \geq \Delta h_z(k + 1) + \Delta h_z(k) \quad . \quad (13)$$

Moreover, we can relate the gain functions of consecutive median indices as follows.
Lemma 6. Let $z \in [m] \setminus \{1\}$ and $k \in [m_{z-1}] \setminus \{1, 2, 3\}$. It holds that

$$\Delta h_{z-1}(k - 2) + \Delta h_{z-1}(k - 3) \geq \Delta h_z(k) + \Delta h_z(k - 1)$$  \hspace{1cm} (14)

$$\Delta h_{z-1}(k) + \Delta h_{z-1}(k - 1) \leq \Delta h_z(k) + \Delta h_z(k - 2)$$  \hspace{1cm} (15)

Proof. For this proof, let us use the same shorthands as in the proof of Lemma 5 and start by noting that for all $i \in [m]$ and $k \in [m_z] \setminus \{1\}$ we have the equality

$$\Delta h_i(k) + \Delta h_i(k - 1) = \frac{2}{n} - \frac{|q_{i}^k - y_i| + |q_{i}^{k-1} - y_i|}{d}$$  \hspace{1cm} (16)

which we can see by extending

$$\Delta h_i(k) + \Delta h_i(k - 1) = h_i(k) - h_i(k - 1) + h_i(k - 1) - h_i(k - 2)$$

$$= \frac{k - k + 2}{n} - \frac{smd(Q_i^k) - smd(Q_i^{k-2})}{d} = \frac{2}{n} - \frac{|q_{i}^k - y_i| + |q_{i}^{k-1} - y_i|}{d}.$$  

We can then show Eq. (14) by applying Eq. (16) two times to

$$\Delta h_{z-1}(k - 2) + \Delta h_{z-1}(k - 3) - (\Delta h_z(k) + \Delta h_z(k - 1))$$

$$= \frac{1}{d} \left( -|q_{z-1}^{k-2} - y_{z-1}| - |q_{z-1}^{k-3} - y_{z-1}| + |q_{z}^{k} - y_{z}| + |q_{z}^{k-1} - y_{z}| \right)$$

and finally by checking separately the case $k$ odd

$$= \frac{1}{d} \left( y_{z-1 - \left\lfloor \frac{k-3}{2} \right\rfloor} - y_{z-1} + y_{z-1} - y_{z-1 + \left\lceil \frac{k-4}{2} \right\rceil} + y_{z} - y_{z - \left\lfloor \frac{k-2}{2} \right\rfloor} + y_{z + \left\lfloor \frac{k-2}{2} \right\rfloor} - y_{z} \right)$$

$$= \frac{1}{d} \left( y_{z - \left\lfloor \frac{k-1}{2} \right\rfloor} - y_{z - \left\lfloor \frac{k-1}{2} \right\rfloor} + y_{z-1 + \left\lceil \frac{k}{2} \right\rceil} - y_{z-3 + \left\lceil \frac{k}{2} \right\rceil} \geq 0 \right)$$

and the case $k$ even

$$= \frac{1}{d} \left( y_{z-1} - y_{z-1 + \left\lceil \frac{k-3}{2} \right\rceil} + y_{z-1 - \left\lfloor \frac{k-4}{2} \right\rfloor} - y_{z-1} + y_{z + \left\lfloor \frac{k-1}{2} \right\rfloor} - y_{z} + y_{z - \left\lfloor \frac{k-2}{2} \right\rfloor} \right)$$

$$= \frac{1}{d} \left( y_{z+1 - \left\lceil \frac{k}{2} \right\rceil} - y_{z+1 - \left\lfloor \frac{k}{2} \right\rfloor} + y_{z + \left\lfloor \frac{k-1}{2} \right\rfloor} - y_{z-2 + \left\lceil \frac{k-1}{2} \right\rceil} \geq 0 \right).$$
Similarly, for Eq. (15) by applying Eq. (16) two times we can write
\[ \Delta h_{z-1}(k) + \Delta h_{z-1}(k-1) - (\Delta h_z(k-2) + \Delta h_z(k-3)) \]
\[= \frac{1}{d} \left( -|q_z^{k-1} - y_z - 1| - |q_z^{k-2} - y_z - 1| + |q_z^{k-3} - y_z - 1| \right) \]
\[= \begin{cases} \frac{1}{d} \left( y_z -\frac{k-2}{2} - y_z +\frac{k-1}{2} + y_z -\frac{k-2}{2} + y_z -\frac{k-1}{2} \right) & \leq 0, \quad k \text{ odd} \\ \frac{1}{d} \left( y_z -\frac{k-2}{2} - y_z +\frac{k-1}{2} + y_z -\frac{k-2}{2} + y_z -\frac{k-1}{2} \right) & \leq 0, \quad k \text{ even} \end{cases} \]

Combining all of the above we can finally proof our main result as follows.

**Theorem 3.** We start by showing that every \( k \in [m_z+1] \) with \( k < k^*_z - 3 \) can not be an optimizer of \( h_{z+1} \). It follows that \( k^*_z - 3 \leq k^*_z + 1 \), and, hence, \( k^*_z \leq k^*_z + 1 + 3 \) as required for the upper bound. Indeed, we have
\[ h_{z+1}(k) = h_{z+1}(k + 2) \]
\[\leq h_{z+1}(k + 2) - (\Delta h_{z+1}(k^*_z - 2) + \Delta h_{z+1}(k^*_z - 3)) \]
\[\leq h_{z+1}(k + 2) - (\Delta h_{z}(k^*_z + 1) + \Delta h_{z}(k^*_z + 1)) < h_{z+1}(k + 2) \quad \text{(by Lm. 6)} \]

Analogously, for the lower bound, we show that every \( k \in [m_z+1] \) with \( k > k^*_z + 3 \) can not be the smallest optimizer of \( h_{z+1} \). It follows that \( k^*_z + 3 \geq k^*_z + 1 \), and, hence, \( k^*_z \geq k^*_z + 1 - 3 \) as required. Indeed, we can write
\[ h_{z+1}(k) = h_{z+1}(k - 2) \]
\[\leq h_{z+1}(k - 2) + \Delta h_{z+1}(k^*_z + 2) + \Delta h_{z+1}(k^*_z + 1) \]
\[\leq h_{z+1}(k - 2) + \Delta h_{z}(k^*_z + 2) + \Delta h_{z}(k^*_z + 1) \leq h_{z+1}(k - 2) \quad \text{(by Lm. 6)} \]

\[\square\]
B. Additional Proofs

Prop. 4. Using $d_{ij}$ as a shorthand for $y_j - y_i$ for $i, j \in [m]$ with $i \leq j$ we can write

$$e_l(z) - e_l(a) - (a - 1)(y_z - y_a) + e_r(z) - e_r(b) - (m - b)d_{zb}$$

$$= \sum_{i=1}^{z-1} d_{iz} - \sum_{i=1}^{a-1} d_{ia} - (a - 1)d_{az} + \sum_{i=z+1}^{m} d_{zi} - \sum_{i=b+1}^{m} d_{bi} - (m - b)d_{zb}$$

$$= \sum_{i=a}^{z-1} d_{iz} + \sum_{i=1}^{a-1} \frac{(d_{iz} - d_{ia})}{d_{az}} - (a - 1)d_{az} + \sum_{i=z+1}^{b} d_{zi} + \sum_{i=b+1}^{m} \frac{(d_{zi} - d_{bi})}{d_{zb}} - (m - b)d_{zb}$$

$$= \sum_{i=a}^{z-1} d_{iz} + (a - 1)d_{az} - (a - 1)d_{az} + \sum_{i=z+1}^{b} d_{zi} + (m - b)d_{zb} - (m - b)d_{zb}$$

$$= \sum_{i=a}^{z-1} d_{iz} + \sum_{i=z+1}^{b} d_{zi} = \sum_{i=a}^{b} |y_z - y_i|$$