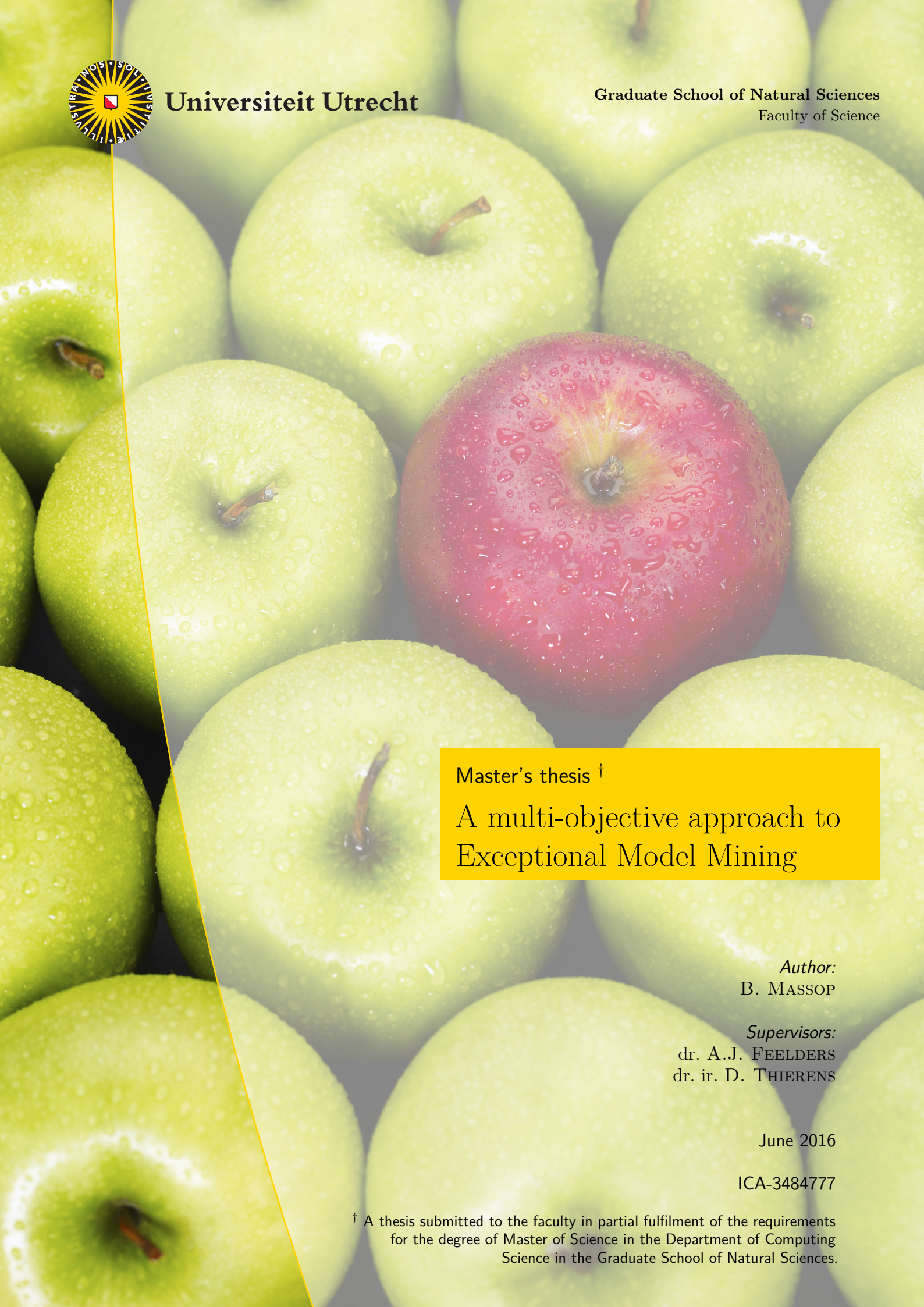

Universiteit Utrecht

Graduate School of Natural Sciences
Faculty of Science

Master's thesis †

A multi-objective approach to Exceptional Model Mining

Author:
B. Massop

Supervisors:
dr. A.J. Feelders
dr. ir. D. Thierens

June 2016

ICA-3484777

† A thesis submitted to the faculty in partial fulfilment of the requirements for the degree of Master of Science in the Department of Computing Science in the Graduate School of Natural Sciences.

Abstract

This thesis investigates the applicability and use of a multi-objective approach to the Exceptional Model Mining knowledge discovery problem under linear regression models. The main contribution of this thesis is twofold.

First, in a single-objective setting, Cook's distance is often used as a quality measure. Empirical results show that this distance measure is biased and unsound for measuring exceptionality. It is shown that this bias arises from repeated sampling from a finite population of observations. A statistical distribution for Cook's distance of random subgroups with fixed support is derived based on asymptotic properties of linear regression, and a modified Cook's distance is proposed for which the derived distribution does not rely on these properties. The latter distribution is shown to be related to Fisher's F-distribution.

Second, the performance of multi-objective strategies based on Pareto Local Search is compared to existing single-objective techniques. Each algorithm's value to the analyst is measured by means of attained hypervolume on evidence, generality and confidence measures. The multi-objective strategies are shown to provide significantly better results on the majority of benchmark datasets. Apart from attaining a greater hypervolume, these strategies are shown to yield solutions with greater evidence as well.

Acknowledgements

I am grateful to my supervisors, dr. Ad Feelders and dr. ir. Dirk Thierens for helping me with their thoughtful comments, questions and an insight into the relevant literature. In particular I am thankful to Ad for his insistence on having the 'why' explained, which led to a particularly interesting journey in statistics and the writing of a major part of this thesis. I am also thankful to Dirk for his friendly yet sceptical remarks, which allowed me to keep the scope of this thesis within reason.

I would also like to acknowledge and thank Thomas Krak for providing his implementation of Tree-Constrained Gradient Ascent for reuse and further inspection, and for taking the time to guide me through it. This saved me plenty of time in recreating his experiments.

For the non-scientific side of this thesis, my thanks go to Ryan Pendavingh for his endless love and encouragement. Furthermore, I would would like to acknowledge Guido Passage, Geertièn de Vries and all other occupants of the BBL-502 study room, who had to bear with my ramblings on statistical distributions and kept my spirits high by providing fruitful discussion on unrelated topics — not to mention that without you all, I would have had to eat all that liquorice by myself. Finally, I would like to thank my parents, friends and all others who have supported me throughout my thesis — your help has been invaluable.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Exceptional Model Mining is a fairly recent contribution to the field of data mining [40]. In Exceptional Model Mining, the goal is to find a concise subgroup of a dataset that is exceptionally different in some way. Consider a dataset, containing observations on various attributes or features in the domain under consideration, and consider a model describing the nature of the relation of a selection of these attributes. A simple linear regression model, for example, would describe a relation between the model's dependent variable and its dependent variable, parametrized by an intercept and a slope.

Exceptional Model Mining aims to identify interesting parts of the dataset, for which this model is exceptional. Here, a part is considered exceptional when the relation described by the estimated model for the full dataset does not quite explain the relation observed for this part of the dataset. To be of value to the analyst, the observations in such a part must be described in a general way that is non-specific to the exact choice of observations. As such, they must be identified by a concise description on attributes in the dataset: for example, the description 'all observations for men under the age of 55' is concise, but the description 'observations 1, 3 and 232 through 310' is not.

More formally, the Exceptional Model Mining problem is concerned with the task of finding a concise description of a *concept* within a dataset, for which the model of the subgroup — all the observations in the dataset described by this concept — is exceptionally different from the model of the entire dataset. Exceptional Model Mining is similar to the Subgroup Discovery data mining problem, however with an important distinction: Subgroup Discovery only considers the distribution of a single target variable, where Exceptional Model Mining is formulated independent of the choice of model. As such, Exceptional Model Mining can be considered a generalization of Subgroup Discovery. Therefore Exceptional Model Mining belongs to the class of supervised descriptive rule learning problems, designed to capture knowledge contained in datasets.

Similar to Subgroup Discovery, Exceptional Model Mining is designed to provide the analyst with insight into peculiarities in the dataset. This capture of knowledge, assisting the analyst in gaining insight into the dataset at hand, will form a central perspective throughout this work. To properly introduce Exceptional Model Mining in its broader context, this chapter will first proceed with an introduction from exactly this perspective.

1.1 Mining for knowledge in databases

Exceptional Model Mining can be considered a strategy for the discovery of knowledge in databases. Klösgen introduces this general *Knowledge Discovery in Databases* data mining task, further denoted as knowledge discovery, as the "search for patterns that exist in databases, but are hidden among the volumes of data" [36]. Klösgen identifies the following general requirements for patterns to supply valuable knowledge:

1. Patterns must consist of interpretable problem-relevant attributes, and
2. Patterns must describe many representative cases of the domain.

Klösgen divides knowledge discovery strategies into the paradigms of user-guided search, visualization, navigation, and low-level strategies for searching and evaluating patterns. The former strategies involve strong user guidance for discovery of patterns of interest. As such, these paradigms are impractical from a computational point of view, and will not be considered in this work.

1.1.1 Automated knowledge discovery

Exceptional Model Mining and other members of the class of supervised descriptive rule learning problems — this class includes Subgroup Discovery and others [46] — rely on strategies of Klösgen's latter paradigm. These automated knowledge discovery strategies, often described as data mining techniques, can colloquially be envisioned as a search through a hypothesis space. Here, hypotheses in the problem domain are specified by means of a pattern language. Such patterns then describe hypotheses in a way comprehensible to the analyst.

Generally, like any hypothesis, patterns take the form of a *concept* or antecedent in what is often a propositional language over problem-relevant independent attributes of the data, and a consequent that models peculiarities in the dependent attributes in the data for this concept. Propositional statements on problem-relevant attributes are often easily interpretable for the analyst. When an appropriate consequent for the pattern is

chosen, propositional antecedents are thus likely to meet Klösgen's first requirement for patterns.[1]

1.1.2 Patterns in Exceptional Model Mining

If we now limit the pattern language to a fixed and appropriately chosen consequent, the search can be described as a search through the concept space of a problem domain. Klösgen's second requirement can then be met by limiting the search to propositions for which we have enough observations in the dataset. In Exceptional Model Mining, we find our patterns to have such a fixed consequent: the distribution of the dependent variables, under some model, is unusual for a concept. Here we define unusual as in some sense significantly different from the population, that is, the entire dataset. For our case of Exceptional Model Mining, the pattern space is thus limited to the concept space, and Exceptional Model Mining can then be defined as the search for concepts with an unusual distribution of the dependent variables. As follows from Klösgen's second requirement, these concepts must adhere to some constraint of minimum support in the dataset.

1.1.3 The analyst's goal

In the sequel of this work, we will assume that the analyst is satisfied if some *nuggets* are identified by the knowledge mining strategy [35]. For Exceptional Model Mining, these nuggets take the form of a subgroup of the observations in the dataset, identified by a concept on problem-relevant attributes of this dataset, featuring an unusual behaviour or an unusual distribution. We explicitly do not assume that the analyst's goal is to describe the entire dataset succinctly, as is often the case for other data mining problems such as classification.

1.2 Scope and focus of the research

The central question addressed in this thesis is as follows: can Exceptional Model Mining benefit from a multi-objective approach? As such, the goals of this thesis are to provide insight into the various dimension of exceptionality of Exceptional Model Mining, and to improve on existing Exceptional Model Mining strategies by leveraging this knowledge.

So far, no restriction on the class of models to be used have been imposed in the definitions of Exceptional Model Mining, as long as in some way exceptionality or un-

[1] Note that this requirement immediately disqualifies concepts that are identified by collections of indices of subsets of the data. This separates the knowledge discovery strategy from other data mining techniques such as clustering.

usualness can be measured on these models. For measures and objectives for this broad range of models would be infeasible to analyse in a single thesis, we limit our scope to the Exceptional Model Mining task under linear regression models where model class specific objectives are discussed.

1.3 Structure of the thesis

In the sequel of this thesis, Section 1.4 first introduces the necessary notation and definitions. Chapter 2 then proceeds with a discussion of the existing literature on Exceptional Model Mining and the quality measures used therein. Here, some shortcomings related to the existing single-objective approaches and in particular their objectives are established. In Chapter 3, an empirical process is outlined where the notion of significance of unusualness of Exceptional Model Mining under linear regression models is investigated in relation to other objectives. The insights gained here form the foundation of further theoretical investigation of the quality measure often used for exceptionality of subgroups. The results of this further investigation are discussed and validated in Chapter 4. Further objectives relevant to Exceptional Model Mining are discussed in Chapter 5.

Having obtained a clearer view on the multi-objective nature of Exceptional Model Mining, an experiment is set up in Chapter 6 to analyse the merits of a multi-objective approach, in contrast to single-objective techniques from the literature. Results and discussion of the results are provided in Chapters 7 and 8. The conclusions on our analysis and experimentation are then presented in Chapter 9.

1.4 Notation and definitions

This section introduces some notational conventions and definitions used throughout this work.

Matrices are written uppercase as $\boldsymbol{M}$, vectors lowercase as $\boldsymbol{v}$, with transpose $\cdot^\mathsf{T}$, inverse $\cdot^{-1}$ and symmetric square root $\cdot^{1/2}$ where applicable. Observations from a finite population of variate A are written as a_i, indexed by their respective position in the population. The mean of a vector, variate or sample thereof is written throughout as $\bar{\cdot}$, where the expected value and (co)variance of a vector or a matrix are written as $\mathrm{E}(\cdot)$ and $\mathrm{Var}(\cdot)$ respectively. Estimates of a variable are written as $\hat{\cdot}$.

Definition 1.1 (Dataset). Let the dataset $\mathcal{D}$ denote a collection of n observations. For notational convenience, the relationship of the variables introduced hereafter to $\mathcal{D}$ is implicitly assumed unless otherwise specified, for in general we consider only a single dataset. For the purpose of Exceptional Model Mining, we decompose the attributes

of the dataset — usually represented as columns therein — into concept attributes and model attributes.

We allow restrictions on the concept attributes of a dataset to be imposed by rules, providing an antecedent of a pattern in the hypothesis space.

Definition 1.2 (Rule). Let a rule on a dataset be an expression in some language, consisting of constraints on the concept attributes of the dataset. A rule serves as a single description of a concept, although multiple rules may describe the same concept. Unless noted otherwise, a propositional language is assumed to be used.

In Exceptional Model Mining, a single model on the model attributes of the dataset is used in the consequent of a pattern in the hypothesis space. Throughout this work we assume this model to be a linear regression model.

Definition 1.3 (Linear regression). Let $\boldsymbol{X}$ be a $n \times p$ matrix of observations for p independent variables and $\boldsymbol{y}$ be a $n \times 1$ vector of observations for a single dependent variable, where columns in $\boldsymbol{X}$ and $\boldsymbol{y}$ are assumed to be disjoint columns in the model attributes of $\mathcal{D}$ from Definition 1.1. Under the linear regression model, $\boldsymbol{X}$ and $\boldsymbol{y}$ are related as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1.1}$$

with $\boldsymbol{\varepsilon}$ a $n \times 1$ vector of random unobservable errors, with $\mathrm{E}(\boldsymbol{\varepsilon}) = 0$ and $\boldsymbol{\beta}$ a $p \times 1$ vector of true linear regression coefficients. Then

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y} \tag{1.2}$$

is the $p \times 1$ least-squares estimate of $\boldsymbol{\beta}$ for given $\boldsymbol{X}$ and $\boldsymbol{y}$. In addition, we denote the fitted values as $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and the residuals under the least-squares estimate as $\hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{y}} - \boldsymbol{y}$.

Definition 1.4 (Subgroup). Define a subgroup G as the selection of $|G| = k < n$ observations in $\mathcal{D}$, with G^C its complement in $\mathcal{D}$ of size $|G^\mathsf{C}| = n - k$. We represent G as a sequence of k unique integers such that $1 \leq G_i \leq n$ for all $1 \leq i \leq k$, where the presence of an integer $j = G_i$ in this sequence represents the inclusion of the jth observation from $\mathcal{D}$ in G. Linear regression on subgroups is analogous to linear regression on the entire dataset as in definition 1.3, however considering $\boldsymbol{X}_G$ and $\boldsymbol{y}_G$ instead of $\boldsymbol{X}$ and $\boldsymbol{y}$ respectively, where $\boldsymbol{X}_G$ and $\boldsymbol{y}_G$ include only the observations in G. Here, $\hat{\boldsymbol{\beta}}_G$ denotes the least-squares estimate of $\boldsymbol{\beta}$ based only on the observations in G, and $\hat{\boldsymbol{\varepsilon}}_G$ their residuals under this estimate.

Optimization problems such as Exceptional Model Mining often require comparison of solutions subject to some quality measure. For multi-dimensional quality measures, only a partial ordering can be established where one solution may or may not dominate the other.

Definition 1.5 (Dominance). Consider two solutions $s_1, s_2 \in S$ in an optimization problem under a dimensional quality measure $\varphi : S \to R^d$ consisting of d objectives.

Without loss of generality, assume that all objectives in the quality measure are to be minimized. Then $s_1 \prec_\varphi s_2$ (say s_1 dominates s_2), if and only if $\forall i \in \{1, \ldots, d\} : \varphi(s_1)_i \leq \varphi(s_2)_i$ and $\exists i \in \{1, \ldots, d\} : \varphi(s_1)_i < \varphi(s_2)_i$.

Definition 1.6 (Incomparable). As implied by the partial ordering of the dominance relation, solutions s_1 and s_2 may be incomparable with respect to φ, if and only if neither one dominates the other: $s_1 \parallel s_2 \iff s_1 \not\prec s_2 \wedge s_2 \not\prec s_1$.

Chapter 2

Related work

A successful attempt at Exceptional Model Mining at the very least requires us to establish the exceptionality of such a model. To be able to judge this quality for a concept in the database, or more specific, a rule under evaluation, we need a measure that can be computed without user interaction. For we only consider linear regression models, we can restrict ourselves to the class of exceptionality measures for linear regression models. An overview of such measures is given in Section 2.1. This chapter then proceeds with a discussion of existing techniques for Exceptional Model Mining under linear regression models.

2.1 Exceptionality measures

Judging the exceptionality of a model boils down to making a comparison between two models, one the exceptional model and the other the base or reference model, establishing their distance in some sense. In Exceptional Model Mining, we generally consider the model on the entire dataset the reference model, where the exceptional model is the model on a subgroup of the dataset. This section proceeds with a discussion of the exceptionality measure most often encountered for linear regression, followed by a brief discussion of possible generalizations and variants.

2.1.1 Cook's distance

Cook's distance, introduced in [10], gives a measure of the influence of exclusion of a single observation on the least squares estimate of the regression coefficients of a linear regression model. Cook's distance is then later extended for measuring the influence of

the exclusion of a subgroup G [11] and the retention of G [21], where the latter distance measure is defined as in Definition 2.1.

Definition 2.1 (Cook's distance). Let $\boldsymbol{X}$, $\boldsymbol{y}$ and $\hat{\boldsymbol{\beta}}$ be defined as in definition 1.3, and G and $\hat{\boldsymbol{\beta}}_G$ as in definition 1.4. Then Cook's distance of G is

$$\mathrm{D}_G = \frac{(\hat{\boldsymbol{\beta}}_G - \hat{\boldsymbol{\beta}})^\mathsf{T} \boldsymbol{X}^\mathsf{T} \boldsymbol{X} (\hat{\boldsymbol{\beta}}_G - \hat{\boldsymbol{\beta}})}{p\hat{\sigma}^2} \tag{2.1}$$

where $\hat{\sigma}^2$ is the variance estimate of the residuals under the of the least-squares estimate with $n - p$ degrees of freedom [10, 11].

Duivesteijn et al. argue that Cook's distance is a perfect fit for Exceptional Model Mining under linear regression models. As will be discussed in Section 4.3, this claim is debatable. Chapter 4 provides further results on the applicability of Cook's distance for Exceptional Model Mining.

2.1.2 Weighted Modified Cook's distance

Krak and Feelders propose a number of modifications to Cook's distance in [38] to allow for non-crisp or soft subgroups in an attempt to make the distance measure differentiable. Krak and Feelders arrive at the quality measure as given in Definition 2.2. Note that the notation here differs slightly from the notation used by Krak and Feelders in [38].

Definition 2.2 (Weighted Cook's distance). Let w be an inclusion weight for every element in the dataset, and $\hat{\boldsymbol{y}}_{\boldsymbol{w}}$ be the weighted equivalent to $\hat{\boldsymbol{y}}_G$ from definition 1.4. Weighted Cook's distance is then defined as

$$\mathrm{D}_w = \frac{\bar{\boldsymbol{w}}}{n\hat{\sigma}^2} ||\hat{\boldsymbol{y}}_{\boldsymbol{w}} - \hat{\boldsymbol{y}}||^2. \tag{2.2}$$

Apart from the introduction of soft subgroups, other modifications become apparent. In an attempt to take relative support of a subgroup into account, a factor $\bar{\boldsymbol{w}}$ is introduced into the equation for Cook's distance, supposedly equivalent to the introduction of a factor $\frac{k}{n}$ for crisp subgroups. In Section 4.2.2 we will show that this correction factor is not optimal. To make the quality measure less dependent on the number of regression coefficients, the factor p^{-1} is dropped from the equation.

The resulting distance measure is used as quality measure for the Tree-Constrained Gradient Ascent algorithm as described in Section 2.2.2.

2.1.3 Other exceptionality measures

It is infeasible to provide a comprehensive overview of exceptionality measures, as there exist infinitely many ways to variate on exceptionality, depending on the specific needs

of the analyst. For instance, the analyst might not — or not exclusively — be interested in the slope or intercept of the model, but in finding some concept where a model's goodness-of-fit, or the magnitude of its regression coefficients differs exceptionally.

As for Cook's distance, Cook and Weisberg give a few variations on their distance measure in [12, Section 3.5.1]. It must however be stressed that other quality measures unrelated to Cook's distance may prove useful for linear regression models as well, as the analyst's requirements vary. The applicability of any variation on exceptionality thus remains specific to the domain of the dataset and the preferences of the analyst.

2.2 Search strategies

Breadth-first search is often the go-to algorithm when a decision space can be searched exhaustively. For common datasets, the decision space in Exceptional Model Mining may however be enormous, and exhaustive search infeasible. Finding concepts that exhibit exceptional models then requires heuristics, pruning, or other strategies to provide a more directed search. In this section, an overview of the existing search strategies for Exceptional Model Mining under linear regression models is given.

2.2.1 Beam Search

When guaranteed optimality can be trade off for decreased complexity, the search space in breadth-first search can be strongly reduced by only considering a limited beam of "best few" solutions for expansion on each depth of the search [42, Section 4.7]. This strategy, later known as Beam Search [43], has been successfully applied in different fields of research where guaranteed optimality is not required [4]. Leman et al. propose using this strategy as a general quality function agnostic search strategy for Exceptional Model Mining [40]. Krak and Feelders also use this approach as their reference algorithm in [38].

When used for knowledge discovery, Beam Search is best described as a breadth-first pruning search for propositional rules in conjunctive form. On each level of search, all refinements of the rules in the beam of the previous level by addition of one single conjunctive clause are enumerated and the best w included in the current level beam, where w is the beam width. The search starts with an empty rule, imposing no restrictions on the concept, and finishes when the search level reaches a predefined search depth d.

2.2.2 Tree-Constrained Gradient Ascent

Where Beam Search is agnostic of the exceptionality measure in use, the exploitation of this domain knowledge may lead to a performance improvement. Tree-Constrained

Gradient Ascent, presented by Krak and Feelders in [38], dynamically partitions the dataset into an included and an excluded part. This is represented as a tree where leaves represent mutually exclusive sets of dataset rows, and each leaf is assigned a soft inclusion weight. The tree's leaves together cover the entire dataset. All positively weighted leaves — leaves with inclusion weight greater than 50% — then represent the "included" partition of the dataset. Growing of the tree — splitting leaves — is interleaved with gradient ascent update steps of the weights of their respective leaves. In this process, a fixed gradient step size is used. This entire process of iterative splitting and updating is repeated until convergence is reached, or the maximum number of iteration is exceeded.

Quality measure

Tree-Constrained Gradient Ascent can, in general, use any quality measure that allows for soft (i.e. weighted) subgroups and is differentiable to the inclusion weights. Krak and Feelders propose a weighted variation on Cook's distance for linear models that takes into account the support of a subgroup. A definition of this quality measure is given in Section 2.1.2.

Krak and Feelders also provide the derivative of this quality function with respect to the weights. It must be noted that the derivation of the derivative of their quality measure introduces an erroneous 2^{-1} and therefore their derivative appears a constant factor 2 too small.

Apparent pitfalls

Apart from the minor error in the derivative of the quality measure in [38], two potential pitfalls become apparent on further analysis of the methods used in Tree-Constrained Gradient Ascent.

Problematic initialization

Krak and Feelders do not describe in [38] how the weights in their initial solution are chosen. From their implementation, generously made available, and Krak's original work [37] we learn that the initial weights are found as

$$\boldsymbol{w}_i = \begin{cases} 1 - \frac{\mathrm{D}_i}{\max_j \mathrm{D}_j} < \boldsymbol{r}_i, & f(\mathrm{D}_i) \\ \text{otherwise}, & 0 \end{cases} \tag{2.3}$$

with $\boldsymbol{r}_i \sim \mathcal{U}(0,1)$ for all $1 \leq i \leq n$, D_i Cook's distance when the ith observation is removed from the dataset as in [10], and $f(x) = (1 + \epsilon^{-x})^{-1}$ the logistic sigmoid function.

As may be noted, the probability of exclusion of any observation from the initial inclusion set is inversely linear proportional to the maximum individual Cook's distance in the dataset. As such, this choice of initialization does severely limit the spread of initial solutions when the dataset has a few very strong outliers.

Biased derivative

In gradient ascent, the gradient is usually ascended iteratively until convergence. Krak and Feelders however opt for using only a single step of fixed step size. The question as to why this choice is made remains unanswered in [37] and [38]. In [38], Krak and Feelders do however mention: "currently only a single gradient ascent update step is performed in between two consecutive splitting steps. It would be interesting to investigate the effect of performing more update steps." While interesting, the effect of performing more update steps is likely to be detrimental. Quick analysis of the derivative in use show that it is not entirely unbiased. Krak and Feelders apparently failed to recognize that any positive scalar multiple of inclusion weights yields the exact same estimates in the linear model. Failing to recognize this, a linear dependence on the magnitude of w is introduced in the derivative, which introduces a bias towards $w = 1$. This is likely to negatively affect the results of greater step sizes or repeated updating.

Chapter 3

Exploring existing techniques under multiple objectives

The previous chapter discussed currently available techniques for Exceptional Model Mining under linear regression, all of which are designed to optimize the single objective of exceptionality. In an effort to create an understanding of the behaviour of these algorithms under this and alternative plausible objective functions, some explorative experimentation is performed. The sequel of this chapter describes these efforts and the results obtained. It must be noted that the results reported in this chapter are not the main contribution of this work, and that given their explorative nature, they may lack proper justification of decisions made.

3.1 Spread of solutions in a multi-objective space

Due to their single-objective nature, the existing techniques introduced in Section 2.2 only seek to optimize for evidence on exceptionality. Still, the spread of their respective solutions in a multi-objective solution space can be measured, providing an interesting insight into the multi-objective behaviour of these techniques. Some techniques may, for instance, tend to find small subgroups, where others tend to find subgroups with a particular goodness-of-fit. We assume Cook's distance, compensated by a factor k for subgroup size as suggested by Krak and Feelders [37, 38], is used as this single objective.

To gain insight in this spread, or rather behaviour under multiple objectives, an experiment is setup where the solutions encountered by the existing techniques are measured for different plausible objective measures on different datasets. Evidence on exceptionality, the single-objective quality, is measured by means of k-scaled Cook's

Dataset	n	**Model**	R^2
Ames Housing	2930	$SalePrice \sim \beta_0 + \beta_1 \cdot LotArea + \beta_2 \cdot OverallQual$	0.675
Windsor Housing	546	$sell \sim \beta_0 + \beta_1 \cdot lot + \beta_2 \cdot bdms + \beta_3 \cdot fb + \beta_4 \cdot sty$	0.536
Wine	9600	$Price \sim \beta_0 + \beta_1 \cdot Cases + \beta_2 \cdot Score + \beta_3 \cdot Age$	0.313

Table 3.1: Datasets with their respective sizes and models in use for explorative experimentation.

distance. The subgroup's support is measured as a measure of generality. Goodness-of-fit is measured as R^2, as a measure of confidence in the resulting model. Finally, the number of terms in a rule is measured to establish the simplicity of a rule (or rather its complexity). The datasets and models to be explored form a small arbitrary sample from those used in Krak and Feelders' comparative experiment in [38]. This selection consists of the *Ames Housing* [15], *Windsor Housing* [1] and *Wine* [14] datasets. These datasets are then used in conjunction with their respective models as in [38]. Table 3.1 provides an overview of the datasets and models used.

3.1.1 Algorithmic settings

For all techniques, a minimum support $k_{min} = 50$ is used, as earlier suggested by Krak and Feelders in [38]. It is expected that this provides reasonable protection against detection of spurious patterns, that would have been expected when the minimum support approached the number of regression coefficients in a model.

Beam Search is used with beam width $w = 50$ and search depths depending on the dataset, chosen such that patterns of reasonable simplicity given the size of the dataset are expected. For Ames Housing and Wine, this amounts to search depths up to 6. For Windsor Housing, search depths up to 4 are considered. Our own implementation of the Beam Search algorithm in the R programming environment [48] is used.

While Krak generously provided his C++ implementation of the Tree-Constrained Gradient Ascent algorithm, as used to obtain the results provided in [38], this implementation proved difficult to adapt and instrument for our specific purposes. Hence for experiments involving Tree-Constrained Gradient Ascent, our own implementation of the algorithm in the R programming environment is used. In this implementation the post-processing step is omitted, as any post-hoc pruning of rules does not affect the solution space the algorithm is able to explore and can only reduce the exceptionality under Cook's distance. Throughout our experiments, a minimum split size $min\_split = 125$ and step sizes $\{0.1, 10\}$ are used, as suggested by Krak and Feelders in [38]. As Tree-Constrained Gradient Ascent returns only a single solution for each run, this experiment is repeated 1000 times. For each run, the step size is chosen randomly.

3.1.2 Results

Figures 3.1, 3.2, 3.3, and 3.4 show the results obtained for the different experiments on Windsor Housing and Wine in all their facets. In these figures, the pairwise relations between evidence (*quality*), confidence (*r.squared*), generality (*support*), and complexity (*terms*) of the results are shown. The figures for Ames Housing are omitted, as an initial bug in the implementation lead to degenerated results for Tree-Constrained Gradient Ascent, providing little insight. This bug was later corrected, but no new figures were generated as the other results appeared sufficiently insightful.

It must be noted that the number of terms for Beam Search solutions cannot be compared to the number of terms for solutions generated by Tree-Constrained Gradient Ascent, since both techniques describe solutions in a different propositional language.

What can readily be seen from the figures is that Tree-Constrained Gradient Ascent yields only a limited number of distinct solutions, a number far smaller than the number of times the algorithm is repeated. This provides tangible evidence for a flawed initialization function, as we earlier suggested in Section 2.2.2.

A notable difference between the solutions generated by Beam Search and Tree-Constrained Gradient Ascent is that the former tends to generate higher-quality solutions for increased rule complexity, where solutions from the latter technique appear to exhibit an opposite quality—complexity relation.

Figures 3.3 and 3.4 show that for the Wine dataset, Tree-Constrained Gradient Ascent exhibits a strong bias towards subgroups with extremely low goodness-of-fit, while Beam Search is able to recover subgroups for which the goodness-of-fit is far above that of the model on the full data set (see Table 3.1).

With both algorithms, a strong bias towards solutions of low generality can be observed for the Wine dataset. For Beam Search on Windsor Housing, in Figure 3.1, goodness-of-fit and support are clearly positively correlated, where standard linear regression theory would predict a negative correlation for random samples. Plenty of the solutions generated, even though they feature a high quality in the sense of a scaled Cook's distance, must be considered degenerate from their low goodness-of-fit and might as well be noise in the dataset. From this it appears that the correction factor k, as suggested by Krak and Feelders in [38], is not sufficient to drive either algorithm away from generating degenerate solutions.

3.2 On evidence and support

From the results outlined in the previous section, we observed that correction by a factor k is not sufficient to prevent a bias of Cook's distance towards degenerate solutions. To

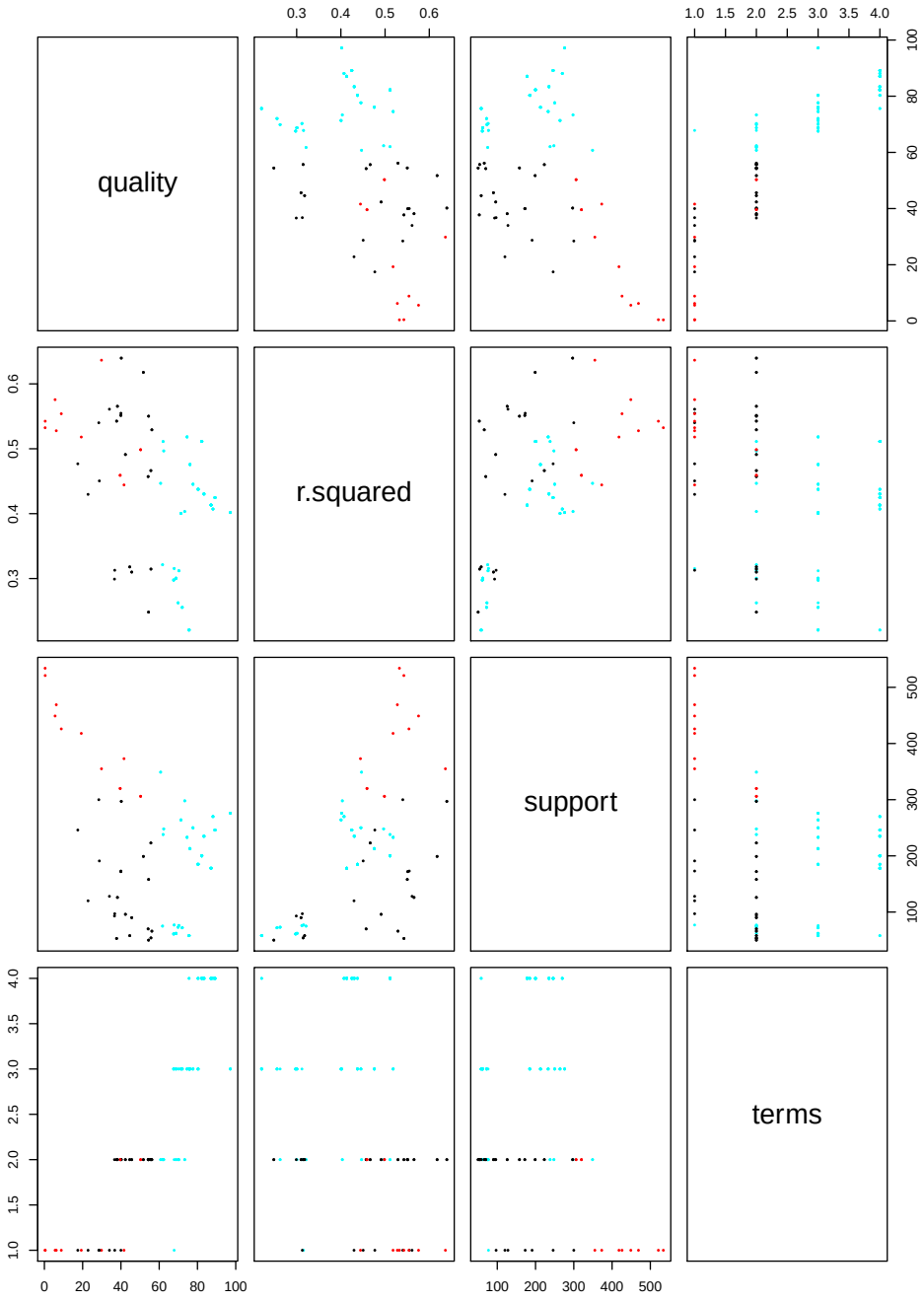

Figure 3.1: Solutions obtained from Beam Search on the Windsor Housing dataset, with search depths $\{1, 2, 3, 4\}$. Results with evidence above 60 are colored blue, results with generality above 50% are colored red.

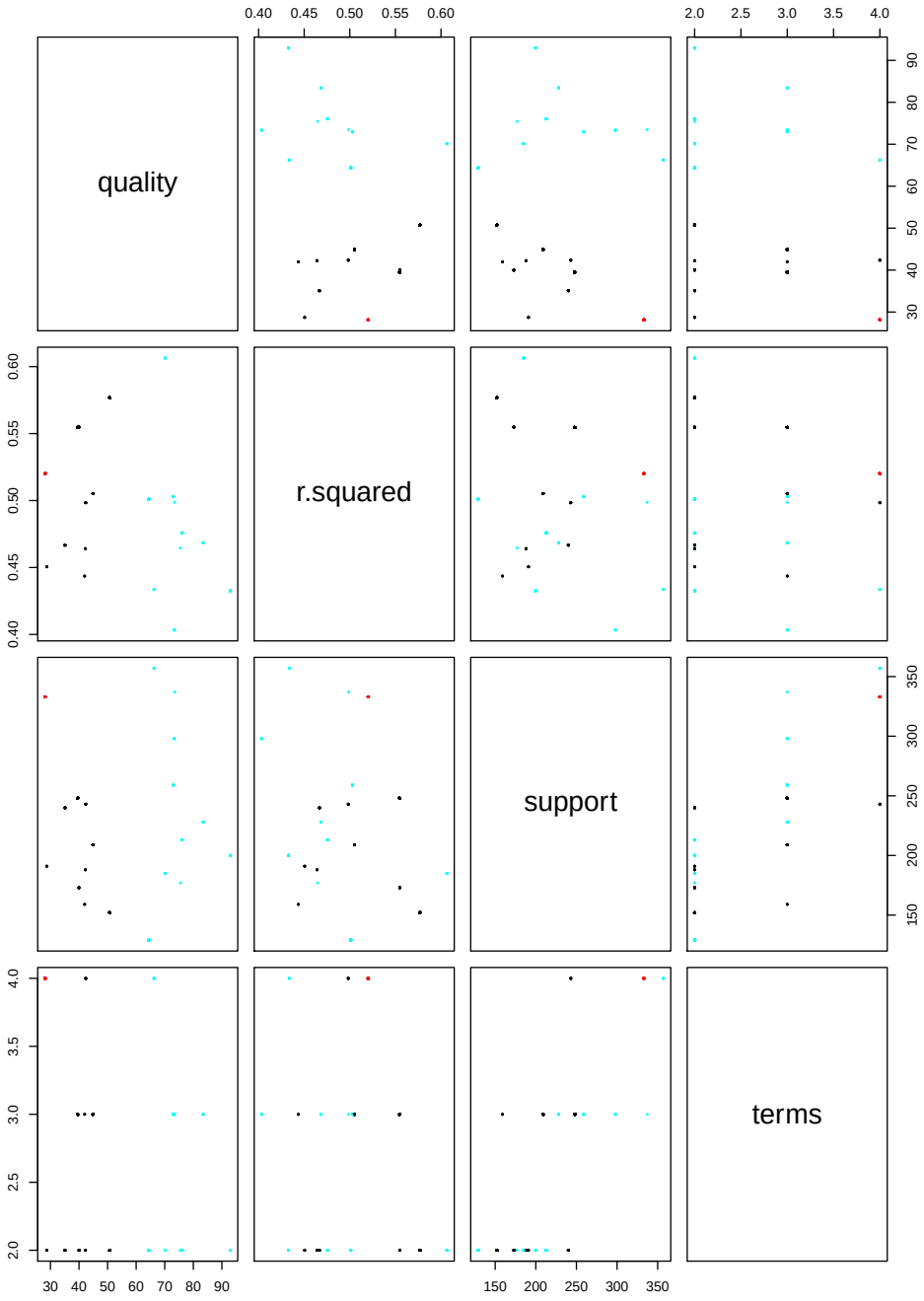

Figure 3.2: Solutions obtained from Tree-Constrained Gradient Ascent on the Windsor Housing dataset. Results with evidence above 60 are colored blue, results with generality above 50% are colored red.

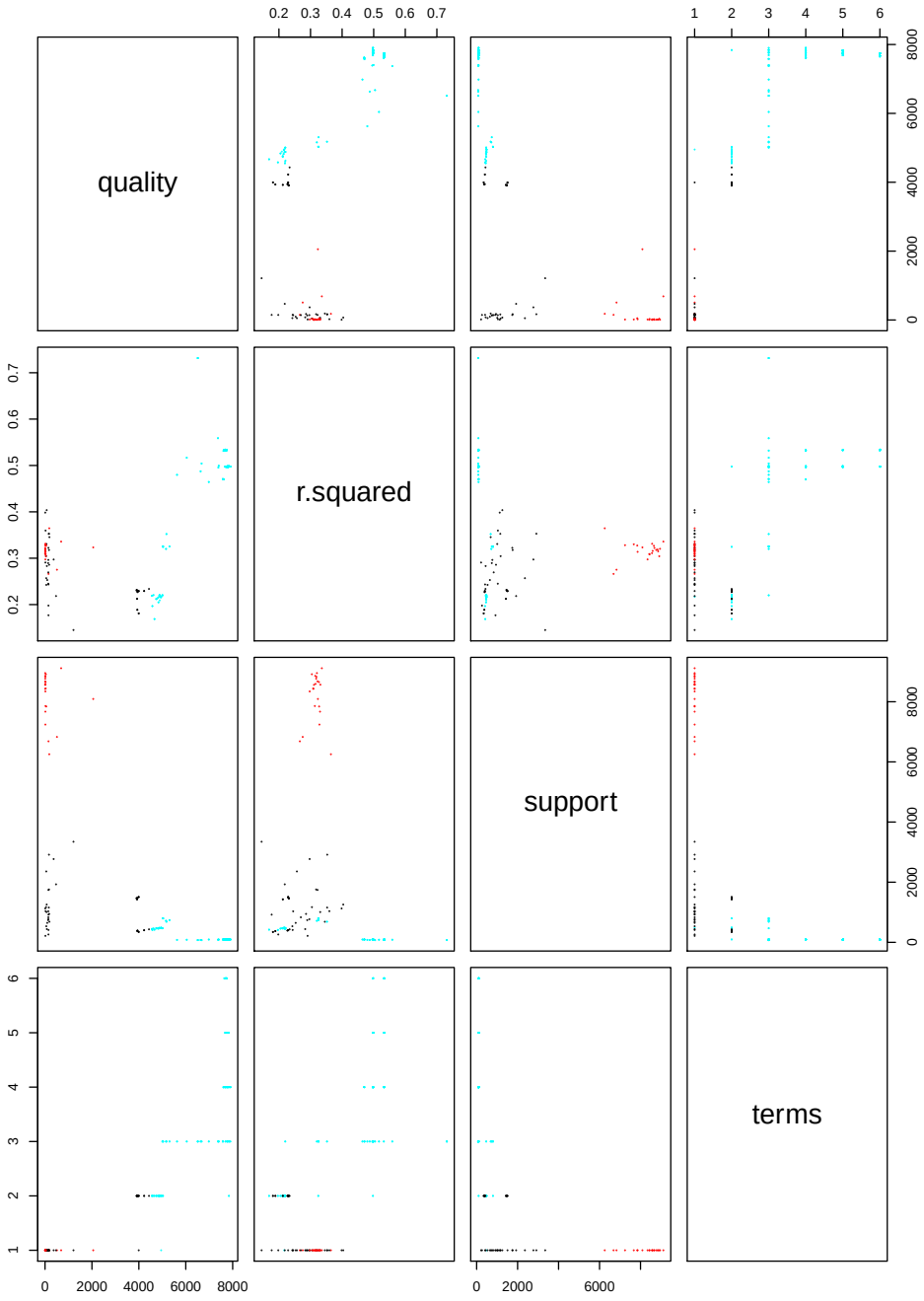

Figure 3.3: Solutions obtained from Beam Search on the Wine dataset, with search depths $\{1, 2, \ldots, 6\}$. Results with evidence above 4500 are colored blue, results with generality above 50% are colored red.

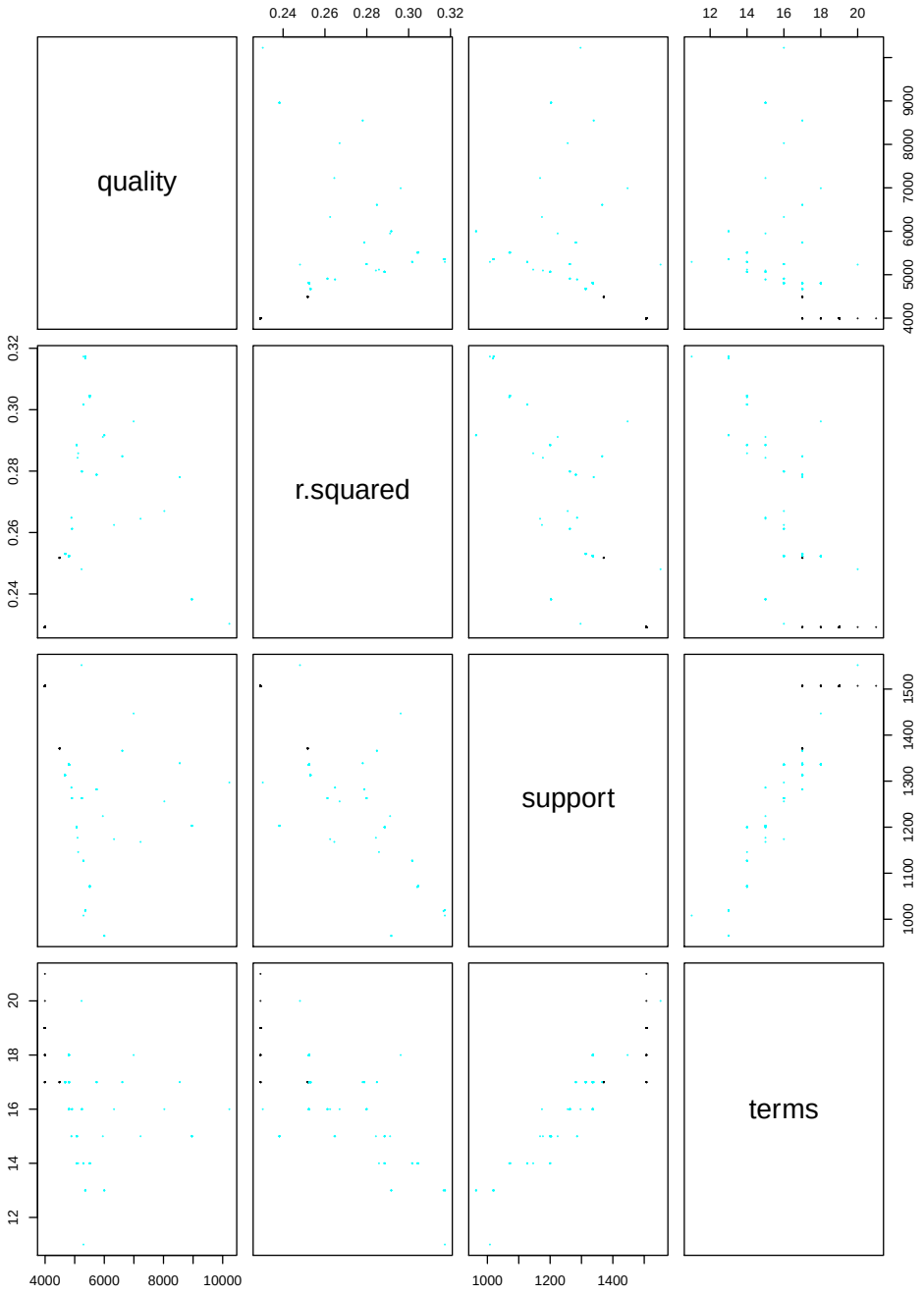

Figure 3.4: Solutions obtained from Tree-Constrained Gradient Ascent on the Wine dataset. Results with evidence above 4500 are colored blue, results with generality above 50% are colored red.

measure exceptionality by means of a statistically sound evidence measure, as intended by Klösgen, such a bias must be eliminated. This gives rise to an obvious question: what model, if any, would provide a sufficient correction? In an initial attempt to answer this question, we start by observing the distribution of Cook's distance for subgroups in our datasets.

Consider the distribution of Cook's distance of a random subset of k data points, chosen uniformly from all observations in the dataset without replacement. We empirically derive the probability density for such a distribution within each of our datasets outlined in the previous section, for each k, by means of sampling and density estimation in the R programming environment.

Figure 3.5 shows the probability densities measured over the Wine dataset for different supports k, on a logarithmic distance scale. On this scale, within each value of k, the measured density has the general appearance of a normal distribution. This allows to guess that within a value of k, Cook's distance follows a lognormal distribution, that is to say $\mathrm{D}_k \sim \ln \mathcal{N}(\mu_k, \sigma_k^2)$ for some μ_k decreasing in k and constant σ_k^2. The probability densities of the maximum likelihood estimate of the lognormal distribution are provided as an overlay in Figure 3.5. Figures for Ames Housing and Windsor Housing show distributions of similar shape, and are omitted for brevity.

Analysis of the mean values for these distributions with respect to their support can now provide some insight in the scaling behaviour of Cook's distance in k. Figure 3.6 shows these mean values and their single-sigma confidence intervals under the maximum likelihood lognormal distribution. Clearly, the scaling behaviour is far from equivalent to k^{-1}. It should now be obvious that scaling of Cook's distance with a factor k does not properly eliminate its bias towards smaller subgroups — let alone to bias it towards larger subgroups that may be of greater interest to the analyst, as was intended by the introduction of the scaling factor in [38].

In order to properly compensate for support bias in Cook's distance, an empirical approach may be taken: a look-up table of means can be constructed in a way similar to how Figure 3.6 is constructed, and the values used as a correction factor. This does however require estimating probability densities from the data, which is not only computationally expensive, but also doesn't provide additional insight in the mechanisms behind the underlying distribution. As such, we proceed with a theoretical approach — rather than the current empirical approach — to this distribution in the following chapter.

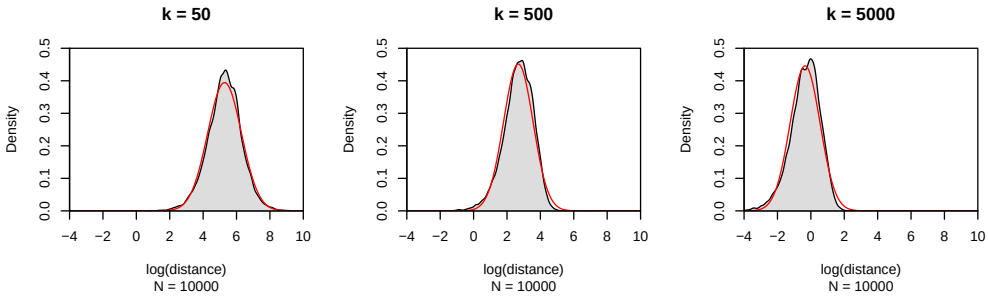

Figure 3.5: Empirical distribution function of Cook's distance for the Wine dataset with $k = 50$, $k = 500$, $k = 5000$ from left to right respectively, based on 10,000 samples. The red overlay represents the maximum likelihood estimate of a lognormal distribution for the empirical distribution.

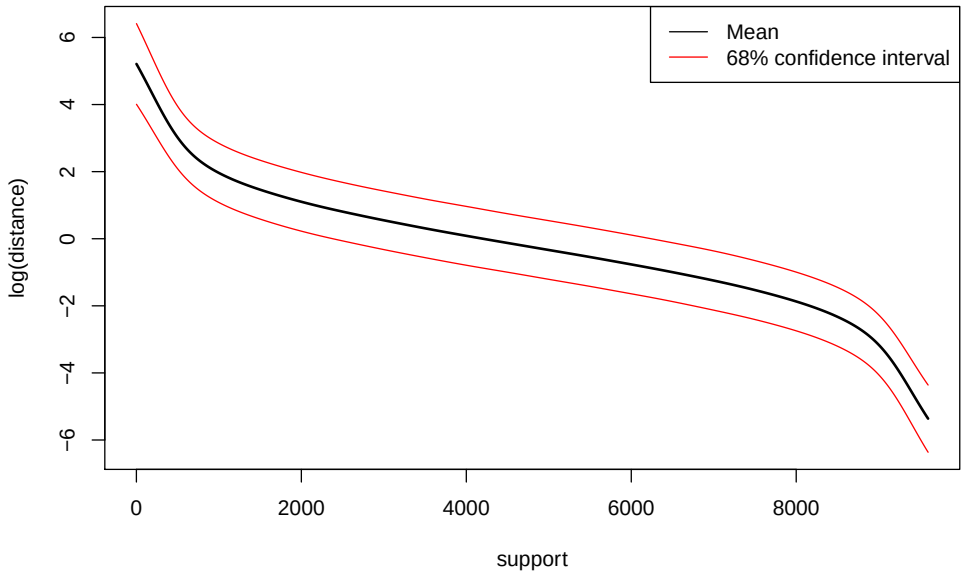

Figure 3.6: Empirically determined scaling behaviour of Cook's distance for the Wine dataset on a logarithmic scale. Based on 10,000 samples each for 50 values of support k, sampled uniformly over the size of the dataset. Mean and single-sigma confidence interval for each k are based on the maximum likelihood lognormal distribution.

Chapter 4

Cook's distance

Originally introduced as a measure for detecting outliers or otherwise influential data points, Cook's distance provides a measure for the deviation in regression coefficient estimate by selection of a subgroup from the full dataset. In the previous chapter, we observed empirically that Cook's distance does not scale inversely linear in k, the support of the subgroup, which might disqualify this distance measure from being used as a statistically sound evidence measure without further compensation than suggested in [38]. However, these empirical results don't provide a model of Cook's distance with any form of proof or theoretical justification. This chapter focuses from a theoretical perspective on the numerical distribution of Cook's distance, in an attempt to establish its suitability for measuring evidence of exceptionality in Exceptional Model Mining.

We first introduce some additional definitions in Section 4.1, then discuss the related work on the subject in Section 4.2. We proceed by briefly reviewing the suitability of Cook's distance and related distance measures for Exceptional Model Mining, based on knowledge from the existing literature, in Section 4.3. Subsequently in Section 4.4, we derive an approximate asymptotic distribution of Cook's distance and propose an alternative distance measure that better suits Exceptional Model Mining. A brief experimental comparison between the two is then given in Section 4.5.

4.1 Definitions

In addition to the notation and definitions introduced in Section 1.4, we introduce two additional definitions that are required for further analysis of Cook's distance.

Definition 4.1 (Generalized Cook's distance). Let the generalized Cook's distance for a $p \times p$ design matrix $\boldsymbol{M}$ be defined by

$$\mathrm{D}_G(\boldsymbol{M}) = (\hat{\boldsymbol{\beta}}_G - \hat{\boldsymbol{\beta}})^\mathsf{T} \boldsymbol{M} (\hat{\boldsymbol{\beta}}_G - \hat{\boldsymbol{\beta}}). \tag{4.1}$$

Specialization to the usual definition of Cook's distance from Definition 2.1 gives the equivalence

$$\boldsymbol{M} = \frac{\boldsymbol{X}^\top \boldsymbol{X}}{p\hat{\sigma}^2}. \tag{4.2}$$

This definition is equivalent to the generalization Cook and Weisberg introduce in [11]. Cook and Weisberg however treat the design matrix and the denominator of the distance measure separately, for which we find no need.

Definition 4.2 (Simple random sampling). Simple random sampling is the random sampling strategy where k units out of a finite population of n units labeled $[1, \ldots, n]$ are obtained by selection with equal probability for each unit and without replacement. We represent a simple random sample as a sequence of the labels comprising the sample, given in their canonical (i.e. numeric) order.

4.2 Related work

Cook immediately refers to the Fisher–Snedecor distribution (F-distribution) when he first publishes about his proposed measure in 1977, proposing the median of the F-distribution as a cutoff level for further analysis [10]. This suggestion has proven to be a great source for later confusion, criticism and debate [27, 30, 45]. In none of his works does Cook derive the distribution of his distance measure. Since, several authors have published on their efforts on deriving a distribution of Cook's distance. We present here a selection of different views on Cook's distance and the derivation of its distribution, and briefly discuss their applicability to Exceptional Model Mining.

Cook's distance as updating in regression

Muller and Mok provide in [45] an extensive review of the distribution of Cook's distance for linear regression problems with independent and identically distributed Gaussian errors, for cases where the subgroup misses only a single data point from the entire dataset, i.e. $|G^{\mathsf{C}}| = 1$.

Their approach relies on their observation that Cook's distance of a single data point removal is a measure combining the leverage, i.e. the corresponding diagonal element of the hat matrix, and the squared residual error of this data point. Assuming that the dataset at hand has independent and identically distributed Gaussian errors, they relate the leverage to an F-distribution and relate the squared residual (conditional to the leverage) to a β-distribution, and then derive a combined distribution function through integration. In addition, Muller and Mok provide useful computational forms of the distribution function that allow for computation by numerical integration, as well as several forms that are inexact but easier to compute.

The results by Muller and Mok are strongly related to the results by Beckman and Trussell, who earlier described the effects of updating a regression by adding an additional data point on the regression coefficient estimates, and the distribution of the studentized residual of the added data point [3]. It follows immediately from their results that change in residuals depends on the data point added, introducing correlation between residuals. Removal of a data point can of course be seen as the reversal of the process described by Beckman and Trussell, having similar impact on the correlation of residuals. This correlation makes the results by Muller and Mok unlikely to allow for generalization to the case where the subgroup is determined by the deletion of multiple data points. For this reason, their results can be considered essentially useless for application in Exceptional Model Mining, where subgroups of mostly any size are considered. In addition, Muller and Mok assume in their analysis that the errors are independent and identically distributed Gaussians, which is an assumption that we cannot reasonably expect to hold for datasets we may encounter "in the wild".

Cook's distance as a function of eigenvalues

Jensen and Ramirez take quite another approach to modelling the distribution of Cook's distance [30, 32]. They first decompose linear regression to a canonical decomposition, that is, by expressing the linear regression on a subgroup G as a function of the eigenvalue decomposition of the subsetted hat matrix of the subgroup's complement, i.e. $\boldsymbol{H}_{G^\mathsf{C}} = \boldsymbol{X}_{G^\mathsf{C}}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}_{G^\mathsf{C}}{}^\mathsf{T}$. This decomposition essentially allows for prediction of $\boldsymbol{X}_{G^\mathsf{C}}$ based on $\hat{\boldsymbol{\beta}}$ from the full data, which forms the foundation of their work. Then, by assuming the random errors $\boldsymbol{\varepsilon}$ to be independent and identically distributed Gaussians, they are able to derive a distribution related to a generalized F-distribution for Cook's distance when $|G^\mathsf{C}| \leq p$, dependent on the eigenvalues of the aforementioned $\boldsymbol{H}_{G^\mathsf{C}}$.

While interesting in nature, this distribution is not tabulated as far as we are aware, and may be hard to calculate. Jensen and Ramirez claim to have developed an algorithm for its computation [31], but their routines appear to have never been published. Fortunately, they show in their original work that said distribution is bound from below by a scaled but otherwise unmodified F-distribution, which may prove a more usable approach when an exact result is not required.

Unfortunately their results do not extend to the case where $|G^\mathsf{C}| > p$. While not explicitly noted in the work by Jensen and Ramirez, it can easily be seen that any $\boldsymbol{H}_{G^\mathsf{C}}$ for $|G^\mathsf{C}| > p$ is degenerate of rank $\leq p$, for it is a product of matrices with rank $\leq p$. Hence, $\boldsymbol{H}_{G^\mathsf{C}}$ has at most p non-zero eigenvalues, from which at most p rows of $\boldsymbol{X}$ can be reliably predicted: the prediction of more rows is underdetermined.

As with the results from Muller and Mok, the techniques of Jensen and Ramirez are hardly applicable to our Exceptional Model Mining problem at hand due to the inherent subgroup size limitation and the assumption of normality.

Cook's distance as a matrix trace

Díaz-García et al. analyze the distribution of Cook's distance (and some proposed variants) for the multivariate multiple linear regression model in [17, 18]. Under this model, a $n \times v$ matrix $\boldsymbol{Y}$ is considered instead of the vector $\boldsymbol{y}$ in the linear regression equation. $\boldsymbol{\beta}$ then becomes a $p \times v$ matrix $\boldsymbol{B}$. If we choose $v = 1$ however, the resulting model is the univariate multiple linear regression model and we can apply their results as usual. In the sequel of this discussion, we will assume $v = 1$ and consider the case where $|G^{\mathsf{C}}| > 1$.

In their work, Díaz-García et al. rewrite Cook's distance as

$$\begin{aligned} \mathrm{D}_G &= \frac{1}{\hat{\sigma}^2 |G^{\mathsf{C}}|} \operatorname{tr}(\hat{\boldsymbol{\varepsilon}}_{G^{\mathsf{C}}} \hat{\boldsymbol{\varepsilon}}_{G^{\mathsf{C}}}^{\mathsf{T}} (\boldsymbol{I} - \boldsymbol{H}_{G^{\mathsf{C}}})^{-1} \boldsymbol{H}_{G^{\mathsf{C}}} (\boldsymbol{I} - \boldsymbol{H}_{G^{\mathsf{C}}})^{-1}) \\ &= \frac{n-p}{|G^{\mathsf{C}}|} \operatorname{tr}({\boldsymbol{P}_G}^{1/2} \boldsymbol{Q}_G {\boldsymbol{P}_G}^{1/2}) \end{aligned}$$

with $\boldsymbol{P}_G^{1/2} = {\boldsymbol{H}_{G^{\mathsf{C}}}}^{1/2} (\boldsymbol{I} - \boldsymbol{H}_{G^{\mathsf{C}}})^{-1/2}$, where they assert without much justification that $\boldsymbol{Q}_G$ then has a matrix-variate beta distribution, under the assumption that $\boldsymbol{\varepsilon}$ are independent and identically distributed Gaussians. Thus, it is said that ${\boldsymbol{P}_G}^{1/2} \boldsymbol{Q}_G {\boldsymbol{P}_G}^{1/2}$ has a generalized matrix-variate beta distribution, and Cook's distance has the distribution of the trace of said generalized matrix-variate distribution.

While their results appear useful for Exceptional Model Mining, as they do not impose any restriction on the size of $|G^{\mathsf{C}}|$, Díaz-García et al. mention that the distribution they have derived has not yet been tabulated. Their analysis also lacks results on values of the parameters of their conjectured generalized matrix-variate beta distribution. Hence, while their results are interesting from an entirely theoretical point of view, they are by themselves unusable in a practical setting.

Cook's distance as an external norm measure

Gray takes a more general approach in his works on influence diagnostics, of which Cook's distance can be considered a special case. In his approach in [27], Gray differentiates between two classes of influence measures: external and internal norms. Both take the form of a generalized Cook's distance as shown in equation 4.1, however differing slightly in the choice of $\boldsymbol{M}$. He pinpoints the difference between both to the choice of reference set considered. For external norms, this reference set is comprised of all possible estimates $\hat{\boldsymbol{\beta}}$ found as least squares estimates by repeated sampling of $\boldsymbol{y}$ for fixed $\boldsymbol{X}$. For internal norms however, this reference set is comprised of the $\hat{\boldsymbol{\beta}}_G$ for all $\binom{n}{|G|}$ subgroups of fixed $\boldsymbol{X}$ and fixed $\boldsymbol{y}$. Put differently, internal norms are conditional on $\boldsymbol{X}$ and $\boldsymbol{y}$, whereas external norms are conditional only on $\boldsymbol{X}$.

In significance testing, the difference between the two classes described by Gray exhibits as a difference in the null hypotheses they test against, and hence in the design matrices $\boldsymbol{M}$ employed. Ignoring any potential shift in mean for the full model (that

is, under the assumption that $\hat{\boldsymbol{\beta}}$ is representative for $\boldsymbol{\beta}$), the external null hypothesis is effectively

$$H_0 : \hat{\boldsymbol{\beta}}_G = \boldsymbol{\beta}, \tag{4.3}$$

for which the external norm then provides a statistic. Internal norms, however, provide a statistic for the conditional null hypothesis

$$H_0 : \hat{\boldsymbol{\beta}}_G = \hat{\boldsymbol{\beta}} \tag{4.4}$$

for fixed $\hat{\boldsymbol{\beta}}$.

Cook's distance, in its original form, clearly measures the deviation in the regression estimate as a function of the covariance estimate on the assumed model, for it defines $\boldsymbol{M} = \hat{\sigma}^{-2}\boldsymbol{X}^\mathsf{T}\boldsymbol{X} = \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})^{-1}$. It follows that Cook's distance is a member of the class of external norm measures. This is not surprising, since Cook's distance was presented as a measure for detecting outliers — data points that deviate significantly from the (estimated) true model.

Gray suggests that internal norm measures are better for finding subsets with high influence or disproportionate effect when compared to other subsets in the same data problem [27]. This suggests that this class of influence norms are better suited for Exceptional Model Mining. Section 4.3 provides further discussion and results on this subject.

A computational approach
Kim and Storer take a computational approach to the problem of finding appropriate cutoff values for Cook's distance in [34]. Instead of deriving a theoretical distribution, they empirically determine their distribution parameters of interest by means of Monte Carlo simulation. Their method operates on a designed hat matrix instead of one computed from $\boldsymbol{X}$, thereby completely ignoring the descriptive information available in the dataset. Their results hence lack descriptive power. Considering this and the computational overhead associated with such an empirical approach, we will not further discuss these results but continue to focus on a more theoretical approach to Cook's distance instead.

4.2.1 Variations on Cook's distance

Gray proposes an internal norm influence measure as an alternative to Cook's distance. In this norm, the matrix $\boldsymbol{M}$ is one of the pre-computed covariance matrix estimates for each deletion size $|G^\mathsf{C}|$ [27]. For $|G^\mathsf{C}| = 1$ this estimate is easy to compute, for there are only n of such subgroups and exhaustive enumeration is possible. Generally, there are $\binom{n}{|G^\mathsf{C}|}$ subgroups for given $|G^\mathsf{C}|$. It is easily seen that exhaustive enumeration is infeasible for $|G^\mathsf{C}| \gg 1$ (or $|G| \gg 1$ by symmetry), and the covariance estimate must always be

approximated, e.g. by sampling. In his later work, Gray gives an approximation of these covariance matrices based on scaling the covariance matrix estimate for $|G^{\mathsf{C}}| = 1$ [26].

Countless other variations on Cook's distance have been proposed, including but not limited to those presented in [27, 30, 45]. These variations often detect slightly different outliers, but are mainly proposed to allow for a simpler analysis of their distance distributions. Surprisingly, none of the proposed alternatives appear to belong to the class of internal norms.

4.2.2 Scaling behavior under subgroup size

It may now be obvious to the reader that the generalized Cook's distance is not invariant under change in subgroup size for fixed design matrix $\boldsymbol{M}$. As such, any derivation of the distance distribution must take this factor into account.

To start with the exceptional case, Muller and Mok need not compensate for this variation, as they consider only subgroups with $|G^{\mathsf{C}}| = 1$.

While Jensen and Ramirez do correct for the scale difference in the expected value of Cook's distance related to different subgroup sizes, they do this by correcting with a factor $|G^{\mathsf{C}}|$ [32], which Gray shows to be flawed for large $|G^{\mathsf{C}}|$ [26] and for which we find results similar to those of Gray in section 4.4. For large n and $|G^{\mathsf{C}}| \leq p$ with $p \ll n$, the scaling proposed by Jensen and Ramirez gives a reasonable approximation. Note however that these conditions usually do not hold for Exceptional Model Mining.

The results reported by Díaz-García et al. are unclear about the exact scaling in terms of $|G^{\mathsf{C}}|$. Although a correction factor $|G^{\mathsf{C}}|$ is introduced in the analysis, Díaz-García et al. provide no clues on how the trace of the matrix variate beta distribution of their choice, for which they have omitted the required parameters, scales in the size of the matrix diagonal and thereby in $|G^{\mathsf{C}}|$.

4.3 Exceptional Model Mining and external norms

Duivesteijn characterizes quality measures for Exceptional Model Mining as follows: "The typical quality measure in EMM indicates how exceptional the model fitted on the targets in the subgroup is, compared [...] to the model fitted on the targets in the whole dataset" [23]. For a usual definition of exceptional as "unlikely to have occurred by chance", a chosen quality measure cannot be appropriate for Exceptional Model Mining unless some statistic can be provided that takes into account the expected quality of a randomly chosen subgroup. Without this statistic, comparison of subgroups of different sizes is infeasible, and the quality measure may not be well-suited for Exceptional Model Mining.

So far, Cook's distance in its original form has been proposed and used for judging the distance of subgroup regression models in Exceptional Model Mining [22, 23, 38]. Quality measures that take the general form of Cook's distance (see Definition 4.1) measure the distance to the model fitted on the full dataset however weighted by some matrix $\boldsymbol{M}$. For models with $p > 1$ however, the choice of this design matrix has a strong influence on which subgroups are considered exceptional.

As we have argued in the discussion of Gray's results, external norm influence measures — and therefore Cook's distance in its original form — measure the deviation of a model based on an estimate of the asymptotic properties of the model for an infinite dataset. How to interpret Cook's distance then depends on how well the regression estimates for the dataset have converged to their true values. Due to the questionable assumptions (such as normality) under which all of the distributions for Cook's distance in the literature have been derived, none of these results give us a statistic to measure the true exceptionality of Cook's distance, or external norm influence measures in general, for real datasets. Lacking such a statistic, external norm influence measures must be considered ill-suited to Exceptional Model Mining.

4.4 Results on Cook's distance

Muller and Mok, Jensen and Ramirez and Díaz-García et al. all fail to recognize [18, 30, 32, 45] that the distribution of Cook's distance for subgroups of a dataset is related to repeated sampling from a finite population without replacement. Gray does recognize this important aspect when distinguishing between internal and external norms [27] and establishing his results on the approximate scaling of the design matrix of his proposed internal norm [26], but he does not give a distribution for either this norm or Cook's distance.

The sequel of this section discusses our results on the distribution of the regression coefficient estimates and Cook's distance where the subgroup G is chosen by simple random sampling. We first show some basic properties of least-squares regression coefficients for these subgroups, and then proceed by showing how this relates to Cook's distance. Finally, we present an alternative internal norm influence measure based on the generalized Cook's distance from Definition 4.1 for which we establish an approximate distribution.

4.4.1 Regression coefficient estimates in samples from a finite population

Cochran has developed an approximation to the expected value and the variance of the ratio of two univariate random variables in [9, theorem 2.5]. In this section, we generalize

Cochran's results to the multivariate case in a way similar to what Gray proposes in the appendix of [26], and show how linear regression relates to the ratio of two variates.

Regression coefficents as a ratio of variates

In order to establish how the least-squares estimate is distributed under random samples of the dataset at hand, we first need to understand how the observations in $\boldsymbol{X}$ and $\boldsymbol{y}$ relate to give rise to this estimate. To this end, we decompose the least-squares estimate into two different components, the ratio of which is the estimate.

Corrolary 4.3. *The regression coefficient estimate $\hat{\boldsymbol{\beta}}$ is a ratio of two variates.*

Proof. Let $\mathsf{a}_i = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{y}_i$, and $\mathsf{b}_i = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{x}_i$, of dimensions $p \times 1$ and $p \times p$ respectively, where $\boldsymbol{x}_i$ denotes the ith *row* vector of $\boldsymbol{X}$. Here, all a_i and b_i are observations for two variates A and B, each observed for the ith sampling unit in the finite population of rows of $\boldsymbol{X}$ and $\boldsymbol{y}$ combined. Let $\bar{\mathsf{A}} = n^{-1}\sum_{i=1}^n \mathsf{a}_i$ and $\bar{\mathsf{B}} = n^{-1}\sum_{i=1}^n \mathsf{b}_i$ denote the population mean for A and B respectively. We can now rewrite the estimate from Definition 1.3 as follows:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y} \\
&= \left(\sum_{i=1}^n \boldsymbol{x}_i^\mathsf{T}\boldsymbol{x}_i\right)^{-1} \sum_{i=1}^n \boldsymbol{x}_i^\mathsf{T}\boldsymbol{y}_i \\
&= \left(\sum_{i=1}^n \mathsf{b}_i\right)^{-1} \sum_{i=1}^n \mathsf{a}_i \\
&= \bar{\mathsf{B}}^{-1}\bar{\mathsf{A}}
\end{aligned} \tag{4.5}$$

Hence, the regression coefficient estimate is the ratio of A to B. □

For it might not be immediately obvious that this notation does indeed represent a ratio, consider a univariate linear regression model, that is, with $p = 1$. Then $\bar{\mathsf{A}}$ and $\bar{\mathsf{B}}$ are scalar and we can write:

$$\hat{\boldsymbol{\beta}} = \frac{\sum_{i=1}^n \mathsf{a}_i}{\sum_{i=1}^n \mathsf{b}_i} = \frac{\bar{\mathsf{A}}}{\bar{\mathsf{B}}}. \tag{4.6}$$

Expected value of the sample estimate

Now that we have established that the least-squares estimate is a ratio of two variates, we can use Cochran's theory on ratio estimates (when generalized to the multivariate case) to derive further properties. For repeated sampling from an infinite population, it is known that $\hat{\boldsymbol{\beta}}$ approaches the true $\boldsymbol{\beta}$ and $\mathrm{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ [39]. In the sequel of this section, we will show that a similar result holds approximately for $\hat{\boldsymbol{\beta}}_G$.

Theorem 4.4. *For subgroup G chosen by simple random sampling,*

$$\mathrm{E}\Big[\hat{\boldsymbol{\beta}}_G\Big] \approx \hat{\boldsymbol{\beta}}. \tag{4.7}$$

Proof. Let $\bar{\mathsf{a}} = k^{-1}\sum_{i\in G}\mathsf{a}_i$ and $\bar{\mathsf{b}} = k^{-1}\sum_{i\in G}\mathsf{b}_i$ denote the sample mean of A and B respectively for simple random sample G, with $k = |G|$. Now, with proof analogous to the one in Corollary 4.3,

$$\hat{\boldsymbol{\beta}}_G = \bar{\mathsf{b}}^{-1}\bar{\mathsf{a}}. \tag{4.8}$$

We then find

$$\begin{aligned}\hat{\boldsymbol{\beta}}_G - \hat{\boldsymbol{\beta}} &= \bar{\mathsf{b}}^{-1}\bar{\mathsf{a}} - \hat{\boldsymbol{\beta}} \\ &= \bar{\mathsf{b}}^{-1}\Big(\bar{\mathsf{a}} - \bar{\mathsf{b}}\hat{\boldsymbol{\beta}}\Big).\end{aligned} \tag{4.9}$$

Note that $\bar{\mathsf{b}}$ is an unbiased estimate of $\bar{\mathsf{B}}$: if k is sufficiently large, $\bar{\mathsf{b}}$ can be considered approximately equal to $\bar{\mathsf{B}}$. For the distribution of the ratio of two random variables, $\bar{\mathsf{a}} - \bar{\mathsf{b}}\hat{\boldsymbol{\beta}}$ and $\bar{\mathsf{b}}$, is hard to establish, we approximate it by replacing the denominator of the ratio by $\bar{\mathsf{B}}$, and find

$$\hat{\boldsymbol{\beta}}_G - \hat{\boldsymbol{\beta}} \approx \bar{\mathsf{B}}^{-1}\Big(\bar{\mathsf{a}} - \bar{\mathsf{b}}\hat{\boldsymbol{\beta}}\Big). \tag{4.10}$$

This introduces a slight bias, which can however be considered negligible for large samples [9, p. 153]. For the expectation of the difference, we can move the constants $\bar{\mathsf{B}}^{-1}$ and $\hat{\boldsymbol{\beta}}$ outside of the expectation [7, p. 25], and use that $\bar{\mathsf{a}}$ and $\bar{\mathsf{b}}$ are unbiased estimators of $\bar{\mathsf{A}}$ and $\bar{\mathsf{B}}$ respectively [9, p. 22].

$$\begin{aligned}\mathrm{E}\Big[\hat{\boldsymbol{\beta}}_G - \hat{\boldsymbol{\beta}}\Big] &\approx E\Big[\bar{\mathsf{B}}^{-1}\Big(\bar{\mathsf{a}} - \bar{\mathsf{b}}\hat{\boldsymbol{\beta}}\Big)\Big] \\ &= \bar{\mathsf{B}}^{-1}\Big(\mathrm{E}[\bar{\mathsf{a}}] - \mathrm{E}[\bar{\mathsf{b}}]\hat{\boldsymbol{\beta}}\Big) \\ &= \bar{\mathsf{B}}^{-1}\Big(\bar{\mathsf{A}} - \bar{\mathsf{B}}\hat{\boldsymbol{\beta}}\Big)\end{aligned} \tag{4.11}$$

Then substituting the definition of $\hat{\boldsymbol{\beta}}$ derived in Corollary 4.3, we find

$$\begin{aligned}\mathrm{E}\Big[\hat{\boldsymbol{\beta}}_G - \hat{\boldsymbol{\beta}}\Big] &\approx \bar{\mathsf{B}}^{-1}\big(\bar{\mathsf{A}} - \bar{\mathsf{B}}\bar{\mathsf{B}}^{-1}\bar{\mathsf{A}}\big) \\ &= \bar{\mathsf{B}}^{-1}\big(\bar{\mathsf{A}} - \bar{\mathsf{A}}\big) \\ &= 0_p.\end{aligned} \tag{4.12}$$

Finally, adding the constant $\hat{\boldsymbol{\beta}}$ to both sides of the equation gives

$$\mathrm{E}\Big[\hat{\boldsymbol{\beta}}_G\Big] \approx \hat{\boldsymbol{\beta}}. \tag{4.13}$$

□

Covariance of the sample estimate

In order to derive a distribution on Cook's distance, it is imperative that we have knowledge on the covariance of the regression coefficient estimates. Again following Cochran's approach, in this section we approximate this covariance and analyze its scaling behaviour on the size of the subgroup at hand.

Theorem 4.5. *For subgroup G chosen by simple random sampling,*

$$\operatorname{Var}\left[\hat{\boldsymbol{\beta}}_G\right] \approx \frac{n^3 - kn^2}{k(n-p)} \left(\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^\mathsf{T} \operatorname{diag}(\hat{\boldsymbol{\varepsilon}})^2 \boldsymbol{X} \left(\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\right)^{-1} \tag{4.14}$$

Proof. Starting with the approximation of the difference from Equation 4.10 derived in Theorem 4.4, first observe that our sample variance is invariant under addition of a constant, say $-\hat{\boldsymbol{\beta}}$. This gives

$$\begin{aligned} \operatorname{Var}\left[\hat{\boldsymbol{\beta}}_G\right] &= \operatorname{Var}\left[\hat{\boldsymbol{\beta}}_G - \hat{\boldsymbol{\beta}}\right] \\ &\approx \operatorname{Var}\left[\bar{\mathsf{B}}^{-1}\left(\bar{\mathsf{a}} - \bar{\mathsf{b}}\hat{\boldsymbol{\beta}}\right)\right]. \end{aligned} \tag{4.15}$$

Chatfield and Colling give us the identity $\operatorname{Var}\left[\boldsymbol{P}^\mathsf{T}\boldsymbol{Q}\right] = \boldsymbol{P}^\mathsf{T} \operatorname{Var}[\boldsymbol{Q}]\boldsymbol{P}$ for constant matrix $\boldsymbol{P}$ [7, p. 25], which we put to use in moving the constant $\bar{\mathsf{B}}^{-1}$ out of the variance. For notational convenience, introduce variate $\mathsf{d}_i = \mathsf{a}_i - \mathsf{b}_i\hat{\boldsymbol{\beta}}$ with sample mean $\bar{\mathsf{d}} = \bar{\mathsf{a}} - \bar{\mathsf{b}}\hat{\boldsymbol{\beta}}$ and population mean $\bar{\mathsf{D}} = \bar{\mathsf{A}} - \bar{\mathsf{B}}\hat{\boldsymbol{\beta}} = 0_p$. From this we obtain

$$\begin{aligned} \operatorname{Var}\left[\hat{\boldsymbol{\beta}}_G\right] &\approx \left(\bar{\mathsf{B}}^{-1}\right)^\mathsf{T} \operatorname{Var}\left[\bar{\mathsf{a}} - \bar{\mathsf{b}}\hat{\boldsymbol{\beta}}\right] \bar{\mathsf{B}}^{-1} \\ &= \left(\bar{\mathsf{B}}^{-1}\right)^\mathsf{T} \operatorname{Var}\left[\bar{\mathsf{d}}\right] \bar{\mathsf{B}}^{-1}. \end{aligned} \tag{4.16}$$

Cochran shows that the sample variance of a simple random sample of given size from a finite population can be obtained by applying the finite population correction $\frac{n-k}{k}$ to the internal variance (or dispersion) of the population [9, p. 25]. This allows us to find an estimate for $\operatorname{Var}\left[\bar{\mathsf{d}}\right]$ by applying this correction to internal covariance of D for our entire dataset, which we will denote ${S_d}^2$. Note that an unbiased estimate of the internal covariance can easily and effectively be computed from the dataset, by means of summation over the squared deviation from $\bar{\mathsf{D}}$ of all d_i, with $n-p$ degrees of freedom. Substitution then gives

$$\operatorname{Var}\left[\hat{\boldsymbol{\beta}}_G\right] \approx \frac{n-k}{k} \left(\bar{\mathsf{B}}^{-1}\right)^\mathsf{T} {S_d}^2 \bar{\mathsf{B}}^{-1}. \tag{4.17}$$

Working out the definition of ${S_d}^2$, we find

$$\begin{aligned}
(n-p){S_d}^2 &= \sum_{i=1}^{n} \big(\mathsf{d}_i - \bar{\mathsf{D}}\big)\big(\mathsf{d}_i - \bar{\mathsf{D}}\big)^{\mathsf{T}} \\
&= \sum_{i=1}^{n} \mathsf{d}_i \mathsf{d}_i^{\mathsf{T}} \\
&= \sum_{i=1}^{n} \Big(\mathsf{a}_i - \mathsf{b}_i \hat{\boldsymbol{\beta}}\Big)\Big(\mathsf{a}_i - \mathsf{b}_i \hat{\boldsymbol{\beta}}\Big)^{\mathsf{T}} \\
&= \sum_{i=1}^{n} \Big(\boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{y}_i - \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{x}_i \hat{\boldsymbol{\beta}}\Big)\Big(\boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{y}_i - \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{x}_i \hat{\boldsymbol{\beta}}\Big)^{\mathsf{T}} \\
&= \sum_{i=1}^{n} \boldsymbol{x}_i^{\mathsf{T}} \Big(\boldsymbol{y}_i - \boldsymbol{x}_i \hat{\boldsymbol{\beta}}\Big)\Big(\boldsymbol{y}_i - \boldsymbol{x}_i \hat{\boldsymbol{\beta}}\Big)^{\mathsf{T}} \boldsymbol{x}_i \\
&= \sum_{i=1}^{n} \boldsymbol{x}_i^{\mathsf{T}} (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^2 \boldsymbol{x}_i \\
&= \boldsymbol{X}^{\mathsf{T}} \operatorname{diag}(\boldsymbol{y} - \hat{\boldsymbol{y}})^2 \boldsymbol{X} \\
&= \boldsymbol{X}^{\mathsf{T}} \operatorname{diag}(\hat{\boldsymbol{\varepsilon}})^2 \boldsymbol{X}.
\end{aligned} \tag{4.18}$$

Substitution of $\bar{\mathsf{B}}^{-1}$ and $\sum_{i=1}^{n} \mathsf{d}_i \mathsf{d}_i^{\mathsf{T}}$ in equation 4.17 then completes the proof.

$$\begin{aligned}
\operatorname{Var}\Big[\hat{\boldsymbol{\beta}}_G\Big] &\approx \frac{n-k}{k} \frac{n^2}{n-p} \big(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\big)^{-1} \boldsymbol{X}^{\mathsf{T}} \operatorname{diag}(\hat{\boldsymbol{\varepsilon}})^2 \boldsymbol{X} \big(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\big)^{-1} \\
&= \frac{n^3 - kn^2}{k(n-p)} \big(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\big)^{-1} \boldsymbol{X}^{\mathsf{T}} \operatorname{diag}(\hat{\boldsymbol{\varepsilon}})^2 \boldsymbol{X} \big(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\big)^{-1}
\end{aligned} \tag{4.19}$$

□

In the estimate of the internal covariance matrix, weights are assigned to the data points based on their residuals. Data points with zero residual have zero weight in this estimate. This is as expected: by least-squares theory, deletion of a data point with zero residual does not move the regression coefficient estimate, and such a data point can thus be seen as not contributing to the internal covariance.

As can be seen from Equation 4.19, the covariance of the regression estimates for a subgroup depends on the residuals of the full model — and thus on the values of $\boldsymbol{y}$, which Cook's distance does not observe — in a non-obvious way. This introduces a major difficulty for the further analysis of Cook's distance. We can however show that asymptotically for large n and under moderate assumptions, the covariance depends on $\boldsymbol{X}$ alone except for scaling.

Theorem 4.6. *For subgroup G chosen by simple random sampling, with random independent and identically distributed errors $\boldsymbol{\varepsilon}$ with $\bar{\varepsilon} = 0$ and variance σ^2, asymptotically*

for $n \to \infty$,

$$\operatorname{Var}\left[\hat{\boldsymbol{\beta}}_G\right] \approx \frac{n^3 - kn^2}{k(n-p)} \sigma^2 \left(\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\right)^{-1}. \tag{4.20}$$

Proof. Asymptotically the least-squares estimates of the regression coefficients converge to the true values [39], thus we find that under the initial assumptions, $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ is exact. It now follows from Definition 1.3 that in this case $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{\varepsilon}$.

$$\operatorname{Var}\left[\hat{\boldsymbol{\beta}}_G\right] \approx \frac{n^3 - kn^2}{k(n-p)} \left(\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^\mathsf{T} \operatorname{diag}(\boldsymbol{\varepsilon})^2 \boldsymbol{X} \left(\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\right)^{-1} \tag{4.21}$$

For errors $\boldsymbol{\varepsilon}$ are assumed to be independent and identically distributed random variables, for constant $\boldsymbol{X}$ asymptotically

$$\begin{aligned} \boldsymbol{X}^\mathsf{T} \operatorname{diag}(\boldsymbol{\varepsilon})^2 \boldsymbol{X} &= \operatorname{E}\left[\boldsymbol{X}^\mathsf{T} \operatorname{diag}(\boldsymbol{\varepsilon})^2 \boldsymbol{X}\right] \\ &= \boldsymbol{X}^\mathsf{T} \operatorname{E}\left[\operatorname{diag}(\boldsymbol{\varepsilon})^2\right] \boldsymbol{X} \\ &= \boldsymbol{X}^\mathsf{T} \left(\sigma^2 \boldsymbol{I}\right) \boldsymbol{X} \end{aligned} \tag{4.22}$$

by linearity of expectations. Substitution then gives

$$\begin{aligned} \operatorname{Var}\left[\hat{\boldsymbol{\beta}}_G\right] &\approx \frac{n^3 - kn^2}{k(n-p)} \left(\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^\mathsf{T} \left(\sigma^2 \boldsymbol{I}\right) \boldsymbol{X} \left(\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\right)^{-1} \\ &= \frac{n^3 - kn^2}{k(n-p)} \sigma^2 \left(\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\right)^{-1} \end{aligned} \tag{4.23}$$

which completes the proof. □

4.4.2 Cook's distance for simple random samples

Under the same asymptotic assumptions as in the previous discussion, we now have all we need to derive an approximate distribution for Cook's distance.

Theorem 4.7. *For subgroup G chosen by simple random sampling, with random independent and identically distributed errors $\boldsymbol{\varepsilon}$ with $\bar{\varepsilon} = 0$ and variance σ^2, asymptotically for $n \to \infty$, approximately*

$$\frac{k(n-p+1)}{n^2 - kn} \, \mathrm{D}_G \sim F(p, n-p+1). \tag{4.24}$$

Proof. Note that $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$ is a constant scalar multiple of the inverse of the variance estimate of $\hat{\boldsymbol{\beta}}_G$ from Theorem 4.6. Slightly rewriting these results, we find

$$\begin{aligned} \boldsymbol{X}^\mathsf{T}\boldsymbol{X} &\approx \frac{n^3 - kn^2}{k(n-p)} \sigma^2 \operatorname{Var}\left[\hat{\boldsymbol{\beta}}_G\right]^{-1} \\ &= \frac{n^3 - kn^2}{k(n-p)} \sigma^2 \operatorname{Var}\left[\hat{\boldsymbol{\beta}}_G - \hat{\boldsymbol{\beta}}\right]^{-1}. \end{aligned} \tag{4.25}$$

Using this approximate identity we can rewrite Cook's distance from Definition 2.1 to

$$\mathrm{D}_G \approx \frac{n^3 - kn^2}{k(n-p)}\sigma^2 \frac{(\hat{\boldsymbol{\beta}}_G - \hat{\boldsymbol{\beta}})^\mathsf{T} \operatorname{Var}\left[\hat{\boldsymbol{\beta}}_G - \hat{\boldsymbol{\beta}}\right]^{-1} (\hat{\boldsymbol{\beta}}_G - \hat{\boldsymbol{\beta}})}{p\hat{\sigma}^2}. \tag{4.26}$$

For asymptotically $n\sigma^2 = (n-p)\hat{\sigma}^2$ is exact, we may simplify to

$$\mathrm{D}_G \approx \frac{n^2 - kn}{pk}(\hat{\boldsymbol{\beta}}_G - \hat{\boldsymbol{\beta}})^\mathsf{T} \operatorname{Var}\left[\hat{\boldsymbol{\beta}}_G - \hat{\boldsymbol{\beta}}\right]^{-1} (\hat{\boldsymbol{\beta}}_G - \hat{\boldsymbol{\beta}}). \tag{4.27}$$

Cochran argues that for the univariate case, $\hat{\boldsymbol{\beta}}_G - \hat{\boldsymbol{\beta}}$ is normally correlated [9, p. 153], Gray gives the same result for the multivariate case in the appendix of [26]. We find $\hat{\boldsymbol{\beta}}_G - \hat{\boldsymbol{\beta}}$ to be p-variate normally correlated with mean 0_p (from Theorem 4.4) and estimated variance $\operatorname{Var}\left[\hat{\boldsymbol{\beta}}_G - \hat{\boldsymbol{\beta}}\right]$.

The right-hand side of Equation 4.27 now has the form of a scaled T^2-statistic as introduced by Hotelling in [29]. As such, we find Cook's distance to be a scaled Hotelling's T^2 statistic:

$$\mathrm{D}_G \approx \frac{n^2 - kn}{pk}\frac{1}{n}T^2. \tag{4.28}$$

Rao further relates Hotelling's T^2 to the F-distribution in [49, p. 458] as

$$\frac{n-p+1}{p}\frac{T^2}{n} \sim F(p, n-p+1). \tag{4.29}$$

From here, substitution of our former result for T^2 leaves us with

$$\frac{k(n-p+1)}{n^2-kn}\,\mathrm{D}_G \sim F(p, n-p+1) \tag{4.30}$$

approximately, which completes the proof. □

4.4.3 A modified Cook's distance

In our previous theorem we have shown that under asymptotic assumptions, we can derive a distribution of Cook's distance. For datasets of limited size however, we cannot be sure that these assumptions hold. Least-squares regression coefficient estimates have been shown to converge in mean square and almost surely under moderate assumptions, but the analysis of the rate of this convergence is complex [39].

Note that the asymptotic assumptions in our proof for Cook's distance are only required for us to be able to find an estimate on the design matrix employed. As such, we may construct a modified Cook's distance based on the generalized Cook's distance, for which we do not have to assume convergence and may still derive a distribution.

Theorem 4.8. *For subgroup G chosen by simple random sampling, approximately*

$$\frac{(n-p+1)(n-p)k}{n^2p(n-k)}\,\mathrm{D}_G\Big(\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\big(\boldsymbol{X}^\mathsf{T}\operatorname{diag}(\hat{\boldsymbol{\varepsilon}})^2\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\Big) \sim F(p, n-p+1). \tag{4.31}$$

Proof. This proof is analogous to the proof for Theorem 4.7, however using the covariance estimate without asymptotic assumptions from Equation 4.19. Observe that the right-hand side of the equation

$$\mathrm{D}_G\left(c\operatorname{Var}\left[\hat{\boldsymbol{\beta}}_G-\hat{\boldsymbol{\beta}}\right]^{-1}\right) = c(\hat{\boldsymbol{\beta}}_G-\hat{\boldsymbol{\beta}})^\mathsf{T}\operatorname{Var}\left[\hat{\boldsymbol{\beta}}_G-\hat{\boldsymbol{\beta}}\right]^{-1}(\hat{\boldsymbol{\beta}}_G-\hat{\boldsymbol{\beta}}) \tag{4.32}$$

has the form of a scaled Hotelling's T^2 statistic:

$$\frac{n}{c}\,\mathrm{D}_G\left(c\operatorname{Var}\left[\hat{\boldsymbol{\beta}}_G-\hat{\boldsymbol{\beta}}\right]^{-1}\right) = T^2 \tag{4.33}$$

We can then reuse the relation Rao gives for Hotelling's T^2 and the F-distribution from Equation 4.29, giving

$$\frac{n-p+1}{cp}\,\mathrm{D}_G\left(c\operatorname{Var}\left[\hat{\boldsymbol{\beta}}_G-\hat{\boldsymbol{\beta}}\right]^{-1}\right) \sim F(p, n-p+1). \tag{4.34}$$

Now substituting c by $\frac{n^3-kn^2}{k(n-p)}$ and $\operatorname{Var}\left[\hat{\boldsymbol{\beta}}_G-\hat{\boldsymbol{\beta}}\right]$ by its approximation from Equation 4.19, we find after simplification that approximately

$$\frac{(n-p+1)(n-p)k}{n^2p(n-k)}\,\mathrm{D}_G\Big(\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\big(\boldsymbol{X}^\mathsf{T}\operatorname{diag}(\hat{\boldsymbol{\varepsilon}})^2\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\Big) \sim F(p, n-p+1). \tag{4.35}$$

□

Theorem 4.8 describes an internal norm influence measure based on the generalized Cook's distance from Definition 4.1 with a known statistical distribution. From this result, we derive our modified Cook's distance.

Definition 4.9 (Modified Cook's distance). Let our modified Cook's distance be defined as

$$\mathrm{D}^*_G = \mathrm{D}_G\Big(\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\big(\boldsymbol{X}^\mathsf{T}\operatorname{diag}(\hat{\boldsymbol{\varepsilon}})^2\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\Big) \tag{4.36}$$

and let its F-measure be

$$\frac{(n-p+1)(n-p)k}{n^2p(n-k)}\,\mathrm{D}^*_G, \tag{4.37}$$

which is distributed according to the F-distribution with p and $n-p+1$ degrees of freedom for random samples, as shown in Theorem 4.8.

Our modified Cook's distance is a true internal norm influence measure: the deviation of the regression coefficient estimates is measured with respect to the internal covariance of the full dataset, instead of the expected sample covariance for repeated sampling

from an infinite population. Asymptotically however, our modified Cook's distance can be shown to be equivalent to the original except for a constant scaling factor, for they asymptotically have identical design matrices, which follows immediately from Equation 4.22.

As our modified Cook's distance measure does not depend on the asymptotic properties of the dataset at hand, we consider this measure more suitable for Exceptional Model Mining than Cook's distance in its original form, for reasons outlined in Section 4.3.

4.5 Empirical verification

Figures 4.1 and 4.2 show the distributions of Cook's distance and our modified Cook's distance, for the Ames Housing and Windsor Housing dataset respectively, both superimposed with the prediction arising from the approximate distribution derived in Theorem 4.7 for Cook's distance and Theorem 4.8 for our modified Cook's distance.

For Cook's distance in its original form on Ames Housing, it is clear that our prediction is way off. The ratio of the measured distance distribution to the predicted distance distribution is approximately 9, exhibiting primarily as an offset error (as well as a smaller scaling error) on the logarithmic scale. Note that the scaling behavior of Cook's distance is still predicted quite well. Further analysis points out that this is caused by the internal covariance matrix differing substantially from the asymptotic estimate, both in magnitude and sign. For Windsor Housing however, these matrices line up substantially better, and the results improve equivalently.

The distribution derived for our modified Cook's distance predicts very well on the troublesome Ames Housing dataset, and even better on the Windsor Housing dataset. The small deviation in mean for small subsets on Windsor Housing can be attributed to the approximation introduced in Theorem 4.4, but is deemed negligible for larger subgroups. Note that in practice, following from the requirements identified by Klösgen, Exceptional Model Mining will not be used for finding small subgroups, as such subgroups will provide almost no insight in the dataset at hand due to their small relative size. As such, this approximation is unlikely to be problematic in practice.

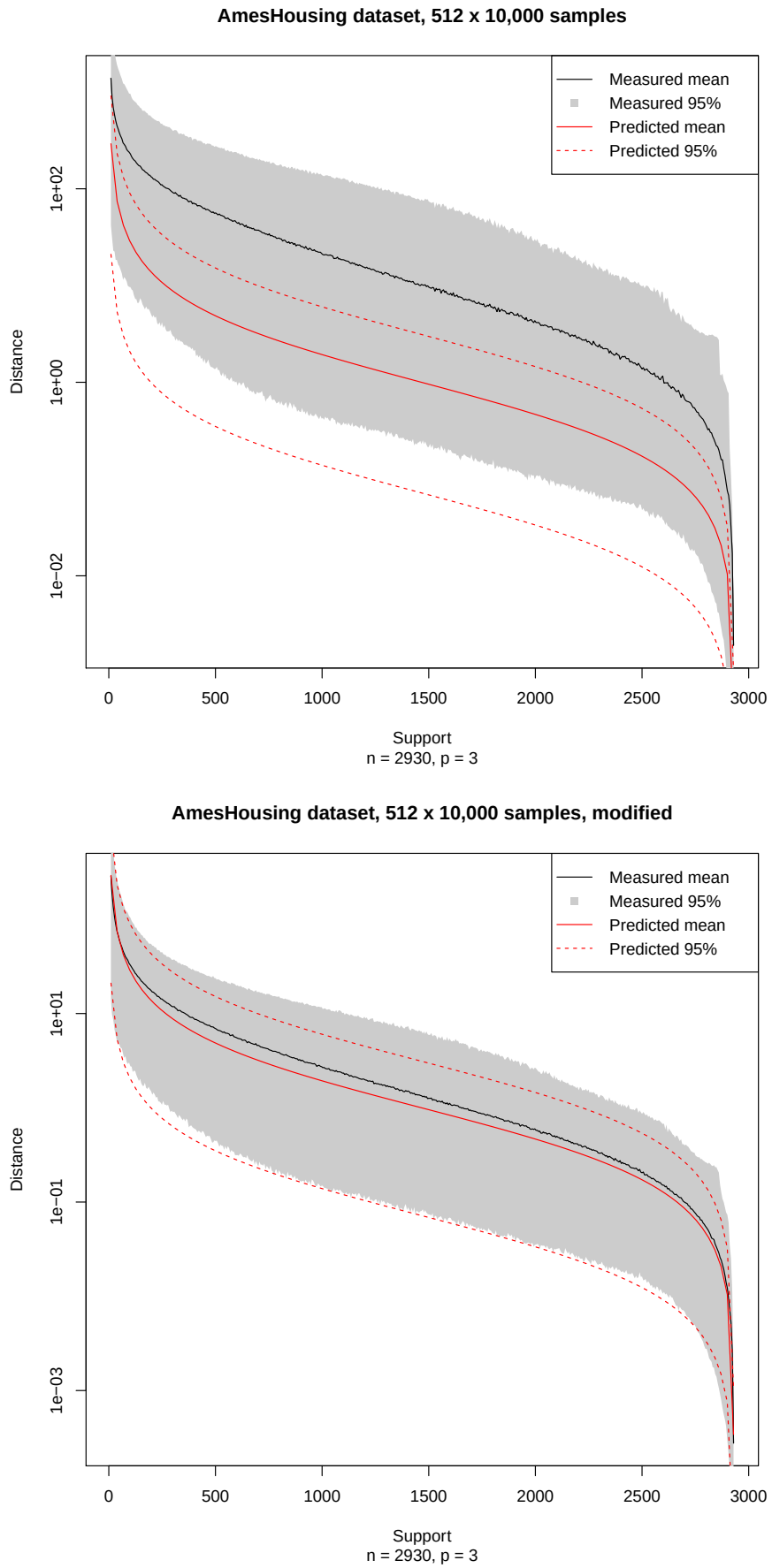

Figure 4.1: The distribution of Cook's distance D_G (top) and our modified Cook's distance D_G^* (bottom) for randomly chosen subgroups from the Ames Housing dataset. Sampled for 512 different subgroup sizes (supports), based on 10,000 samples per subgroup size. The black line and gray area indicate the measured mean and 95% confidence interval, where the red lines indicate the predicted equivalents following Theorem 4.7 and 4.8. Results are presented on a logarithmic vertical scale for legibility.

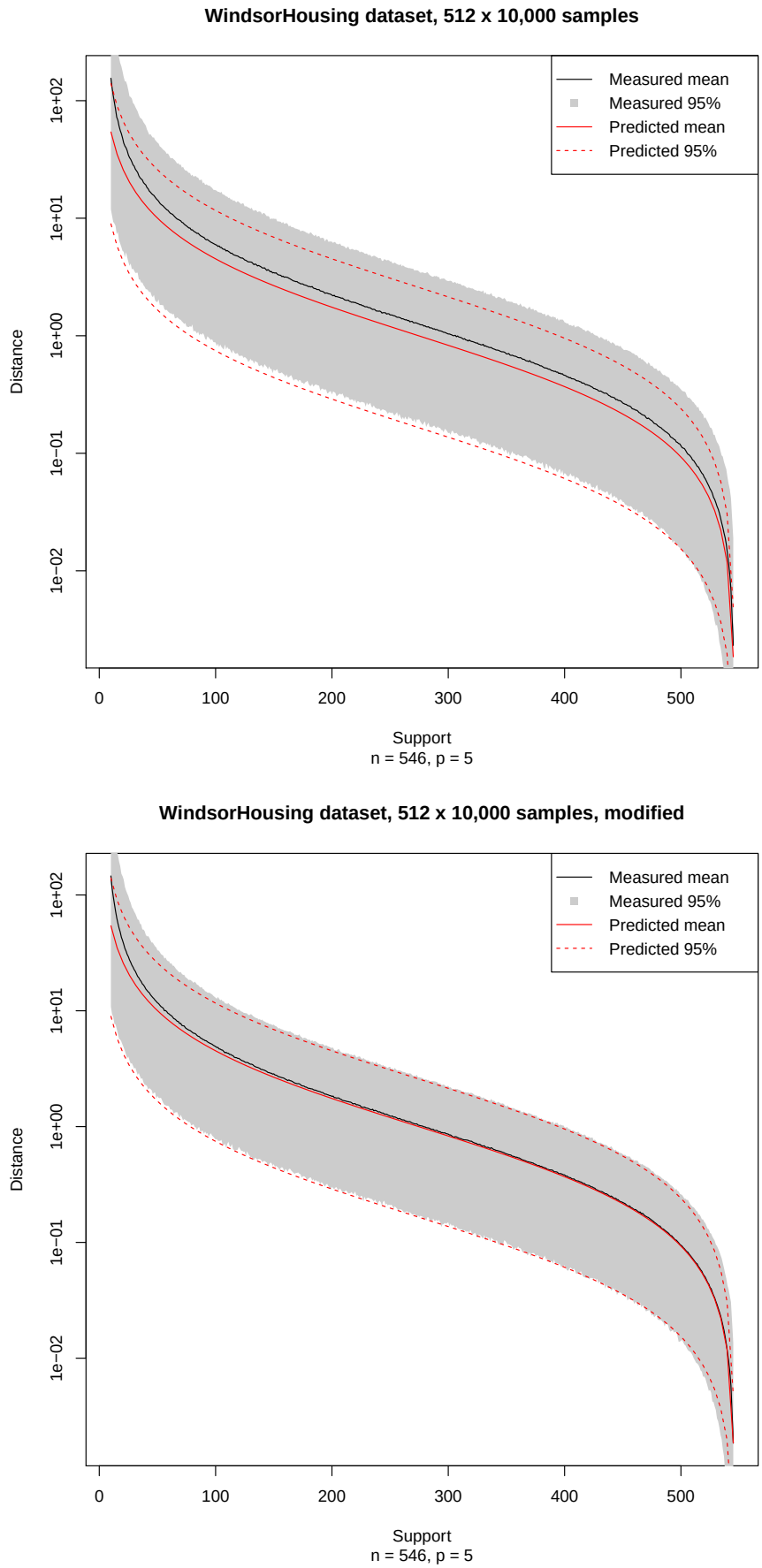

Figure 4.2: The distribution of Cook's distance D_G (top) and our modified Cook's distance D_G^* (bottom) for randomly chosen subgroups from the Windsor Housing dataset. Sampled for 512 different subgroup sizes (supports), based on 10,000 samples per subgroup size. The black line and gray area indicate the measured mean and 95% confidence interval, where the red lines indicate the predicted equivalents following Theorem 4.7 and 4.8. Results are presented on a logarithmic vertical scale for legibility.

4.6 Concluding remarks

The expected distance of the model parameters for a subgroup, compared to those for the full dataset, strongly depends on the subgroup's support. Near-complete subgroups will usually not differ much from the full dataset, while it is reasonable to expect the existence of extremely small subgroups that differ substantially. In the trade-off between evidence and generality, the analyst may consider both to be interesting as long as a subgroup's evidence is substantially greater than expected given its support. To establish this interestingness, proper knowledge on the distribution of the evidence with respect to support is required. This does not only involve the scaling behaviour in mean value, but also the scaling behaviour in variance and skewness of the distribution, albeit to a lesser extent.

For the original Cook's distance, only non-tabulated generalized distributions or distributions subject to strong assumptions have been published in the literature. Theorem 4.7 shows that we can only derive a distribution on Cook's distance D_G when certain asymptotic properties hold on the dataset and model under consideration, which can be observed in Figure 4.1 to be quite off. For our modified Cook's distance D_G^* from Definition 4.9 however, a reasonable approximation can be derived as has been shown in Theorem 4.8. This knowledge allows the data miner to judge a subgroup's evidence of exceptionality in a statistically sound way, regardless of the subgroup's support, as the F-measure for D_G^* from Definition 4.9 can be compared against a known F-distribution. As such, the use of our modified Cook's distance is more suitable for Exceptional Model Mining and other data mining tasks than Cook's distance in its original form.

Chapter 5

Mining for knowledge in multiple objectives

In this chapter, we further investigate the objectives relevant to Exceptional Model Mining, and the potential merits of a multi-objective approach to this data mining problem. Traditional search strategies for knowledge discovery tend to optimize only a single objective. Depending on how the resulting solutions are chosen, the analyst may end up with only a single solution or a set of usually highly similar solutions, leaving little to no choice at all. This traditional approach ignores the fact that interestingness of a solution to the analyst cannot be based on a single objective function, for many of the analyst's interests are inherently conflicting.

The goal of this chapter is to establish a roadmap for subsequent experimentation on Exceptional Model Mining, which will be discussed in Chapter 6. First, an introduction to data mining in a multi-objective setting is given and relevant terminology is established. Previous multi-objective algorithmic results on problems closely related to Exceptional Model Mining are discussed subsequently. Finally, general strategies for dealing with multi-objective problems from related work are discussed.

5.1 On interesting subgroup objectives

According to Klösgen's criteria, as discussed in-depth in Section 1.1, subgroup concepts must be interpretable, application relevant, and "interesting" in a statistically significant way [35]. Interpretability and application relevance can be enforced by choice of the concept space. Subsequently, all solutions to be found by an automated discovery strategy are to be judged on their interestingness to the analyst. Klösgen identifies several

criteria of interestingness [36], of which only the following can be determined without further user interaction:

Evidence The significance of the peculiarities in the consequent.

Simplicity The syntactical complexity of a pattern.

Generality The fraction of all observations represented by the concept.

Klösgen concentrates primarily on evidence and generality, and leaves it to the analyst to implement simplicity constraints in the chosen concept language.

Klösgen argues in [36] that exhaustive search of a large concept space without strong user interaction for intermediate introduction of search bias is infeasible. Additionally, Klösgen describes knowledge discovery as "an interactive process which depends on the goals of the analyst", and argues that no automated system can replace an analyst completely. This forms a direct case against methods that attempt to return a single best result, or a set of results that has little internal variance: the analyst must be given a choice.

Generally applicable heuristic strategies such as beam search can be considered, not requiring domain knowledge for their workings, to alleviate the need for user interaction. However, single-objective methods usually converge to sets of similar solutions that are most interesting on that specific choice of objective, but may not capture the real interest of the analyst, let alone provide choice. For this exact reason, traditional single-objective algorithms have the tendency to excel on certain datasets, yet completely fail on others. García et al. identify this as a consequence of the 'no free lunch' theorem: one single algorithm cannot have the best behaviour for all problems [25].

While free lunch cannot be had, instead of compromising interestingness to the analyst, a knowledge discovery strategy can at least make sure that the analyst has plenty of choice. The general solution is to always include one of the best attainable results in the trade-off between the analyst's conflicting interests. This is where a multi-objective Pareto-based approach may come in. Multi-objective approaches in the Pareto sense allow to find a set of best trade-offs, i.e. a highly varied set of results, allowing the analyst a post-hoc choice.

Novak et al. provide an analysis of various algorithms for strategies in the Supervised Descriptive Rule Discovery framework in [46] — this framework unifies various automated knowledge discovery strategies such as Contrast Set Mining, Emerging Pattern Mining and Subgroup Discovery. As a generalization of Subgroup Discovery, Exceptional Model Mining also fits in this framework. Novak et al. conclude that all algorithms they considered "aim at optimizing a trade off between rule coverage and precision"[1] [46].

[1] In the terminology used in this work, 'coverage' is denoted as generality and 'precision' as evidence.

This emphasizes the multi-objective nature of knowledge discovery — and thereby Exceptional Model Mining — as earlier identified by Klösgen [36]. However, none of the considered algorithms truly exploit this characteristic by taking a multi-objective approach. In the following sections we will discuss existing algorithms that are claimed to feature a multi-objective approach.

5.2 Multi-objective Subgroup Discovery

Several attempts at leveraging multidimensional objectives have been made for the Subgroup Discovery knowledge discovery task. For Exceptional Model Mining however, we are not aware of any such attempts in the literature. With Subgroup Discovery being closely related to Exceptional Model Mining, the latter being a generalization of the former, results on Subgroup Discovery are likely to be somewhat applicable to Exceptional Model Mining. As such, we proceed by discussing these previous results from the literature, while taking into consideration their applicability to Exceptional Model Mining.

5.2.1 Linear-weighted "multi-objective" Subgroup Discovery

Del Jesus et al. present SDIGA, a novel approach to Subgroup Discovery based on an evolutionary strategy [33]. To overcome limits imposed by discretization of the data — Del Jesus et al. note this to be often required for reasons of search space reduction and interpretability of the resulting rules — they represent subgroups as fuzzy rules. These fuzzy rules are concepts in disjunctive normal form, where all numerical attributes are replaced by overlapping linguistic labels from "low" to "high", avoiding what they describe as unnatural boundaries. Del Jesus et al. claim that this leads to simple rules that are highly actionable, and thus of high interest to the analyst.

Del Jesus et al. introduce a hybrid evolutionary strategy to finding simple and general concepts with significant evidence. Their approach is based on a combination of local search, where newly found solutions are improved by means of hill-climbing, and evolution, where promising solutions are mutated and recombined to introduce variation and find entirely new solutions. Concepts are encoded as binary strings, where groups of consecutive bits represent the inclusion (or exclusion) of linguistic terms in the concept, each group encoding for one independent variable. Two-point crossover is used as a recombination operator. Two mutation strategies are employed, namely either eliminating all constraints on one variable in the concept, or assigning it a random choice of linguistic terms.

Fitness of solutions is evaluated as a linear combination of generality and evidence, where the former is measured as relative support: the fraction of cases represented by

the concept, not represented by any of the existing rules. The idea behind this approach to generality is that this leads to a good covering of the dataset by all rules in the population. One might however argue that, as already pointed out by Klösgen, the analyst is not interested in a complete description of the dataset, but in the nuggets contained therein instead. Also, as the multiple objectives are considered in a linearly weighted fashion, the algorithm is strongly biased in the direction of the trade-off line represented by the weights. It is therefore doubtful that this approach is helpful to the analyst, or that it will provide him the desired choice.

Pachón et al. take another evolutionary approach to Subgroup Discovery, where both categorical and continuous attributes in the concept space can be used without further prior discretization [47]. While the authors repeatedly label their method as "multi-objective", in reality, only a single objective that is a linear combination of several objectives is optimized. Hence, their algorithm is again not a *true* multi-objective evolutionary algorithm in the Pareto sense, and may not always give the analyst an appropriate choice of concepts.

The genetic encoding employed by Pachón et al., however, allows for a lack of discretization. Each gene, of which there are a fixed number, encodes for either the minimum and maximum value of the interval for an ordered variable, or the set of possible values for an unordered variable. By fixing the number of genes, a lower bound is introduced on the simplicity of the concepts considered.

The genetic operators used are uniform crossover (by which both possible candidates are evaluated, and the best chosen), and point mutation where either the upper and lower bounds of an interval are modified or values are added or removed from the set of possible values for that variable. The intervals are always reduced to values that belong to the database.

As mentioned before, the fitness measure employed is a linear combination of several measures, including significance of evidence and generality. In addition, measures for confidence, negative terms for size of the intervals, and a negative term for covering already covered observations are employed.

Here, the measure for *confidence* in Subgroup Discovery — representing the fraction of observations matched by the antecedent of the pattern with the target attribute equal to the consequent rule — sparks particular interest, as it determines a goodness-of-fit for which the R^2 measure can be substituted in linear regression. Is is worth noting that this objective is missing in the domain analysis of Klösgen [36], as it should be obvious to the reader that high-confidence patterns should be preferable to the analyst. As such, we may extend the objectives in the overview from Section 5.1 as follows:

Confidence The accuracy of the consequent for all observations represented by the concept.

Pachón et al. report that their method outperforms existing methods for Subgroup Discovery. We must however note that their algorithm consistently finds either more significant results (at some times combined with high confidence), or more general results, but never both for the same dataset. This again depends on the dataset at hand, and as such highlights the potential merits of a true multi-objective approach.

5.2.2 Pareto multi-objective Subgroup Discovery

González et al. later extend the SDIGA algorithm to a true multi-objective approach in NMEEF-SD [6]. In their report, they recognize that no single quality measure may be sufficient for Subgroup Discovery, and that the best trade-off between the conflicting quality measures can be obtained by means of a true multi-objective evolutionary algorithm. The aim of their approach is to improve diversity in the population, which we have earlier considered to be invaluable to the analyst. González et al. realize this by using the NSGA-II elitist evolutionary model, the aim of which is to find a broad approximation of the entire Pareto front by partitioning the objective space into niches and adopting selection strategies that take into account the crowdedness of a niche. The reader is referred to [16] for further details on NSGA-II.

As is the case for SDIGA, the concept space in NMEEF-SD is also based on fuzzy linguistic rules, and uses the same genetic representation for concepts, except for employing integers instead of bit sequences to represent variables. The same genetic operators as in SDIGA are employed. In addition to the NSGA-II scheme, a re-initialization step is used when the Pareto front estimate is detected to have stopped evolving. The objective functions under evaluation again include evidence, generality and confidence. González et al. report that their algorithm outperforms all existing algorithms they evaluated.

5.3 Multi-objective search for Exceptional Models

Drugan and Thierens describe a generally applicable approach to multi-objective optimization in [19]. Their approach is based on a Pareto local search mechanism: the exploration of the neighbourhood of non-dominated solutions in a population of solutions by means of local search. To avoid exploration of already explored regions, a deactivation mechanism is introduced. Drugan and Thierens describe a variety of neighbourhood exploration strategies or improvement strategies, and provide stochastic perturbation mechanisms to introduce new solutions when the entire neighbourhood is explored. These perturbation strategies include a multi-restart strategy based on random reinitialization, and a genetic strategy that is biased towards preserving common structures in the set of non-dominated solutions.

It must be noted that except for a domain-specific neighbourhood definition, Drugan and Thierens' Pareto Local Search strategies do not rely on domain knowledge and can as such be applied to any multi-objective problem. As such, for a proper definition of neighbourhood and encoding of solutions, these Pareto Local Search strategies can be expected to be applicable to Exceptional Model Mining as well. As previously discussed in Section 5.2.1, Pachón et al. have shown that concept encoding based on limits for continuous variables is useful in Subgroup Discovery. Since concept encoding does not necessarily differ for other tasks in supervised descriptive rule discovery, this approach is promising for all other problems in this class as well, including Exceptional Model Mining. While the strategy by Pachón et al. is not a true multi-objective approach, this does not constrain further usability of their concept encoding. When encoding concepts in knowledge discovery as intervals, we observe that care must be taken to properly handle the dependencies between the minimum and the maximum of the interval. One potential solution to drop the dependence between the two values, is to encode an interval center and an interval width, both of which can be adjusted independently without a strong bias towards infeasible solutions.

Chen et al. propose a new hybrid local search-based optimization algorithm for multi-objective problems on continuous decision spaces [8] names NSLS. Similar to evolutionary strategies, a generational approach is taken, and existing solutions are used to generate offspring. NSLS is not entirely an evolutionary strategy however, for it does not feature recombination. Rather, it uses Gaussian perturbation where the severity of perturbation on a variable is based on its variance in the population. NSLS is structurally similar to NSGA-II. Its main contribution however, is that it replaces the spread-biased selection mechanism based on crowding distance in NSGA-II, by selection of farthest candidates. Chen et al. claim superiority of their approach to many existing algorithms on several synthetic test problems. While promising, its potential for use on Exceptional Model Mining may be limited unless it is assumed that datasets feature only continuous (or otherwise ordinal) attributes, or a different perturbation strategy is devised that allows for nominal attributes in the decision space.

Chapter 6

Experimental setup

In the previous chapter, an outline is given for experimentation on Exceptional Model Mining in a multi-objective setting. This chapter provides the design of an experiment to evaluate the effectiveness of a multi-objective local search approach to exceptional model mining, compared to existing single-objective techniques.

6.1 Benchmark datasets

The algorithms evaluated in this experiment are judged on their relative performance on a selection of datasets and models describing this data. To account for possible performance bias on number of records in the dataset, dataset dimensionality, model complexity, underlying "true" mechanism of subgroup appearance and such, care is taken to compile a set of varied datasets with respect to these properties. No further preferences are used in the compilation of this set, as for the purposes of this experiment, the semantics of the found subgroups themselves are not of particular interest. As such, the selection of datasets on other properties is done rather arbitrarily, and primarily based on the choice of datasets in [38]. Again, we must stress that the exact choice of datasets is not important for this experiment, as long as they are substantially different in general. Models are selected as to provide a reasonable global fit, wherever possible. Care is taken to not introduce overly complex models, as that would reduce the number of attributes available for the description of concepts, and decrease the descriptive quality and expected maximum evidence of subgroups.

The benchmark datasets chosen are *Ames Housing*, *Windsor Housing*, *Wine*, *Student Performance* (both the 'mathematics' and the 'Portuguese' class, further abbreviated as 'mat' and 'por' respectively) and *Census 90*. Tables 6.1 and 6.2 list the benchmark models used on these datasets and the expected mechanisms behind subgroups in the

datasets, while Table 6.3 lists some descriptive statistics on these datasets and their models. This section proceeds with a brief introduction and some justification of our choice of datasets.

Ames Housing and Windsor Housing

Ames Housing [15] and Windsor Housing [1] are datasets of similar nature, both containing pricing and additional information from house sales in the cities of Ames and Windsor respectively. As such, it can be expected that both datasets may show subgroups as a result of hedonistic pricing as described by Anglin and Ramazan in [1] for Windsor Housing. Ames Housing, however, is of far greater dimensionality than Windsor Housing — both in number of records and in number of attributes. For this reason, both datasets are included in the benchmark set. Their benchmark models then predict the sale price of houses based on a small selections of independent attributes that appear to provide a reasonable fit for either dataset (see Table 6.3), corresponding to the selection chosen by Krak and Feelders in [38] for these datasets.

Wine

The Wine dataset [14] again measures pricing — of wines instead of housing — influenced by hedonism, as reported by Constanigro et al. in [14]. This dataset is remarkable for having a vast number of binary attributes, yet only a single non-binary attribute (except for those used in the regression model). Again, a simple model is chosen, corresponding to that in [38].

Student Performance

One pair of datasets — the Student Performance datasets [13, 41] — has not been used as a benchmark in [38]. These datasets describe various properties of secondary school students following courses in Portuguese or mathematics (or both). For these courses, students are assigned grades for three different exams at different times during the course. These exam grades are also included in the dataset. This pair gives a wonderful possiblity to benchmark the algorithms using a simple model with great goodness-of-fit, by the expected correlation between the test results, hence predicting the final exam grade based on the previously obtained grades. This good fit is demonstrated by the respective R^2 values in Table 6.3.

Census 90

Krak and Feelders also include the *Adult* dataset in their benchmarks [38], which consists of census data on the income of various individuals in the United States of America [41]. This dataset however is ill-suited for linear regression, as the target variable (income) is discretized into less than and more than $50,000, making it more of a dataset suitable for classification.

Still, census datasets are a viable candidate for benchmarking. Census datasets are expected to be interesting due to the expectation of a discrimination subgroup mechanism (such as discrimination on gender [5] and race [28]) yielding high-quality subgroups. Also, the high quantity of data usually available from the census gives the opportunity to benchmark on a larger dataset.

For the original data that this dataset is based on, that is, prior to discretization of income, is to our best knowledge unavailable, we introduce a new dataset, Census 90. Our Census 90 dataset is derived from a 20,000-household sample of the 1990 1% unweighted IPUMS census data [50]. Data is preprocessed similarly to the "Adult" dataset [41], however without discretization of income. As such, only representative employed persons are included: these are the cases with age, total personal income, person inclusion weight and usual hours of work per week satisfying

$$(AGE > 16) \wedge (INCTOT > 100) \wedge (PERWT > 1) \wedge (UHRSWORK > 0).$$

Total income $INCTOT$ is then scaled logarithmically to increase spread. An attempt has been made to include most of the other variables provided in the Adult dataset, where they were available in the IPUMS data. Descriptive properties of this dataset under our simple benchmark model (see Table 6.1) are listed in Table 6.3.

6.2 Comparison and analysis

For every dataset, each algorithm under comparison is evaluated r times until the stop criterion is reached. For each such algorithm evaluation, for the resulting rules and their respective subgroups the following criteria are measured[1]:

Evidence (unusualness) by means of the F-measure of our modified Cook's distance D^*_G from Definition 4.9, or 0 for rules matching the entire dataset.

Generality (support) by means of the fraction $\frac{k}{n}$.

Confidence (goodness-of-fit) by means of R^2.

Here, evidence and generality are taken from Klösgen's framework as discussed more in-depth in Section 5.1, and confidence as the trivial analog to the confidence measure in Subgroup Discovery as discussed in Section 5.2.1. The simplicity measure from Klösgen's framework is not included, as this would require all algorithms under comparison to use concept representations of comparable expressivity. This would impose undue limits on the choice of algorithms under comparison.

[1] Note that the terminology used for these objective functions in the literature differs between authors, in the sequel of this chapter, the choice of terminology typeset in bold will be used.

Dataset	Model
Ames Housing	$SalePrice \sim \beta_0 + \beta_1 \cdot LotArea + \beta_2 \cdot OverallQual$
Windsor Housing	$sell \sim \beta_0 + \beta_1 \cdot lot + \beta_2 \cdot bdms + \beta_3 \cdot fb + \beta_4 \cdot sty$
Wine	$Price \sim \beta_0 + \beta_1 \cdot Cases + \beta_2 \cdot Score + \beta_3 \cdot Age$
Student Performance (por)	$G3 \sim \beta_0 + \beta_1 \cdot G1 + \beta_2 \cdot G2$
Student Performance (mat)	$G3 \sim \beta_0 + \beta_1 \cdot G1 + \beta_2 \cdot G2$
Census 90	$\log(INCTOT) \sim \beta_0 + \beta_1 \cdot AGE + \beta_2 \cdot UHRSWORK + \beta_3 \cdot EDUC$

Table 6.1: Overview of benchmark datasets and models.

Dataset	Prediction	Subgroup mechanism
Ames Housing	house pricing	hedonism[a]
Windsor Housing	house pricing	hedonism [1]
Wine	wine pricing	hedonism [14]
Student Performance (por)	student grades	*unknown*
Student Performance (mat)	student grades	*unknown*
Census 90	income	discrimination [5, 28][b]

[a]: under the assumption that the results in [1] apply to house prices in Ames as well
[b]: mechanisms are known for income in general, but not reported for this dataset in specific

Table 6.2: Overview of prediction targets and subgroup mechanisms known from the literature for the benchmark datasets.

Dataset	R^2	n	p	Attr. (bin, nom, ord)
Ames Housing	0.675	2930	3	77 (2, 21, 54)
Windsor Housing	0.536	546	5	7 (6, 0, 1)
Wine	0.313	9600	4	26 (25, 0, 1)
Student Performance (por)	0.848	649	3	30 (13, 4, 13)
Student Performance (mat)	0.822	395	3	30 (13, 4, 13)
Census 90	0.370	26623	4	13 (2, 10, 1)

Table 6.3: Descriptive statistics for the datasets used under the models outlined in Table 6.1. Shown are global goodness-of-fit measured by R^2, number of records n, number of concept attributes and their types (binary, other nominal/categorical, ordered/numeric), and number of regression coefficients p.

To compare different runs of algorithms within a dataset, the dominated hypervolume (or Lebesgue measure of the non-dominated front) of each non-dominated result set is used as a scalar performance measure. The hypervolume not only increases when single objectives are improved, but also when the solutions are spread more widely. This makes the hypervolume an indicator for the potential of the entire solution set in terms of interestingness to the analyst, making attained hypervolume a suitable measure for the performance of the algorithms under comparison.

For the criteria used have ranges of $[0, \infty)^2$, $(0, 1]$, and $(0, 1]$ respectively, the dominated hypervolume is calculated with respect to the reference vector $(0, 0, 0)$ and the axes of the identity basis. Note that this measure is ill-suited to compare results from different datasets, for no knowledge is assumed on the actual Pareto front of each dataset. Accordingly, the reference vector is chosen pessimistically. Hypervolume is calculated in the R programming environment [48] by means of the `mco` package version 1.0-15.1 [44], implementing the efficient hypervolume indicator algorithm presented by Fonseca et al. in [24].

For each dataset, the results for each algorithm are compared following the routine outlined by Bader in [2]. Here it is established whether there is a significant difference between algorithms by means of a Kruskal–Wallis rank sum test with $\alpha = 0.05$, and if so, the Conover–Inman post-hoc procedure is applied at the same confidence level. The former test establishes whether there is a pair of algorithms that performs significantly different, and the latter procedure determines for each pair of algorithms whether their performance difference is significant. Here, the null hypothesis is that all algorithms perform equally well, that is, the choice of algorithm does not significantly affect the hypervolume obtained.

Apart from results on the final resulting sets, we are also interested in establishing an overview of the speed of convergence of our algorithms under comparison. To this end, not only the hypervolume for each resulting set is measured, but also intermediate sets are measured intermittently before the stop criterion is reached.

6.3 Algorithms and variants

The performance of different variants of multi-objective algorithms is compared to two existing single-objective approaches.

[2]Note that the range of the F-measure for our modified Cook's Distance is limited by the dataset at hand. One can establish an upper bound on the attainable distance by enumerating all subgroups consisting of p records, and finding the maximum of their respective distances, as the regression estimates are always a linear combination of the estimates for these subgroups [26]. This is however irrelevant to our calculations, so infinity is assumed instead.

6.3.1 Reference single-objective algorithms

Two algorithms are included in this experiment as single-objective references. The generically applicable Beam Search strategy is chosen for its simplicity and for not requiring use of any domain knowledge at all. In addition, the Tree-Constrained Gradient Ascent algorithm, a state of the art algorithm that is reported to outperform Beam Search for a wide variety of datasets [38], is compared against.

The sequel of this section primarily discusses their implementation details. For a more general overview of these algorithms the reader is referred to Section 2.2.

Beam Search

Usually Beam Search is fixed to a certain search depth, but for rules are constrained further at each search level — thereby reducing generality — solutions encountered before the final level must also be considered. Therefore a slightly modified Beam Search algorithm is used. This algorithm will be denoted Archiving Beam Search.

Where Beam Search only returns the final beam, with the top solutions found at the final search level, Archiving Beam Search keeps the beam found at the end of each search level in an archive, including the final beam, and yields the union of this archive. Pseudocode for Archiving Beam Search is given in Algorithm 1. Arguments are the attributes $\mathcal{A}$ to refine on, the refinement operator $\mathcal{N}$ that refines a rule on a given attribute, a quality function φ, search depth d and beam width w.

Note that in implementation, computational complexity can be reduced by keeping beam's rules sorted by quality.

For this experiment our own implementation of this algorithm in the R programming language is used. Let the refinement operator $\mathcal{N}(a, r)$ be defined such that it returns the complete set of rules adding a single conjunctive clause to r. To this end, clauses c^+ and c^- are generated for each value v in $\pi_a(\sigma_r(\mathcal{D}))$, where σ_r denotes the selection of observations matching r, and π_a the projection on a. For ordinal attributes, these clauses are '$> v$' and '$< v$' respectively; for unordered attributes, these clauses are '$= v$' and '$\neq v$'. The union of the set of all $r \wedge c^+$ and $r \wedge c^-$ is then returned as the set of refinements.

The quality measure $\varphi(r)$ is chosen to be our modified Cook's distance for the subgroup of $\mathcal{D}$ by selection on r.

Tree-Constrained Gradient Ascent

For this experiment our own implementation of the Tree-Constrained Gradient Ascent algorithm from [38] without post-processing is used, implemented in the R programming

Algorithm 1 The Beam Search algorithm.

```
1: function ArchivingBeamSearch(𝒜, 𝒩, φ, d, w)
2:     B_0 ← {⊤}                          ▷ Start with a base rule matching all observations
3:     for i ← 1 up to d do
4:         B_i ← ∅
5:         for r ∈ B_{i−1} do
6:             for a ∈ 𝒜 do
7:                 for r′ ∈ 𝒩(a, r) do
8:                     r⁻ ← arg min_{r∈B_{i−1}} φ(r)
9:                     if |B_i| = w and φ(r′) > φ(r⁻) then
10:                        B_i ← {r′} ∪ B_i \ {r⁻}
11:                    end if
12:                end for
13:            end for
14:        end for                        ▷ B_i contains the top-w refinements of B_{i−1}
15:    end for
16:    return ⋃_{i=1}^{d} B_i
17: end function
```

language. The reader is referred to Section 2.2.2 and Krak and Feelders [38] for a detailed discussion of the algorithm and pseudocode of its operation.

Krak and Feelders do not specify the exact initialization used in [38], so our initialization is based on the initialization procedure outlined in the earlier work of Krak on this subject in [37, Section 4.4]. This way, random initial solutions that are already biased towards possible exceptionalities in the dataset are generated. As pointed out earlier (see Section 2.2.2), this initialization can be problematic on datasets that contain a few strong outliers, leading to initialization procedure always generating the same subgroup (that is, the subgroup where only those outliers are removed), or only slight variations thereon. To overcome this limitation, the uniformity of the spread of distances in the calculation is improved by taking the square root of the normalized distances when computing the selection probabilities.

6.3.2 Pareto Local Search

As the benchmark multi-objective strategy in this experiment, Pareto Local Search is chosen. Several stochastic variants of Pareto Local Search, as outlined by Drugan and Thierens in [19], are evaluated. All algorithmic variants use the same base algorithm for local search, but use other criteria to choose which solution to explore, and use different restart mechanisms. The base algorithm for the Pareto Local Search experiment is first

defined, proceeded by an outline of the variants to be evaluated. This section heavily draws on the results of Drugan and Thierens in [19], however with rewritten pseudocode to match notational conventions elsewhere in this chapter. For all algorithms described here, our own implementation in the R programming environment is used.

Preliminaries

The base algorithm for Pareto Local Search, as defined by Drugan and Thierens in [19], is given in Algorithm 2.

Algorithm 2 The Pareto Local Search algorithm.

1: **function** ParetoLocalSearch($\mathcal{I}, \mathcal{N}, \varphi, V, R$)
2: $R' \leftarrow R$
3: **while** $\exists r \in R' \setminus V$ **do**
4: $R'' \leftarrow \mathcal{I}(\mathcal{N}, \varphi, R', r)$ ▷ Search neighbourhood
5: $R' \leftarrow \text{NonDominated}(\varphi, R' \cup R'')$ ▷ Update non-dominated front
6: $V \leftarrow V \cup \{r\}$ ▷ Mark rule as visited
7: **end while**
8: **return** V, R'
9: **end function**

Here, $\mathcal{I}(\mathcal{N}, \varphi, R, r)$ denotes a Pareto improvement strategy that yields a set of local improvements from the neighbourhood $\mathcal{N}$ of r with respect to the current front R, measured on dominance relations on φ. Here, $\text{NonDominated}(\varphi, R) = \{r \in R \mid \neg\exists(r' \in R) : \varphi(r') \prec \varphi(r)\}$ yields the non-dominated solutions in R with respect to φ.

Improvement strategies

As the improvement strategy plays an important role in local search, two different strategies, i.e. variants of $\mathcal{I}$, are benchmarked. The performance of the 'first improvement' and 'neutral improvement' strategy, as given by Drugan and Thierens, is tested. Drugan and Thierens give a 'best improvement' strategy as well, but we consider this strategy prohibitively expensive to use in a local search setting for our problem. Their respective pseudocode is listed in Algorithms 3 and 4. The reader is referred to Drugan and Thierens [19] for an in-depth discussion of these strategies and their properties.

Neighbourhood of a rule

We define the neighbourhood $\mathcal{N}(r)$ for our Pareto Local Search experiments as the smallest possible changes to r, under the constraint that the rules satisfy a given minimum support k_{min}. While the search space for ordinal attributes is often continuous, our dataset has a finite number of elements, so the search space can losslessly be discretized

Algorithm 3 The first improvement strategy $\mathcal{I}_F(\mathcal{N}, \varphi, R, r)$

$R' \leftarrow R$
for all $r' \in \mathcal{N}(r)$ in random order **do**
    if $\forall r'' \in R' : \varphi(r') \prec \varphi(r'') \vee \varphi(r') \parallel \varphi(r'')$ **then**
        $R' \leftarrow \textsc{NonDominated}(\varphi, R' \cup \{r'\})$
        if $\varphi(r') \prec \varphi(r)$ **then**
            return $R' \setminus R$
        end if
    end if
end for
return $R' \setminus R$

Algorithm 4 The neutral improvement strategy $\mathcal{I}_N(\mathcal{N}, \varphi, R, r)$

$R' \leftarrow R$
for all $r' \in \mathcal{N}(r)$ in random order **do**
    if $\forall r'' \in R' : \varphi(r') \prec \varphi(r'') \vee \varphi(r') \parallel \varphi(r'')$ **then**
        $R' \leftarrow \textsc{NonDominated}(\varphi, R' \cup \{r'\})$
        return $R' \setminus R$
    end if
end for
return $\emptyset$

for the purpose of this experiment. This limits the neighbourhood to a manageable size, linear in the number of attributes.

The smallest possible changes are considered to be:

- The upper or lower bound of exactly one of the ordinal variables in r is either increased or decreased. Values must always be present in the dataset, except for sentinels positive or negative infinity. After choice of direction, the new value is chosen such that the difference between the original and the new value is minimized and at least one additional data point is added or removed from the support of the rule.
- One allowed value is added or removed for one of the unordered variables in r.

Multi-restart Pareto Local Search

The Pareto Local Search requires an initial unvisited solution to start its exploration. Two different variants on the generation of such solutions are included in this experiment. The first and simplest variant included in this experiment is the Multi-restart Pareto Local Search algorithm given by Drugan and Thierens in [19]. The respective pseudocode is given in Algorithm 5.

Algorithm 5 The Multi-restart Pareto Local Search algorithm.

function MULTIRESTARTPLS($\mathcal{I}, \mathcal{N}, \varphi$)
  $R \leftarrow \emptyset$
  $V \leftarrow \emptyset$
  while the stopping criterion is not met **do**
    $r \leftarrow$ RANDOMSOLUTION()
    $R' \leftarrow$ DEACTIVATE(φ, R, r)
    $V, R' \leftarrow$ PARETOLOCALSEARCH($\mathcal{I}, \mathcal{N}, \varphi, V, R'$)
    $R \leftarrow$ NONDOMINATED($\varphi, R \cup R'$)
  end while
  return V, R
end function

The deactivation function DEACTIVATE(φ, R, r) = NONDOMINATED($\varphi, \{r\} \cup \{r' \in R \mid \varphi(r') \parallel \varphi(r)\}$) helps to reduce the number of solutions explored, without loss of completeness when the 'best improvement' strategy is used [19]. While this improvement strategy is not used in this experiment, we still consider deactivation beneficial to the speed of convergence, as suggested by Dubois for strategies that reduce the breadth of the search in general. [20]

Initialization

At initialization on restart, as denoted in Algorithm 5 by the RandomSolution procedure, a random rule is generated subject to the constraint of minimum support k_{min}. Rules are chosen by iteratively adding constraints to an initially empty rule (a rule that matches all records). At each iteration, a restrictive clause on one of the attributes — sampled uniformly random without replacement — is added to the rule, the value of which is also sampled uniformly random, such that the expected decrease in support is 50%. After each iteration, the algorithm stops with a probability chosen such that the expected support of a generated rule is $4k_{min}$, or continues for another iteration. The algorithm stops immediately when the rule voids the minimum support constraint, returning the last-valid rule instead.

Genetic Pareto Local Search

Assuming mutual information in the population, genetic algorithms may outperform random restarts. To this end, a genetic variant on Multi-restart Pareto Local Search is included in this experiment, as a second variant on the generation of initial solutions for Pareto Local Search. The pseudocode for Genetic Pareto Local Search, based on that from Drugan and Thierens in [19], is given in Algorithm 6.

Algorithm 6 The Genetic Pareto Local Search algorithm.

```
1: function GeneticPLS(I, N, φ, α)
2:     V, R ← MultiRestartPLS(I, N, φ)
3:     while the stopping criterion is not met do
4:         Select r randomly from R
5:         if α > U(0, 1) or |R| < 2 then
6:             r′ ← Mutate(r)
7:         else
8:             Select r″ ≠ r randomly from R
9:             r′ ← Recombine(r, r″)
10:        end if
11:        R′ ← Deactivate(φ, R, r′)
12:        V, R′ ← ParetoLocalSearch(I, N, φ, V, R′)
13:        R ← NonDominated(φ, R ∪ R′)
14:    end while
15:    return R
16: end function
```

Note that the internal stopping criterion for the call to MultiRestartPLS must be one that is met far earlier than the stopping criterion for the entire algorithm.

Genetic encoding
Encoding of 'chromosome = rule' is used, where rules are in conjunctive normal form. The population then describes a collection of rules. For each ordinal attribute, a constraint is encoded as a tuple of (*center*, *width*). Unordered (i.e. nominal/categorical) attributes are represented as sets of possible values.

Mutation
We define the mutation operator MUTATE as one or more moves of center or width, or addition(s) or deletion(s) of a possible value. The number of such moves is drawn from the exponential distribution with rate $\lambda = 1$, rounded towards positive infinity. This gives a number of mutations expected between 1 and 2, but never smaller than 1, and with a slight, but strongly decreasing, probability of a larger number of mutations (thereby being more influential but also more destructive).

In contrast to the perturbations in the neighbourhood definition, mutations must not necessarily be the smallest possible change. Instead, we define a mutation on the center of a constraint as addition with a sample from the normal distribution with mean $\mu = 0$ and standard deviation $\sigma = \mathit{width}/2$. A mutation on the width is then defined as multiplication with a sample from the exponential distribution with rate $\lambda = 1$, in order to prevent the width becoming negative. We finally define a mutation on possible values to be a uniform random sample of uniform random length of not-yet-possible (or currently possible) values for a single attribute to be added (or deleted) to (or from) the set of possible values. Care is taken to disqualify such samples that would lead to a violation of k_{min}.

Recombination
No structure is assumed in the positions of attributes in the dataset at hand, so the recombination operator RECOMBINE is based on uniform crossover. However, a uniform choice between all constraints of both parent rules is highly likely to result in a rule with support below k_{min}. Therefore one parent, p_1, is randomly picked as the base, and constraint attributes (either center, width, or a single possible value) are picked uniform randomly and for each such attribute, a value is selected with equal probability from either parent. This process continues until there are no attributes left, or the current choice would lead to a violation of the minimum support.

6.4 General setup

Results from $r = 10$ independent runs are collected for each combination of dataset and algorithm. The minimum support is fixed to $k_{min} = 50$, representing a sound minimum considering the number of dependent variables in the regression models used (see Table

Experiment	Algorithm	Configuration
ABS	Archiving Beam search	$d = 20$ $w = 50$
TCGA	Tree-Constrain Gradient Ascent	$\eta = 0.1$ $min\_split = 2k_{min}$
MPLS f	Multi-restart Pareto Local Search	$\mathcal{I} = \mathcal{I}_F$
MPLS n	Multi-restart Pareto Local Search	$\mathcal{I} = \mathcal{I}_N$
GPLS f	Genetic Pareto Local Search	$\mathcal{I} = \mathcal{I}_F$ $\alpha = 0.5$ $T_{MPLS} = 90\,\mathrm{s}$
GPLS n	Genetic Pareto Local Search	$\mathcal{I} = \mathcal{I}_N$ $\alpha = 0.5$ $T_{MPLS} = 90\,\mathrm{s}$

Table 6.4: Algorithm-specific settings for the algorithms under comparison.

6.3). To not put any algorithm at an inherent advantage, the stop criterion is chosen in time units instead of a particular number of evaluations, the definition of which may vary depending on the algorithm at hand. All algorithms are set to stop after 30 minutes of computation time each, or earlier when they terminate beforehand (as may be the case with Archiving Beam Search).

Table 6.4 lists all algorithmic configurations under consideration to be evaluated, as well as their settings.

For Archiving Beam Search, a beam width of $w = 50$ is chosen, allowing for a spread of different solutions, and the search depth is set to $d = 20$. The latter is chosen artificially high in order not to put Beam Search at an inherent disadvantage for the higher dimensional datasets. Note that we don't expect many results in the final beams due to the subgroup size being strictly monotonically decreasing in search level in Beam Search, soon reaching the limits of minimum support k_{min}.

Tree-Constrained Gradient Ascent is configured with ascent step size $\eta = 0.1$, a value already suggested by its author ([37]) and chosen to prevent immediate overfitting that is aggravated by the algorithm's suboptimal initialization procedure (see Section 6.3.1 and 2.2.2).

Finally, for the Genetic Pareto Local Search variants, the mutation probability is chosen as $\alpha = 0.5$, lacking prior knowledge on the effects of mutation and recombination on this class of problems under the encoding described in Section 6.3.2. The time spent on initialization of the Pareto front by means of Multi-restart Pareto Local Search is limited to $T_{MPLS} = 90$ seconds, representing a 5% share of the total running time of the Genetic Pareto Local Search algorithm.

Each algorithm run is performed in a single thread on an Intel(R) Core(TM) i5-3470 CPU @ 3.20GHz, running the GNU/Linux kernel version 4.1.8. The execution environment is R version 3.2.3 [48], enhanced by the `compiler` package for just-in-time compilation of program code. Time is measured as actual CPU time ("user time") allocated to the execution of the algorithm.

Chapter 7

Results

For each dataset, the quality of the Pareto non-dominated set found by each algorithm is measured in 10 independent runs, and the results ranked by hypervolume of this set. A summary of the results obtained is outlined in Table 7.2. Either Archiving Beam Search or a Pareto Local Search variant performs best on average, Tree-Constrained Gradient Ascent is consistently outperformed by either one of the former; as well on the measure of hypervolume as well as on the more direct measure of evidence, that is, the modified Cook's Distance. It must be noted that the number of solutions on the Pareto non-dominated front reported varies wildly among the algorithms.

To the collection of hypervolumes obtained for all runs of the algorithms under comparison, the Kruskal–Wallis rank sum test is applied to test for existence of a difference between the results obtained by each of the algorithms, that is, whether the choice of algorithm significantly impacts the hypervolume obtained. The results for this test are outlined in Table 7.1. For all datasets a significant result is obtained (at confidence level $\alpha = 0.05$ with 5 degrees of freedom). We must thus reject the null hypothesis of the Kruskal–Wallis test for all datasets, and may assume that the choice of algorithm influences the obtained hypervolume.

As such, we can proceed with the Conover–Inman post-hoc procedure to establish in more detail which algorithms perform better than others for each dataset. The results for this procedure are outlined in Table 7.3, along with the mean rank of each algorithm. Here, each row algorithm is measured against each worse column algorithm and tested for significance of their difference in results (at confidence level $2\alpha = 0.05$ with 54 degrees of freedom[1]). The better algorithms consistently outperform their lesser algorithms significantly, except within the class of Pareto Local Search algorithms, where

[1]The Cononver–Inman procedure, which is essentially a t-test, finds $p = 0.5$ when the ranks are exactly equal. As such, α must be chosen half that of the Kruskal–Wallis test to obtain results at the same confidence level.

Dataset	**Best algorithm**	p**-value**
Ames Housing	ABS	$\mathbf{6.58 \times 10^{-9}}$
Windsor Housing	any *PLS	$\mathbf{2.29 \times 10^{-11}}$
Wine	MPLS f	$\mathbf{8.82 \times 10^{-9}}$
Student Performance (por)	MPLS f	$\mathbf{5.08 \times 10^{-10}}$
Student Performance (mat)	MPLS n	$\mathbf{4.54 \times 10^{-8}}$
Census 90	ABS	$\mathbf{3.74 \times 10^{-6}}$

Table 7.1: Results of the Kruskal–Wallis test on the influence of the choice of algorithm on the hypervolume results for each dataset. For each dataset, the best algorithm and its significance (e.g. the best algorithm performs significantly better than another algorithm at a confidence level of $\alpha = 0.05$) is given. Significant values are typeset bold.

results are often close. A final exception is Tree-Constrained Gradient Ascent in *Census 90*, that matches up closely with the Pareto Local Search class of algorithms under the neutral improvement strategy.

For *Windsor Housing*, Table 7.2 indicates that all Pareto Local Search-based algorithms obtain a Pareto non-dominated front of identical quality. As such, we cannot identify a unique best algorithm for this dataset, as demonstrated in Table 7.1. It is however clear from Table 7.3 that the Pareto Local Search class of algorithms all perform significantly better than either Archiving Beam Search or Tree-Constrained Gradient Ascent, no matter the choice of particular Pareto Local Search algorithm tested.

Then, Figure 7.1 shows the development of the obtained hypervolume for each algorithm, averaged over all runs. As can be seen from the figures, Archiving Beam Search often terminated long before its maximum execution time, producing no new results. For *Windsor Housing*, results consistently appear to have converged after less than 10 minutes of execution. For all other combinations of algorithms and datasets, the rate of convergence has strongly dropped near the end of execution, but the solutions do not yet appear to be stable at that point.

Ames Housing

	Front	Hypervolume	Evidence	Generality	Confidence
ABS	35	1172.4	8340.5	0.995	0.744
TCGA	20.6	176	544.3	0.572	0.814
MPLS n	2560.4	105.4	2601.2	0.713	0.981
MPLS f	17594.6	67.6	229.7	0.535	0.96
GPLS f	17040.6	67.1	228.1	0.532	0.962
GPLS n	1587.9	43.3	1512.6	0.34	0.97

Windsor Housing

	Front	Hypervolume	Evidence	Generality	Confidence
MPLS n	58	12.4	29.9	1	0.777
MPLS f	58	12.4	29.9	1	0.777
GPLS n	58	12.4	29.9	1	0.777
GPLS f	58	12.4	29.9	1	0.777
ABS	28	11.9	29.9	0.978	0.652
TCGA	7.3	10.1	28	0.689	0.577

Wine

	Front	Hypervolume	Evidence	Generality	Confidence
MPLS f	4869.7	784.5	17382.1	1	0.817
GPLS n	3116.2	787.3	19445.6	1	0.817
MPLS n	2964.5	781.2	18364.6	1	0.817
GPLS f	4602.8	706.2	9798.3	1	0.797
ABS	14	599.9	4096.9	0.949	0.732
TCGA	8.5	371.8	7054.5	0.166	0.406

Student Performance (por)

	Front	Hypervolume	Evidence	Generality	Confidence
MPLS f	4728.1	18.2	138.5	1	0.988
GPLS f	4390.6	17.5	127.7	1	0.988
MPLS n	1794.8	16.6	94.3	1	0.987
GPLS n	2409.7	15.9	99.4	1	0.987
ABS	44	13.3	124.6	0.998	0.899
TCGA	11.2	8.5	13.7	0.889	0.921

Student Performance (mat)

	Front	Hypervolume	Evidence	Generality	Confidence
MPLS n	1410.6	30.8	60.8	1	0.991
MPLS f	2019.3	30.1	55.6	1	0.992
GPLS n	1550.8	30.1	57.2	1	0.991
GPLS f	2227.2	28.5	44.6	1	0.99
ABS	38	27.7	48.5	0.98	0.945
TCGA	4.9	25.7	37.2	0.853	0.939

Census 90

	Front	Hypervolume	Evidence	Generality	Confidence
ABS	49	459.1	22443.6	0.994	0.508
GPLS f	7335.2	244.1	1451.7	0.801	0.667
MPLS f	6515.1	229.2	1135.3	0.8	0.661
GPLS n	3602.7	124.8	2217.2	0.427	0.911
TCGA	2.1	109.1	705	0.458	0.338
MPLS n	3128.8	109.1	2729.2	0.458	0.876

Table 7.2: Result summary over the runs of all algorithms on all datasets. The rows (algorithms) are ordered by their respective mean ranks as outlined in Table 7.3. Front and Hypervolume denote the mean size of the Pareto non-dominated front of the final solution of all runs, Evidence, Generality and Confidence denote the mean over all runs of the maximum of each respective property of the solutions on the Pareto non-dominated front.

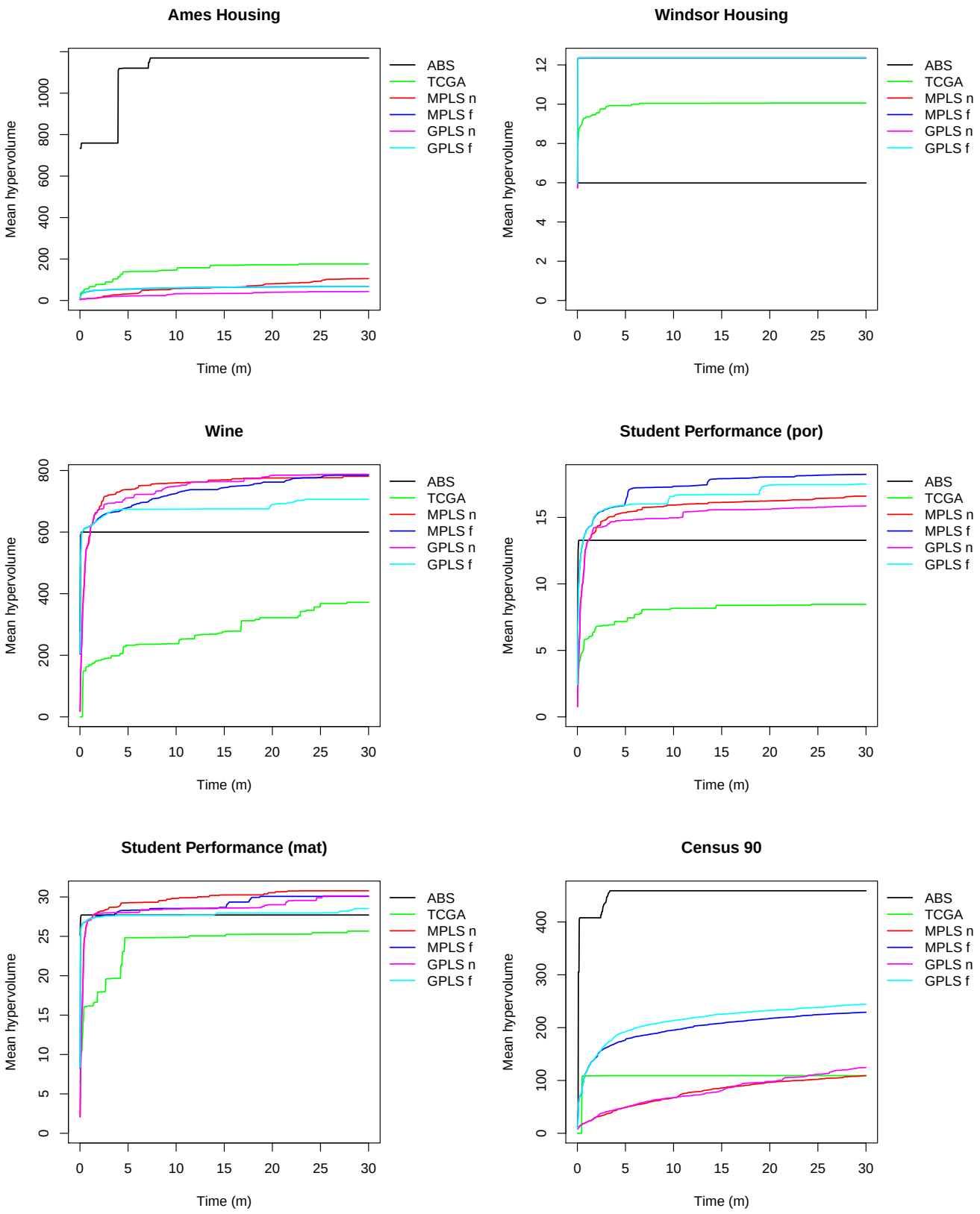

Figure 7.1: Mean results of attained hypervolume over all runs of each algorithm during execution.

Ames Housing

	Mean rank	TCGA	MPLS n	MPLS f	GPLS f	GPLS n
ABS	5.5	**0.005**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**
TCGA	15.5		**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**
MPLS n	29.9			**0.001**	**< 0.001**	**< 0.001**
MPLS f	41.8				0.384	0.069
GPLS f	42.9					0.116
GPLS n	47.4					

Windsor Housing

	Mean rank	MPLS f	GPLS n	GPLS f	ABS	TCGA
MPLS n	20.5	0.5	0.5	0.5	**< 0.001**	**< 0.001**
MPLS f	20.5		0.5	0.5	**< 0.001**	**< 0.001**
GPLS n	20.5			0.5	**< 0.001**	**< 0.001**
GPLS f	20.5				**< 0.001**	**< 0.001**
ABS	45.5					**< 0.001**
TCGA	55.5					

Wine

	Mean rank	GPLS n	MPLS n	GPLS f	ABS	TCGA
MPLS f	11.1	**0.010**	**0.002**	**< 0.001**	**< 0.001**	**< 0.001**
GPLS n	20.2		0.249	**0.024**	**< 0.001**	**< 0.001**
MPLS n	22.8			0.093	**< 0.001**	**< 0.001**
GPLS f	27.9				**< 0.001**	**< 0.001**
ABS	45.5					**0.006**
TCGA	55.5					

Student Performance (por)

	Mean rank	GPLS f	MPLS n	GPLS n	ABS	TCGA
MPLS f	8.2	**0.005**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**
GPLS f	15.7		**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**
MPLS n	26.6			0.042	**< 0.001**	**< 0.001**
GPLS n	31.5				**< 0.001**	**< 0.001**
ABS	45.5					**< 0.001**
TCGA	55.5					

Student Performance (mat)

	Mean rank	MPLS f	GPLS n	GPLS f	ABS	TCGA
MPLS n	16.2	0.364	0.155	**0.003**	**< 0.001**	**< 0.001**
MPLS f	17.7		0.251	**0.008**	**< 0.001**	**< 0.001**
GPLS n	20.6			0.036	**< 0.001**	**< 0.001**
GPLS f	28.5				**< 0.001**	**< 0.001**
ABS	44.5					**0.007**
TCGA	55.5					

Census 90

	Mean rank	GPLS f	MPLS f	GPLS n	TCGA	MPLS n
ABS	5.5	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**
GPLS f	26.3		0.405	**0.006**	**0.006**	**0.002**
MPLS f	27.6			**0.012**	**0.010**	**0.003**
GPLS n	40.2				0.478	0.309
TCGA	40.5					0.329
MPLS n	42.9					

Table 7.3: Results of the Conover–Inman post-hoc procedure on each algorithm's performance measured by hypervolume. Algorithms are ordered by their mean ranks. Significant values, where the row algorithm performs significantly better than the column algorithm, are typeset bold at the same confidence level as used for Kruskal–Wallis in Table 7.1 (e.g. $2\alpha = 0.05$).

Chapter 8

Discussion

From the results in Table 7.2, the potential of our multi-objective approach over existing single-objective approaches immediately becomes clear. Here it can be seen that for most of the datasets under evaluation, not only does our multi-objective approach find a solution front with a larger hypervolume — an indicator for the diversity and hence usefulness of the solution front — but also does it find a solution front that contains solutions with higher evidence, that is, more exceptional models. This indicates that the inclusion of more than a single objective allows the strategy to break out of local optima, where the existing single-objective algorithms would get stuck. An illustrative example of this can be found in Figure 7.1, the graph for Student Performance (por) in particular. Here, the single-objective Archiving Beam Search gets stuck in a local optimum almost immediately, and Tree-Constrained Gradient Ascent fails to find substantial improvements after about 15 minutes of computation time. Our multi-objective Pareto Local Search strategies however continue to find better solutions, clearly escaping local optima, as can in particular be observed in the strong bumps in the mean hypervolumes for the Pareto Local Search variants under the first-improvement strategy in Figure 7.1.

In the subsequent sections, we provide a more in-depth discussion on the performance of the algorithms under comparison, and the factors contributing to their respective differences.

8.1 Convergence of the algorithms

All classes of algorithms show greatly varying rates of hypervolume convergence, as can be seen in the mean hypervolume progressions in Figure 7.1.

Except for its results on Windsor Housing, Archiving Beam Search reaches a nearly converged hypervolume in the shortest amount of time. This is to be expected from the

way the algorithm operates. Archiving Beam Search first explores all single-clause rules in the concept space, and keeps just a few (in this case, 50) of them having maximum evidence. The number of possible single-clause rules is bound by the number of unique values in the dataset, which cannot exceed n times the number of concept attributes. These initial solutions can thus be explored in a limited amount of time. As the F-measure of our modified Cook's distance is used as a quality measure for Archiving Beam Search, more specific rules have equal expected evidence compared to less specific rules, i.e. the quality measure is expected to be invariant under generality. Therefore, a reasonable spread over generality is expected after optimizing for evidence solely. The effect of generality on goodness-of-fit is inversely proportional in general, but as k_{min} is chosen much higher than p for any dataset, no great effect can be expected. As Archiving Beam Search optimizes solely on evidence, addition of clauses to any rule, as done in every sequential search level, is not expected to greatly increase the attained hypervolume.

The Pareto Local Search algorithms under the first-improvement strategy can be seen to exhibit similar behaviour. In these algorithms, again a large number of neighbourhood solutions is explored and added to the non-dominated front, even when they do not strictly dominate the base solution under exploration. In support of a multi-objective approach, on most datasets the Pareto Local Search algorithms soon reach hypervolume parity with Archiving Beam Search, and continue to improve where Archiving Beam Search fails to find further improvements.

For the Pareto Local Search variants, before the solution front can be considered converged, the maximum attained generality must be nearing 1 — under the assumption that the local search starts from rules of low generality, as is the case in this experiment. This follows immediately from the definition of non-domination under the objective functions in this experiment: if the most general rule in the population is relaxed, increasing its generality, the resulting rule will always be non-dominated in the population, and the hypervolume is increased[1].

It can be observed in Table 7.2 that for all datasets where Pareto Local Search variants perform well, their solution Pareto non-dominated front has maximum generality of (approximately) 1. On the other datasets, the maximum generality stays far from 1. From this we must assume that on the datasets where the Pareto Local Search variants did not perform well, this is at least partially explained by lack of convergence. In the following section, we will discuss one of the factors that we find to strongly influence the rate of convergence.

[1]The single exception to this is when the empty rule — a rule that imposes no restrictions on the concept, thereby matching all observations in the dataset — is added to the population: the evidence of the empty rule is zero, and hence adding the rule does not increase the hypervolume.

8.1.1 Impact of concept space dimensionality

Notable exceptions to the convergence of the Pareto Local Search variants can be observed for the Ames Housing and Census 90 datasets, where Archiving Beam Search and Tree-Constrained Gradient Ascent solidly outperform our multi-objective strategies. Table 7.2 and Figure 7.1 provide some insight in the mechanism behind this. From Table 7.2 we learn that the Pareto Local Search strategies, in particular those using the first-improvement strategy, generate an immense Pareto non-dominated front. As for each solution on the front, its neighbourhood must be explored, this inherently leads to slow convergence. This is further aggravated by the increased computation complexity of checking for non-domination on larger fronts.

The vast size of the non-dominated fronts on some of the datasets may be explained by the relative size of their concept spaces. Ames Housing, in particular, has a vast number of ordinal attributes. Census 90 features only a small number attributes, however most of them nominal with large dimensionality. Both datasets therefore sport a large concept space, i.e. the decision space represented by all possible constraints on the attributes. The other datasets feature either a smaller number of attributes, or attributes of small dimensionality, and hence have considerably smaller concept spaces. When the concept space of a dataset is of large enough dimensionality, an almost arbitrary selection of observations can be represented by a set of constraints on the attributes. In such a case, the minimal relaxation of a constraint will lead to the addition of only small number of observations in the subgroup, and a proper choice of constraint allows for nearly all selections of observations to add. An equivalent observation can be made when a constraint is made slightly more specific instead of more relaxed. In general, an appropriately chosen observation can be omitted from the subgroup without decreasing the subgroup's evidence. This smaller subgroup will feature a worse generality, but often a slightly better confidence, as R^2 is slightly inversely proportional to subgroup size. As such, the original subgroup and its smaller variant will often both be non-dominated by the other, and the smaller subgroup will have a high likelihood of being non-dominated in the population if the original subgroup was non-dominated there as well.

Note that these symptoms of slow convergence are related to the choice of the neighbourhood operator, as minimal changes do not necessarily have to belong to the neighbourhood of a rule. Pareto Local Search strategies with the three-dimensional objective of evidence, generality and confidence, can be expected to converge slowly on datasets with high-dimensional concept spaces under the neighbourhood operator used in this experiment. It remains an open question whether a different choice of neighbourhood operator may alleviate the severity of slow convergence.

8.1.2 Impact of bias towards complex rules

As becomes clear from Figure 7.1, on Ames Housing, the Archiving Beam Search reaches a hypervolume far higher than that of the other algorithms, already at its first search level. The mean evidence, listed in Table 7.2, follows this observation. Considering the bottom-up mode of rule exploration of Archiving Beam Search, this means that a small collection of simple rules — rules with a small number of clauses — can already span a large hypervolume and have high evidence in this dataset. Apparently, Ames Housing features simple yet high-quality subgroups: a small selection of its attributes appears to have a strong influence on the regression estimates. The same observations appear to hold for the Census 90 dataset.

Due to the initialization procedure chosen for the Pareto Local Search variants, their mode of operation is more top-down, thus biased towards complex rules. This bias proves detrimental on datasets such as Ames Housing and Census 90, for it results in failure to explore simple rules in an early stage. Given this strong clue as to why the Pareto Local Search variants are consistently outperformed on these datasets, another initialization procedure that is more biased towards simple rules may be considered to overcome this limitation. Tree-Constrained Gradient Ascent appears to be affected by the same method of failure, but here the bias towards complex rules is inherent to the algorithm and cannot be fixed by another choice of initialization.

8.1.3 Sensitivity to implementation efficiency

While care has been taken to implement each algorithm as computationally efficient as possible, there may still be possible improvements to each implementation that the author is unaware of. All algorithms are implemented in the same programming language, and experiments performed under identical conditions on the same machine. Still, it cannot be denied that seemingly minor choices in implementation may impact the quality of the results as long as an algorithm has not fully converged when the stop criterion is reached. This is inherent to measuring an algorithm's performance against time, instead of some fixed and comparable measure such as number of fitness function evaluations. Due to the different natures of the algorithms under consideration, there exists no such a common measure other than computation time. To minimize the impact of any discrepancies in computational efficiency, an attempt was made to choose the maximum computation time to be sufficiently high. Even though most experiments do not appear to have converged when the stop criterion is reached, this choice allows us to observe from the rate of hypervolume increase in Figure 7.1, that significant differences in the relative performance of the algorithms are unlikely to arise within this order of magnitude of computation time.

	Hypervolume	Front	Evidence	Generality	Confidence
ABS	1172.4	35	8340.5	0.995	0.744
TCGA	176	20.6	544.3	0.572	0.814
MPLS n EGC	105.4	2560.4	2601.2	0.713	0.981
MPLS n EG	141	361	1573.5	0.723	0.842
MPLS n EC	38.6	48	2385	0.717	0.974
MPLS f ECG	67.6	17594.6	229.7	0.535	0.96
MPLS f EG	122.2	849.4	990	0.666	0.793
MPLS f EC	33.4	464.9	893.8	0.524	0.955
GPLS n ECG	43.3	1587.9	1512.6	0.34	0.97
GPLS n EG	226.7	525.6	3606.2	0.721	0.894
GPLS n EC	34.2	72.1	2001	0.617	0.961
GPLS f ECG	67.1	17040.6	228.1	0.532	0.962
GPLS f EG	161.7	783	2054.8	0.663	0.832
GPLS f EC	39.6	551.1	1199.4	0.44	0.951

Table 8.1: Result summary over the runs of all considered algorithms and objective functions. Front and Hypervolume denote the mean size of the Pareto non-dominated front of the final solution of all runs, Evidence, Generality and Confidence denote the mean over all runs of the maximum of each respective property of the solutions on the Pareto non-dominated front. Hypervolume is measured in the three-dimensional evidence – generality – confidence objective regardless of the objective function used in the experiment on Ames Housing.

8.2 Choice of objective functions

As pointed out in Section 8.1.1, some datasets may feature enormous Pareto non-dominated fronts under the choice of objective functions used in this experiment, leading to slow convergence. To get an impression of how influential the choice of objective functions is, and with that choice the dimensionality of the objective, an additional experiment is performed. To this end, all Pareto Local Search variants are again evaluated on Ames Housing, the results of which were previously hypothesized to be affected by the choice of objective functions, however with different choices of objective functions. For all Pareto Local Search variants, the complete three-dimensional evidence – generality – confidence (EGC) as well as the two-dimensional evidence – generality (EG) and evidence – confidence (EC) objective functions are evaluated. As in the original experiment, each experiment variant is measured for 10 runs. To allow us to compare the results for all experiments, we still evaluate the original three-dimensional hypervolume as described in Section 6.2.

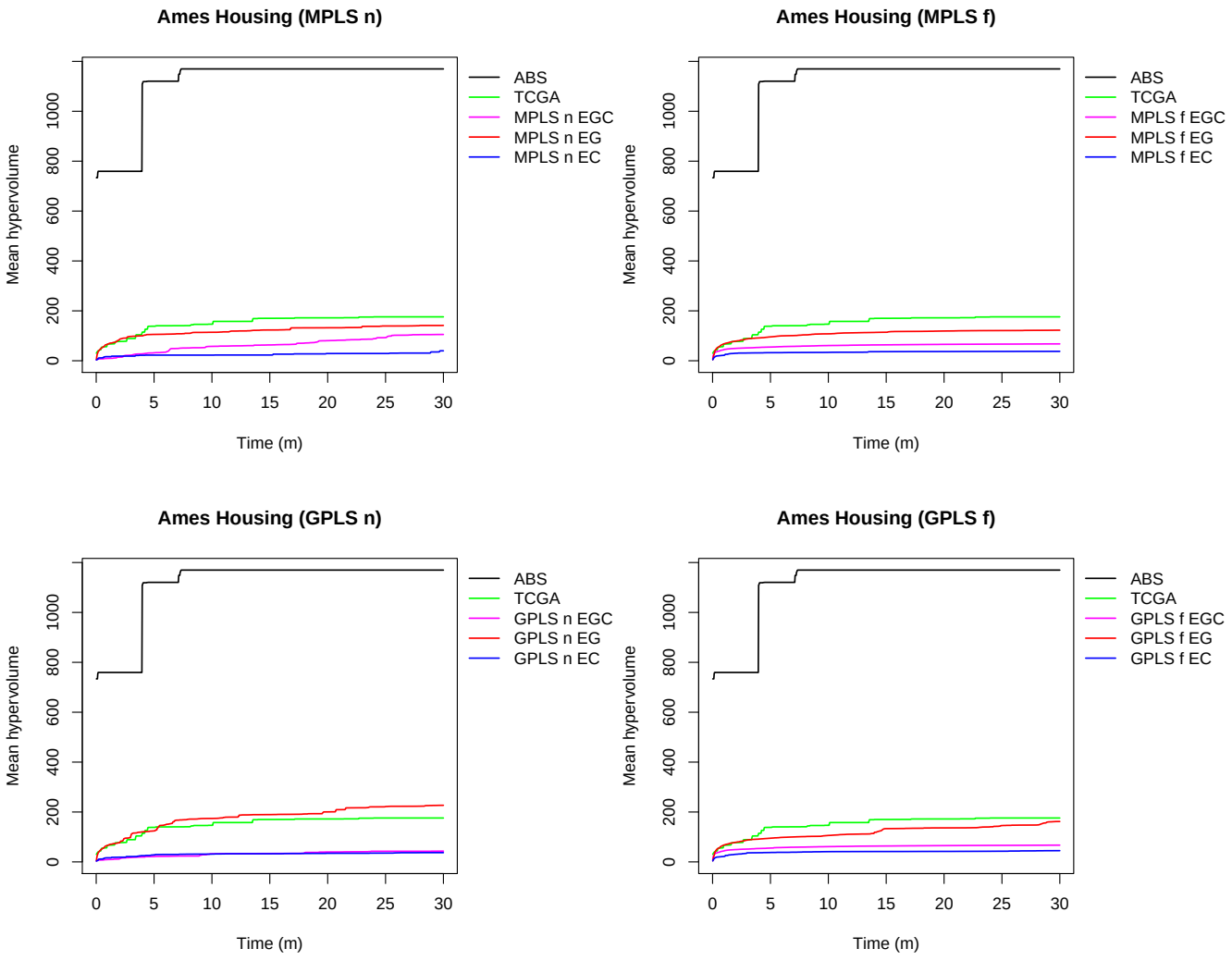

Figure 8.1: Mean results of attained full-dimensional hypervolume over all runs of Pareto Local Search during execution, for different selections of objective functions. Results are grouped by their Pareto Local Search variant and improvement strategy, and include the reference results from the original experiment for Archiving Beam Search and Tree-Constrained Gradient Ascent.

The mean progression of hypervolume over time for these variants is shown in Figure 8.1. From this figure, it immediately becomes clear that the choice of objective functions strongly impacts the results. As previously hypothesized in Section 8.1.1, the combination of the generality and confidence objectives is likely to strongly increase the size of the Pareto non-dominated front and hence reduce the speed of convergence to a point where it impacts the quality of the results obtained when the stop criterion is reached. Table 8.1 supports this observation: the Pareto non-dominated front size is consistently larger when both objectives are included. For all Pareto Local Search variants, the difference in hypervolume between the EG and either the EGC or EC objective function of the same variant shows to be significant in the Conover–Inman post-hoc procedure at a confidence level of $2\alpha = 0.05$ (MPLS n: $p = 0.024$, MPLS f: $p = 0.003$, GPLS n: $p < 0.001$, GPLS f: $p < 0.001$). Figure 8.1 shows that the mean hypervolume of Genetic Pareto Local Search with the neutral-improvement strategy is greater than that of Tree-Constrained Gradient Ascent. This difference however is not significant ($p = 0.446$).

Comparing the mean maximum attained generality for the different objective functions, it become clear that the experiments under the evidence – generality objective function get closer to convergence than those under the other objective functions. The merits of this increased convergence are demonstrated by both a vastly greater mean hypervolume and greater mean maximum evidence, represented by far less solutions in the Pareto non-dominated solution front. Multi-restart Pareto Local Search under the neutral-improvement strategy forms a single exception: here, all mean maximum generalities are roughly equal and convergence can be considered comparable as well.

8.3 Effectiveness of a genetic approach

Judging from Table 7.3, the Genetic Pareto Local Search variants never perform significantly better than their multi-restart counterparts. From this, we learn that the mutation and recombination operators as used in this experiment must be unlikely to yield high-quality offspring. Considering that the recombination operator used is a simple uniform crossover (only modified to bail out early when the minimum support k_{min} is void), the expected performance of the genetic approach is based on the assumption of similarity of solutions in the population. The lack of a performance improvement over random restarts hence implies that solutions on the Pareto non-dominated front encountered in the experiment have little in common. Note that the mutation operator, when applied generously, can with probability mimic random restarts, which explains why Genetic Pareto Local Search does not always perform significantly worse than its multi-restart counterpart. Without the expected improvement over random restarts, we must judge the genetic approach taken in this experiment to be ineffective when com-

pared to the multi-restart approach. As our results suggest the absence of a common denominator in rules on the Pareto non-dominated front, we may assume that a genetic approach for this class of data mining problem is unlikely to be effective in general.

8.4 Tree-Constrained Gradient Ascent initialization anomalies

Tree-Constrained Gradient Ascent can almost universally be seen to slowly but gradually reach a point where the hypervolume stabilizes. A notable exception here is the Census 90 dataset, where a sudden increase is seen after a minute of computation, followed by no substantial improvement at all. Table 7.2 provides a clue as to what happened here: the non-dominated Pareto front contains only 2.1 solutions on average. This means that Tree-Constrained Gradient Ascent was usually unable to find new non-dominated solutions after a short amount of time, where only a few solutions had made it to the front. Given the 3-dimensional nature our objective function, any three reasonable solutions would likely have been non-dominated by each other, and would thus have been included in the front. From the lack thereof we can conclude that on Census 90, the initialization procedure is probably at fault, generating (almost) always the same initial solutions. Another choice of power — a square root has been introduced in Section 6.3.1 to slightly alleviate this problems already — or another choice of initialization altogether might have alleviated this. Further analysis of this problem is however not within the scope of this work.

8.5 Directions for future research

Our results point out that Tree-Constrained Gradient Ascent can be a useful strategy for some datasets, but fails on others. As discussed earlier, Tree-Constrained Gradient Ascent features potential for improvement in its initialization procedure, the correctness of the derivative of its objective function and the choice of step size. It remains to be investigated whether any such improvements would lead to considerably better performance.

As for Pareto Local Search, other strategies may be investigated to increase the speed of convergence (and possibly decrease the size of the solution front), which we observed to be one of the bottlenecks of our current multi-objective approach. Various algorithmic variations may lead to such a decrease. In particular, we suggest to investigate the impact of niching mechanisms such as employed in NSGA-II [16], as this has been reported to work well on Subgroup Discovery [6]. Other solutions to directly reduce the

size of the non-dominated front without impacting its spread, such as farthest-candidate pruning [8], may be investigated as well.

We must note that Pareto Local Search is just one of many imaginable multi-objective search strategies. As such, we suggest future research on other strategies that leverage the multi-objective nature of Exceptional Model Mining. In particular, strategies that use more domain knowledge about the models at hand, as is the case in Tree-Constrained Gradient Ascent, may be worth investigating.

As our small-scale experimental results in Section 8.2 show, our current choice of objectives leaves room for improvement. Further investigation may also reconsider the choice of objective functions in this experiment.

Chapter 9

Conclusion

Existing single-objective approaches to Exceptional Model Mining leave much to be desired. Exceptional Model Mining is inherently a multi-objective problem, and so is knowledge discovery in general. The most pressing issue with single-objective approaches to knowledge discovery is the relation between generality and evidence, both of which are features of interest to the analyst, but can often not be optimized simultaneously due to their inversely proportional relationship. For Exceptional Model Mining on linear regression models, the trade-off between these two properties has not yet seen much analysis in the existing literature.

Our analysis of Cook's distance measure for subgroups in Exceptional Model Mining has shown that the usual correction applied to compensate for subgroup size cannot be considered valid from a statistical point of view. We have developed a modified Cook's distance that is a proper internal influence measure, and have derived its statistical distribution. This distribution correctly recognizes the finite population sampling properties of Cook's distance for subgroups, taking subgroup size into account. Empirical results have confirmed that the derived distribution on this measure is indeed correct.

Furthermore we have demonstrated that multi-objective Pareto Local Search strategies yield significant improvements over existing single-objective approaches on the majority of the datasets under evaluation. Here, the multi-objective strategies yield a greater hypervolume, and thus a greater choice for the analyst. As their results also feature greater evidence, the multi-objective strategies outperform the existing evidence-based single-objective algorithms, even when only considering the objective they tend to optimize. This demonstrates the ability of the multi-objective strategies to escape from local optima. For the minority of datasets where our strategies did not perform well, a reasonable explanation for their failure can be found in the exact choice of objectives and initialization procedure. Hence, we must conclude that Exceptional Model Mining can certainly benefit from a multi-objective approach.

Bibliography

[1] P. M. Anglin and R. Gençay. "Semiparametric estimation of a hedonic price function". In: *Journal of Applied Econometrics* 11.6 (1996), pp. 633–648.

[2] J. M. Bader. "Hypervolume-Based Search for Multiobjective Optimization: Theory and Methods". PhD thesis. ETH Zurich, 2009.

[3] R. Beckman and H. Trussell. "The distribution of an arbitrary studentized residual and the effects of updating in multiple regression". In: *Journal of the American Statistical Association* 69.345 (1974), pp. 199–201.

[4] R. Bisiani. "Beam Search". In: *Encyclopedia of Artificial Intelligence*. Ed. by S. C. Shapiro and D. Eckroth. Vol. 1. John Wiley & Sons, 1987, pp. 56–58.

[5] F. D. Blau and L. M. Kahn. "Swimming upstream: Trends in the Gender Wage Differential in the 1980s". In: *Journal of Labor Economics* (1997), pp. 1–42.

[6] C. J. Carmona et al. "NMEEF-SD: Non-Dominated Multiobjective Evolutionary Algorithm for Extracting Fuzzy Rules in Subgroup Discovery". In: *Fuzzy Systems, IEEE Transactions on* 18.5 (2010), pp. 958–970.

[7] C. Chatfield and A. J. Collins. *Introduction to Multivariate Analysis*. Chapman and Hall, 1980.

[8] B. Chen et al. "A New Local Search-Based Multiobjective Optimization Algorithm". In: *Evolutionary Computation, IEEE Transactions on* 19.1 (2015), pp. 50–73.

[9] W. G. Cochran. *Sampling techniques*. 3rd ed. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1977.

[10] R. D. Cook. "Detection of influential observation in linear regression". In: *Technometrics* 19.1 (1977), pp. 15–18.

[11] R. D. Cook and S. Weisberg. "Characterizations of an empirical influence function for detecting influential cases in regression". In: *Technometrics* 22.4 (1980), pp. 495–508.

[12] R. D. Cook and S. Weisberg. *Residuals and influence in regression*. Monographs on statistics and applied probability. Chapman and Hall, 1982.

[13] P. Cortez and A. M. G. Silva. "Using data mining to predict secondary school student performance". In: *Proceedings of 5th Annual Future Business Technology Conference, Porto, 2008*. Ed. by A. Brito and J. Teixeira. EUROSIS. 2008, pp. 5–12.

[14] M. Costanigro, R. C. Mittelhammer, and J. J. McCluskey. "Estimating class-specific parametric models under class uncertainty: local polynomial regression clustering in an hedonic analysis of wine markets". In: *Journal of Applied Econometrics* 24.7 (2009), pp. 1117–1135.

[15] D. De Cock. "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project". In: *Journal of Statistics Education* 19.3 (2011).

[16] K. Deb et al. "A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II". In: *Evolutionary Computation, IEEE Transactions on* 6.2 (2002), pp. 182–197.

[17] J. A. Díaz-García and G. González-Farías. "A note on the Cook's distance". In: *Journal of Statistical Planning and Inference* 120.1 (2004), pp. 119–136.

[18] J. A. Díaz-García, G. González-Farías, and V. M. Alvarado-Castro. "Exact distributions for sensitivity analysis in linear regression". In: *Applied Mathematical Sciences* 1.22 (2007), pp. 1083–1100.

[19] M. M. Drugan and D. Thierens. "Stochastic Pareto local search: Pareto neighbourhood exploration and perturbation strategies". In: *Journal of Heuristics* 18.5 (2012), pp. 727–766.

[20] J. Dubois-Lacoste, M. López-Ibáñez, and T. Stützle. "Anytime Pareto local search". In: *European Journal of Operational Research* 243.2 (2015), pp. 369–385.

[21] W. Duivesteijn. "Exceptional Model Mining". PhD thesis. Leiden Institute of Advanced Computer Science (LIACS), Faculty of Science, Leiden University, 2013.

[22] W. Duivesteijn, A. Feelders, and A. Knobbe. "Different slopes for different folks: mining for exceptional regression models with Cook's distance". In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2012, pp. 868–876.

[23] W. Duivesteijn, A. J. Feelders, and A. Knobbe. "Exceptional model mining". In: *Data Mining and Knowledge Discovery* (2013), pp. 1–52.

[24] C. M. Fonseca, L. Paquete, and M. López-Ibánez. "An improved dimension-sweep algorithm for the hypervolume indicator". In: *IEEE Congress on Evolutionary Computation, 2006*. IEEE. 2006, pp. 1157–1163.

[25] S. García et al. "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power". In: *Information Sciences* 180.10 (2010), pp. 2044–2064.

[26] J. B. Gray. “Approximating the internal norm influence measure in linear regression”. In: *Communications in Statistics—Simulation and Computation* 22.1 (1993), pp. 117–135.

[27] J. B. Gray. “The internal norm approach to influence diagnostics”. In: *Communications in Statistics—Theory and Methods* 18.3 (1989), pp. 943–958.

[28] E. Grodsky and D. Pager. “The structure of disadvantage: Individual and occupational determinants of the black-white wage gap”. In: *American Sociological Review* (2001), pp. 542–567.

[29] H. Hotelling. “The Generalization of Student’s Ratio”. In: *Annals of Mathematical Statistics* 2.3 (Aug. 1931), pp. 360–378.

[30] D. R. Jensen and D. E. Ramirez. “Some exact properties of Cook’s D_I”. In: *Handbook of Statistics*. Ed. by N. Balakrishnan and C. R. Rao. Vol. 16. Elsevier, 1998, pp. 387–402.

[31] D. R. Jensen and D. E. Ramirez. *Computing the* cdf *of Cook’s D_I statistic*. Unpublished. 1996. URL: `http://people.virginia.edu/~der/pdf/der59.pdf` (visited on 02/03/2016).

[32] D. R. Jensen and D. E. Ramirez. *The distribution of Cook’s D_I statistic*. Unpublished. 1998. URL: `http://people.virginia.edu/~der/pdf/der61.pdf` (visited on 02/03/2016).

[33] M. J. del Jesus et al. “Evolutionary Fuzzy Rule Induction Process For Subgroup Discovery: A Case Study in Marketing”. In: *IEEE Transactions on Fuzzy Systems* 15.4 (2007), p. 578.

[34] C. Kim and B. E. Storer. “Reference values for Cook’s distance”. In: *Communications in Statistics—Simulation and Computation* 25.3 (1996), pp. 691–708.

[35] W. Klösgen. “Applications and research problems of subgroup mining”. In: *Foundations of Intelligent Systems: 11th International Symposium, ISMIS’99 Warsaw, Poland, June 8–11, 1999 Proceedings*. Ed. by Z. W. Raś and A. Skowron. Springer, 1999, pp. 1–15.

[36] W. Klösgen. “Explora: A Multipattern and Multistrategy Discovery Assistant”. In: *Advances in Knowledge Discovery and Data Mining*. Ed. by U. M. Fayyad et al. AAAI Press / MIT Press, 1996, pp. 249–271.

[37] T. E. Krak. “Mining Exceptional Linear Regression Models with Tree-Constrained Gradient Ascent”. Master’s thesis. Universiteit Utrecht, 2013.

[38] T. E. Krak and A. J. Feelders. “Exceptional Model Mining with Tree-Constrained Gradient Ascent”. In: *Proceedings of 2015 SIAM International Conference on Data Mining (SDM 2015)*. SIAM. 2015.

[39] T. L. Lai, H. Robbins, and C. Z. Wei. “Strong consistency of least squares estimates in multiple regression II”. In: *Journal of Multivariate Analysis* 9.3 (1979), pp. 343–361.

[40] D. Leman, A. Feelders, and A. Knobbe. “Exceptional model mining”. In: *Machine Learning and Knowledge Discovery in Databases*. Springer, 2008, pp. 1–16.

[41] M. Lichman. *UCI Machine Learning Repository*. 2013. URL: `http://archive.ics.uci.edu/ml`.

[42] B. T. Lowerre. “The HARPY Speech Recognition System”. PhD thesis. Carnegie Mellon University, Apr. 1976.

[43] B. T. Lowerre and R. Reddy. “The HARPY Speech Understanding System”. In: *Trends in Speech Recognition*. Ed. by W. A. Lea. Prentice-Hall Signal Processing Series. Prentice-Hall, 1980, pp. 340–360.

[44] O. Mersmann et al. *mco: Multiple Criteria Optimization Algorithms and Related Functions*. R package version 1.0-15.1. 2014. URL: `https://cran.r-project.org/package=mco`.

[45] K. E. Muller and M. C. Mok. “The distribution of Cook’s D statistic”. In: *Communications in Statistics—Theory and Methods* 26.3 (1997), pp. 525–546.

[46] P. K. Novak, N. Lavrač, and G. I. Webb. “Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining”. In: *Journal of Machine Learning Research* 10 (2009), pp. 377–403.

[47] V. Pachón et al. “Multi-objective Evolutionary Approach for Subgroup Discovery”. In: *Hybrid Artificial Intelligent Systems: 6th International Conference, HAIS 2011, Wroclaw, Poland, May 23-25, 2011, Proceedings, Part II*. Ed. by E. Corchado, M. Kurzyński, and M. Woźniak. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 271–278.

[48] R Core Team. *R: A Language and Environment for Statistical Computing*. Version 3.2.3. R Foundation for Statistical Computing. Vienna, Austria, 2015. URL: `https://www.r-project.org/`.

[49] C. R. Rao. *Linear Statistical Inference and Its Applications*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1965.

[50] S. Ruggles et al. *Integrated Public Use Microdata Series*. Version 6.0 [Machine-readable database]. 2015.