A Goemans-Williamson Type Algorithm for Identifying Subcohorts in Clinical Trials

Pratik Worah
Google Research

Abstract

We design an efficient algorithm that outputs tests for identifying predominantly homogeneous subcohorts of patients from large in-homogeneous datasets. Our theoretical contribution is a rounding technique, similar to that of Goemans and Wiliamson (1995), which approximates the optimal solution within a factor of 0.82. As an application, we use our algorithm to trade-off sensitivity for specificity to systematically identify clinically interesting homogeneous subcohorts of patients in the RNA microarray data set for breast cancer from Curtis et al. (2012). One identified subcohort suggests a link between LXR over-expression and BRCA2 and MSH6 methylation levels for patients in that subcohort.

1 INTRODUCTION

A linear separator in $\mathbb{R}^d$ is a basic tool used in statistics and machine learning to partition a set of points into two subsets. Given a set of Red and Blue points to be separated, there may not exist a linear separator that separates most of the Red and Blue points. Moreover, even if one exists, then it can be expected to have most coefficients to be far from zero, i.e., it is not sparse, unless there are structural assumptions on the dataset. On the other hand, if we want a linear separator that separates a small (but non-trivial) percentage of Red points from Blue points, then a sparse linear separator may exist. Such sparse separators also have the advantage that they are likely to be interpretable. Therefore, the focus shifts to identifying a subset of Red points to carve out from the rest, so that: (1) the number of Blue points in the identified subset is relatively small (specificity is high); (2) the relative number of Red points in the subset, compared to the total number of Red points, is large (sensitivity is high); and (3) the number of non-zero coefficients in the separator is small (the separator is sparse).

In this paper, we devise a semidefinite programming (SDP) based algorithm (see Algorithm 1) for computing sparse linear separators that identify homogeneous subgroups. We prove theoretical guarantees on the specificity and sensitivity of the separator (see Theorems 3.2 and 3.3). We apply our algorithm to design a test to identify subgroups of metastatic breast cancer cases from a large breast cancer data set (Curtis et al., 2012; Pereira et al., 2016; Rueda et al., 2019). We empirically study the trade-offs between the sensitivity, specificity and sparsity of the tests output by our algorithm in Section 4.1. Finally, in Section 4.2, we use our algorithm to identify clinically interesting subgroups of breast cancer patients.

1.1 Related literature

Perhaps the most well-known subgroup identification algorithm is the Patient Rule Induction Method (PRIM) Friedman and Fisher (1999) and its variants Fürnkranz et al. (2012); Ott and Hapfelmeier (2017). These are local search algorithms with heuristics designed to gradually trade-off sensitivity for specificity. At its crux, the better-known PRIM algorithm relies on identifying one or more hyper-rectangles, where points in the interior of a hyper-rectangle constitute the identified subgroup. While such bump hunting heuristics are hard to analyze theoretically, explainable clustering algorithms using axis-aligned thresholding cuts have been theoretically analyzed (for e.g. Sieling (2008), Moshkovitz et al. (2020) and Gupta et al. (2023)). There are at least two significant differences between algorithms using axis aligned hyper-rectangles (equivalent to thresholding rules on coordinates) and our approach. First, such rule-based algorithms are almost always restricted to axis-aligned thresholding cuts, as opposed to sparse cuts (as in

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

this paper); even though the latter are also intuitively interpretable. Second, each axis aligned half-space can also be thought of as node in a decision tree, so computing subgroups that are identified by one or more axis-aligned hyper-rectangles corresponds to computing a decision tree, where the underlying objective is chosen to balance sensitivity and specificity. It is well-known that computing optimal decision trees for most reasonable objectives is not just a NP-hard problem, but most versions are even hard to approximate within any constant factor (see for e.g. Gupta et al. (2010); Koch et al. (2024)). Intuitively, this is because of close connections between decision trees and circuit complexity. In contrast, in this paper, we translate subgroup identification with sparse linear cuts to a MAX-CUT like problem (cf. Goemans and Williamson (1995)), which admits constant factor approximation guarantees via convex optimization (an approximation factor of 0.82 in our case). Moreover, there are practical examples where a sparse linear hyperplane, as opposed to a rule based algorithm, is used to identify interesting subcohorts. For example, the papers Ratziu et al. (2006) and Kim et al. (2021) describe diagnostic scores, essentially sparse regression models, that are widely used in gastroenterology.

2 OVERVIEW OF RESULTS

Motivation: Suppose we are given a set of patients in different disease states. For example, in case of cancer patients, some of them may have metastasis but others may not. We want to identify a small set of genes or proteins (or some other markers) that allows us to test and select patients that are in the same state, i.e., we want to design tests that identify predominantly homogeneous subcohorts (subsets) of breast cancer patients with metastasis. For example, if roughly 10% of the cancer patients have metastasis, then we want to design an algorithm that uses a small number of gene expressions to identify a large enough subcohort, say 2% of the overall patient population, of which 70% have metastasis.

Clearly, there are trade-offs involved. If we want the identified subcohort of patients to be a larger fraction of the overall patient population, i.e., if we want the *sensitivity* to be large, then the probability that most patients in the subcohort will have metastasis, will be smaller, i.e., the *specificity* will be smaller (equivalently the subcohort will be more inhomogeneous). Furthermore, if we want the set of identifying markers to be small, i.e., if we want the designed test to be *sparse*, then the specificity is going to be lower (assuming sensitivity is kept constant). On the other hand, if we want to have sparse tests and high specificity in our subcohort, then the sensitivity (which corresponds to subcohort size) will be reduced.

Contribution: In this paper, we formalize and solve the subcohort identification problem using Algorithm 1,which is a generalization of the Goemans-Williamson semidefinite programming (SDP) algorithm for solving MAX-CUT (Goemans and Williamson, 1995). Our theoretical contribution can be summarized as follows:

- We prove that the ratio of the size of subcohorts that are identified by Algorithm 1 vs. the optimal subcohorts is lower bounded by an absolute constant $\alpha \simeq 0.82$ (see Section 3 for details). This implies bounds on the sensitivity and specificity as well (Theorems 3.2 and 3.3).

The theoretical result should be independently interesting, as it generalizes the Goemans-Williamson calculations. While similar SDP rounding algorithms for similar problems have been analyzed before (for e.g. Feige and Goemans (1995); Lewin et al. (2002) and Brakensiek et al. (2020)), we are not aware of any similar analysis for the problem in our paper.[1]

Furthermore, we empirically study the trade-offs mentioned above, for Algorithm 1, in the context of identifying subcohorts of metastatic breast cancer using the METABRIC dataset (Curtis et al., 2012; Pereira et al., 2016; Rueda et al., 2019).[2]

For the empirical part, our first result identifies subcohorts of predominantly metastatic breast cancer cases using sparse linear tests based on expression of nuclear receptors, i.e., we want the ratio of the metastatic cases to all breast cancer cases to be high for the identified subcohort. We denote this ratio as the *specificity* of our test. The number of metastatic cases in the identified subcohort as a percentage of all breast cancer cases is denoted as the *sensitivity* of the test. The number of genes used is roughly proportional to the sparsity parameter of Algorithm 1. Our first empirical result relates the specificity, sensitivity

[1] In a nutshell, the crux of the Goemans-Williamson guarantee relies on computing the maximum ratio of the chord length to the minor arc length in a unit circle, while in our case it hinges on bounding the ratio of the "perimeter" of a triangle to the perimeter of the circumcircle, the latter is assumed to be of unit radius. It is surprising that a Goemans-Williamson SDP rounding works here, as our objective function is not sign definite. Usually more complicated rounding techniques, like Alon and Naor (2006) are required in such cases.

[2] The dataset contains RNA microarray data of approximately 2,000 breast cancer patients, and methylation data for approximately 13K genes in about 1,500 breast cancer patients. About a thousand patients had one or more positive lymph nodes upon biopsy (metastasis), while the remaining did not (non-metastasis).

and the number of genes used, as follows:

- We show that specificity and sensitivity both increase with the number of genes used in the test (see Figure 1(a) and Subsection 4.1 for details). In particular, we obtained subcohorts of size $\sim 20\%$ of the subsampled patient population, with about 57% specificity, when we use roughly 15 out of 48 nuclear receptors in the designed test.

- Moreover, if we define our subcohort as an intersection of half-spaces from two different runs of Algorithm 1 (over two subsamples of the dataset), then we can increase the specificity by sacrificing on sensitivity (see Figure 1(b)). In particular, we obtained subcohorts of size $\sim 7\%$ of the subsampled patient population, with about 67% specificity, when we use roughly 45 out of 48 nuclear receptors in the test.

Our second empirical result, is an application of Algorithm 1 to identify clinically interesting large homogeneous subcohorts of patients. One example of a clinically interesting problem is to identify statistically significant associations between changes in methylation percentages and changes in nuclear receptor expressions. Methylation of tumor suppressor genes is known to be a potential cause of progression of breast cancers Shiovitz and Korde (2015), while nuclear receptors are often used as targets of many drugs Burris et al. (2013). Therefore, we are attempting to associate the likely cause of the pathology to a likely treatment target, for a specific subcohort of patients.

The question arises: why is identifying such associations difficult or novel, after all, it involves only standard deviation computation? A naive average over the entire cohort of patient population, where one indiscriminately averages over non-homogeneous data, may make confidence intervals larger than the separation between the means. Thus deviations from the mean are no longer statistically significant. This is exactly what happens in Figure 2, where we compute the average expression of nuclear receptors and methylation percentages in metastatic and non-metastatic breast cancer patients, over the entire cohort of about 2,000 patients. Based on the figure, one can not differentiate between metastatic vs non-metastatic breast cancer using deviations and simple Z-scores of any of the 90 odd genes plotted – the population is inhomogeneous – confidence intervals are too wide to differentiate between the two classes of patients with statistical significance. Similarly, if one computes averages over many small subsets of the population, across subcohorts, then the confidence intervals are likely to be too large due to small sample size. Therefore, to deduce clinically interesting conclusions, we need to identify and compute the deviations over homogeneous subcohorts.

In Subsection 4.2, we trade-off sensitivity for specificity using Algorithm 1 to make the identified subcohorts more homogeneous. One clinically interesting subcohort is shown in Figure 3. For this subcohort, we observe:

- There is significant methylation increase in genes BRCA2 and MSH6, and nuclear receptor NR1H2 (equivalently LXRB) is significantly over-expressed, both relative to the baseline population. Therefore, one may hypothesize that methylation of the former leads to downstream over-expression of the latter.

- Furthermore, significant over-expression of LXRB receptor gene suggests that patients in the subcohort may be good candidates for LXR-inverse agonists, if supported by clinical evidence.

  Here, it is worthwhile to note that both LXR-inverse agonists (Carpenter et al., 2019) and LXR agonists (Hutchinson et al., 2019) have been proposed to be beneficial for various subsets of breast cancer patients. Our algorithm, and its output test, may offer a principled way to identify genomic pathways underlying the disease. Moreover, it may also help identify patients, who could benefit from use of LXR-inverse agonists or LXR agonists (or neither).

Currently, clinically homogeneous subgroups of breast cancer patients are often identified by single genes. For example, ER+ patients, i.e., those breast cancer patients with ER nuclear receptor expression above a certain threshold, are treated with a certain protocol and vice versa (Mestres et al., 2016). This is a likely suboptimal determination of patient subcohorts and resulting treatment protocols. In contrast, our algorithm presents a systematic approach with a potentially better handle on sensitivity, specificity and sparsity trade-offs.

Organization: We begin with formal algorithmic description and theoretical details in Section 3, which is followed by empirical results on the breast cancer dataset in Section 4.

3 UNDERLYING THEORETICAL PROBLEM

We begin with an abstract description of our problem. We are given a complete bipartite graph $G \equiv (U, V)$, where the vertices are embedded in a high dimensional

space $\mathbb{R}^d$. Our goal is to compute a hyperplane (h, θ), i.e., $\langle h, x\rangle \leq -\theta$,[3] with few non-zero coefficients in h, that separates a subset $S \equiv (S_U, S_V) \subset U \cup V$ of vertices from the rest of the vertices in G. In addition, we want S_U to be much larger than S_V, and S to be large.

It is easy to see that the computed subset S satisfying the above requirements is precisely the subcohort meeting the requirements from the previous discussion.

Without loss of generality, we may assume that the input also specifies a vertex $u_0 \in S_U$, since one can always iterate over all vertices in U.[4] Note that, S is large and $|S_U| \gg |S_V|$, together imply that the cut $(S_U, V \setminus S)$ is large. Therefore, one may formulate the above as a MAX-CUT type integer program, where one associates a ± 1 variable z_u with each of the n vertices in G, and wants to maximize the objective:

$$\max_{\substack{h\in\mathbb{R}^d,\\ \vec{z}\in\{\pm 1\}^n}} \sum_{\substack{u\in U,\\ u'\in V}} \frac{(z_u - z_{u'})^2 + (z_{u_0} - z_{u'})^2 - (z_u - z_{u_0})^2}{3}. \tag{1}$$

Maximizing the size of such a cut is equivalent to maximizing the size of S, while making $|S_U|$ larger than $|S_V|$. Without any further constraints, the trivial solution $S_U \equiv U, S_V \equiv V$ is optimum. However, the separation and sparsity constraints make the optimum non-trivial. These constraints may be enumerated as follows.

- Separating edges between $U \setminus \{u_0\}$ and V: These edges should only contribute to the determination of the separating hyperplane if $u \in U \setminus \{u_0\}$ is not cut, i.e., u and u_0 are on the same side (u is in the subcohort). Hence, for every such $u \in U \setminus \{u_0\}$, we add the constraint:

$$\langle h, x_u\rangle \leq -\theta z_u z_{u_0} \tag{2}$$

- Separating edges between u_0 and V: These edges should only contribute to the determination of the separating hyperplane if $u' \in V$ is cut, i.e., u' and u_0 are on different sides (u' is not in the subcohort). Hence, for every such $u' \in V$, we add the constraint:

$$\langle h, x_{u'}\rangle \geq \theta(1 - z_{u'} z_{u_0}) \tag{3}$$

- Sparsity: We want the designed test to use few genes, i.e., we want its coefficients to be sparse. Hence we bound the ℓ_1 norm of its coefficients (see for example Foucart and Rauhut (2013)) by an input parameter s, which is expressed as:

$$\|h\|_1 \leq s. \tag{4}$$

Remark 3.1 *It is worth noting that the formulated problem above is very similar to maximum margin classification using linear support vector machines. Thus, it is possible to generalize it to non-classifiers using the "kernel trick" (see for example Boser et al. (1992)). The only exception being the sparsity constraint above, which needs to be adapted for the chosen kernel. It may be possible to generalize the sparsity constraint for low degree polynomial kernels using standard compressed sensing methods (see for example Foucart and Rauhut (2013)). However, such non-linear extensions are not the focus of this work.*

One may relax the discrete objective function and the linear constraints on the embedding into a SDP, as specified below. Note that the SDP relaxation can be efficiently solved using a standard solver unlike the discrete objective function.

$$\begin{aligned}
&\max_{\substack{h\in\mathbb{R}^d,\\ \forall i:\ \|v_i\|_2=1}} && \textstyle\sum_{\substack{u\in U,\\ u'\in V}} \frac{2-2\langle v_u, v_{u'}\rangle - 2\langle v_{u_0}, v_{u'}\rangle + 2\langle v_u, v_{u_0}\rangle}{3},\\
s.t.\quad &\forall u\in U: && \langle h, x_u\rangle \leq -\theta\langle v_u, v_{u_0}\rangle\\
&\forall u'\in V: && \langle h, x_{u'}\rangle \geq \theta(1 - \langle v_{u'}, v_{u_0}\rangle)\\
& && \|h\|_1 \leq s,
\end{aligned}$$

where x_u are the embeddings in $\mathbb{R}^d$, s is the sparsity parameter, and θ is the separation threshold parameter, all to be specified with the input. Algorithm 1 essentially maps the vector solution of the SDP to ± 1 values, to obtain the subcohort (corresponding to $\{i \in U \cup V :\ z_i = 1\}$).

In order to show that the sensitivity and specificity of the test output by the SDP solution above are not too far from the optimum, we prove the following approximation guarantee (see the appendix, at the end, for proofs).

Theorem 3.2 (Sensitivity) *Let $S^* = (S^*_U, S^*_V)$ denote the optimal subcohort such that $|S^*_U| \ll |U|$, then Algorithm 1 computes a subcohort $S = (S_U, S_V)$, such that $|S_U| \geq \alpha|S^*_U|$, where the constant $\alpha \simeq 0.82$. Equivalently, $\frac{|S_U|}{|U|+|V|} \geq \alpha \cdot \frac{|S^*_U|}{|U|+|V|}$, i.e., the sensitivity of the output test is within 0.82 of the optimum.*

Theorem 3.3 (Specificity) *For $\kappa \in (0, 1)$, let $\kappa \cdot |U||V|$ denote the fraction of edges in the optimal*

[3] Here $x \in \mathbb{R}^d$ is the input specified embedding of a given vertex in G into $\mathbb{R}^d$.

[4] In the experiments in this paper, we used a uniformly random choice of u_0. Changing u_0 did not change the sensitivity and specificity significantly. The standard deviation of the specificity and sensitivity obtained by picking different random u_0 was much smaller than their means.

Algorithm 1 Identify subcohort

Input: Two sets of vertices (U, V) labeled 0 and 1 respectively, embedded in $\mathbb{R}^d$; vertex $u_0 \in U$, sparsity parameter $s \in (0, \infty)$, and separation threshold $\theta \in \mathbb{R}$.
Output: If feasible, compute a hyperplane $h \in \mathbb{R}^d$ and a subcohort $S \equiv (S_U, S_V) \subset U \cup V$ such that: (1) $u_0 \in S_U$, (2) $\forall u \in S_U : \langle x_u, h\rangle \leq \theta$ and $\forall v \in S_V : \langle x_v, h\rangle \geq 0$, (3) $\|h\|_1 \leq s$, and (4) $|S_U| \gg |S_V|$. Note: Here x_u and x_v are the $\mathbb{R}^d$ embeddings corresponding to vertices u and v (from the input) and $\|h\|_1$ denotes the ℓ_1 norm of the coefficients of h.

1: Solve the SDP relaxation (Equation 5) to obtain the solution vectors $\{v_i\}_n$.
2: If SDP infeasible, **return** infeasible.
3: $\forall i \in \{1, ..., n\} : \; z_i \leftarrow \text{sign}(\langle v_i, \gamma_i\rangle)$, where $\gamma_i \sim \mathcal{N}(0, 1)$.
4: **return** $\vec{z}, h$

sparse cut separating U from V. The specificity of the test output by Algorithm 1 is lower bounded by the solution to the following optimization problem:

$$\min_{x,y\in(0,1)} \frac{x \cdot |U|}{(1-y)|V| + x|U|} \qquad \text{s.t.} \quad x \cdot y \geq \tilde{\alpha} \cdot \kappa, \tag{5}$$

where $\tilde{\alpha} = \frac{\alpha}{2} \simeq 0.41$.

Note that the solution of the optimization problem in Equation 5 grows as $\Omega(\sqrt{\kappa})$, when $|U|$ is assumed equal to $|V|$ (this can be seen by substituting $y = \frac{\tilde{\alpha}\cdot\kappa}{x}$ in the objective). A plot of the solution surface, together with the proof, is given in the appendix (Section B).

While the optimum hyperplane separator h remains unchanged after rounding though the input specified threshold θ is perturbed during rounding. However, we may bound the perturbation amount by using Grothendieck's identity, i.e.,

$$\mathbb{E}\left[\text{sign}(z_u)\text{sign}(z_{u_0})\right] = \frac{2}{\pi}\arcsin\left(\langle v_u, v_{u_0}\rangle\right). \tag{6}$$

as follows.

Theorem 3.4 *The rounded solution from Algorithm 1 obeys the (slightly weaker) separation constraints:*

$$\langle h, x_u\rangle \leq -\frac{2\theta}{\pi} z_u z_{u_0} \tag{7}$$

$$\langle h, x_{u'}\rangle \geq \theta\left(1 - \frac{2}{\pi}(z_{u'} z_{u_0})\right) \tag{8}$$

as opposed to the corresponding constraints in Equations 2 and 3.

4 GENOMICS APPLICATIONS

In this subsection, we use Algorithm 1 to generate tests[5] that identifies subcohorts with predominantly metastatic cases in a breast cancer dataset. Our empirical results may be divided into two parts:

- In Subsection 4.1, we examine the quality of the identified subcohorts. A random test, i.e., a random half-space, can be expected to generate subcohorts with less than 50% metastasis cases (as the number of non-metastasis patients is higher in the dataset). In Figure 1(a), we show that the linear tests generated by Algorithm 1 identify subcohorts where greater than 50% of the cases are metastatic (thus outperforming random). Moreover, if instead of using one hyperplane to identify a subcohort, we use two hyperplanes to identify a subcohort then the identified subcohorts have 60-70% of their cases as metastatic (see Figure 1(b)). We also compare our algorithm's specificity (for a given sensitivity) with a version of the well-known PRIM algorithm for subgroup discovery (this comparison is in the Appendix Section D).
- In Subsection 4.2, we use Algorithm 1 to identify a clinically interesting subcohort of breast cancer patients – those that may respond to LXR inverse agonists. In the process, we also associate epigenetic changes with changes in nuclear receptor expressions. It is known Shiovitz and Korde (2015), that about 40 genes act as tumor suppressors, and their methylation leads to reduced expression and potentially cancer progression. Moreover, in Burris et al. (2013); Overington et al. (2006), the authors discuss the role of nuclear receptors as drug targets. Therefore, it is natural to look for associations between significant changes in nuclear receptor expressions and significant changes in methylation levels of breast cancer related genes (over population baseline levels) – the latter can guide treatment drug targets from the former. We show that, even though at the entire dataset level such associations do not exist, but at the subcohort level they do exist (see Figure 3).

Dataset: The results in this section are based on the METABRIC dataset (Curtis et al., 2012; Pereira et al., 2016; Rueda et al., 2019), which contains RNA microarray data from biopsy samples of approximately 1904 (911 with positive lymph nodes and 993 without positive lymph nodes) breast cancer patients, and

[5] A test is just the output hyperplane of Algorithm 1 – a weighted average of gene expression levels – that evaluates to a score less than the given threshold θ for metastatic cases, and above θ for non-metastatic cases.

methylation data for approximately 13K genes in about 1406 (669 with positive lymph nodes and 737 without positive lymph nodes). In this paper, we (crudely) denote patients with one or more positive lymph nodes upon biopsy as metastasis cases, while the rest as non-metastasis cases.

4.1 Predicting breast cancer metastasis

Recall that, the *sensitivity* of Algorithm 1 is the percentage of metastatic cases that are correctly classified to be in the subcohort, while the *specificity* is the percentage of metastatic cases in the identified subcohort. The number of genes used by the test is defined as the number of coefficients of h whose magnitude is at least $\frac{1}{10}$ of the maximum magnitude coefficient in h. Due to well known results from ℓ_1 norm minimization Foucart and Rauhut (2013), the number of genes used should be roughly proportional to the sparsity parameter input to the algorithm (Equation 4).

Figure 1 shows the sensitivity, specificity and sparsity trade-offs for our subcohort identification algorithm. To generate Figure 1, Algorithm 1 uses the nuclear receptor expressions of a subset of 50 metastatic and 50 non-metastatic cases from the full dataset, the sparsity parameter s, a threshold θ, and an arbitrary metastasis patient that is assumed to belong to the subcohort to be identified. The algorithm outputs a test h, where h is just a tuple of weights for the nuclear receptors. On a separate hold-out set of a 1000 patients (500 with metastasis and 500 without), the test simply calculates the weighted average value of the given nuclear receptor expression levels for each patient in the hold-out, and if this value is below the threshold then the patient is added to the subcohort, otherwise not. The above exercise is repeated 10 times for various sparsity values, and sensitivity and specificity measured each time, to arrive at Figure 1.

Interpretation of Figure 1: First, as the number of genes increases, so do specificity and sensitivity in Figure 1. Second, sensitivity can be traded-off for specificity: We run Algorithm 1 twice, on two different subsampled sets of patients, and obtain two half-spaces/tests h_1, h_2 that correspond to two subcohorts. The new subcohort is defined to be the intersection of the previously identified half-spaces, i.e., both tests h_1, h_2 have to place patients in the subcohort for the patient to be in the new subcohort. Clearly, specificity should increase for this more conservatively defined subcohort and sensitivity should decrease. Indeed that is the case – specificity increases to 70% while sensitivity decreases by a factor of 3 (see Figure 1(b)). One may even combine three or more tests in this way. Thus one can use Algorithm 1 to increase the specificity of the subcohort at the cost of decreased sensitivity, i.e., decreased size of the identified subcohort.

Comparison: We compare our algorithm with a version of the well-known PRIM algorithm in Appendix (Section D). It is clear from our experiments that the specificity of the MAX-CUT based approach outperforms that of the local search based heuristic when sensitivity is held constant. This is despite the fact that the local search heuristic was allowed an order of magnitude more time for computations.

4.2 Methylation and gene expression

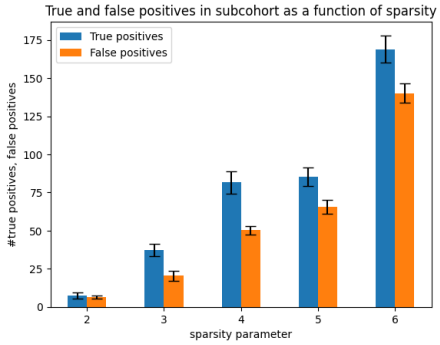
Algorithm 2 Identify Critical Genes

Input: Dataset $D \subset \mathbb{R}^d$, and parameter $\epsilon \in (0, 1)$ (the d coordinates are the genes)
Output: A subcohort with critical genes that are significantly differently expressed from the population

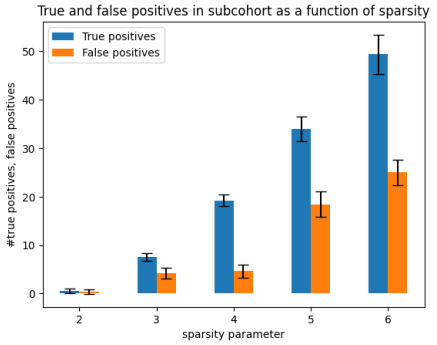
1: Uniformly sample $S \subset D$ with $|S| \simeq \epsilon \cdot |D|$
2: Use Algorithm 1 on S to compute a half-space h corresponding to a likely homogeneous subcohort.
3: Compute means and standard deviation of each of d coordinates in population and subcohort.
4: **If** any coordinate has mean which is one standard deviation away from population mean, then **return** the coordinate (gene), subcohort and h (test)
5: **Else** generate a new sample (Step 1) and repeat

One application of Algorithm 1 is to identify genes which differ significantly between a subcohort and the population. Algorithm 2 presents the basic idea. We will use a slight variation of it to relate epigenetic changes (nucleotide methylation changes) to clinically significant genetic changes downstream – an important question in genomics (see for example Cavalli and Heard (2019)). The use of subsets is critical as averaging over the entire dataset leads to large confidence intervals.

Background: Methylation of certain tumor suppressor genes can potentially lead to breast cancer progression (Cavalli and Heard, 2019; Shiovitz and Korde, 2015), and nuclear receptors are a target for a large number of drugs Burris et al. (2013); Overington et al. (2006). Therefore, if we compute the mean methylation and expression levels of critical genes over the entire patient population in the dataset to find a tumor suppressor gene that is significantly more methylated and a nuclear receptor that is significantly over-expressed; then we can test whether a drug targeting that over-expressed nuclear receptor is effective in patients that show a similar pattern of excessive

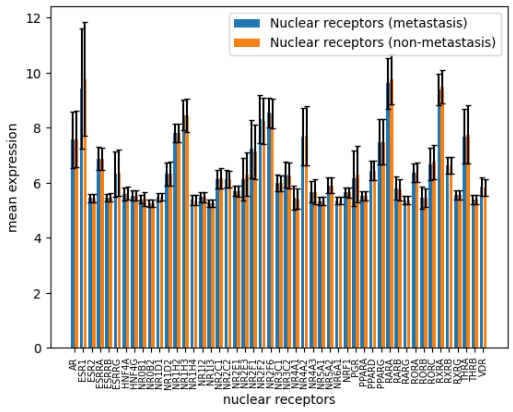


(a) The number of metastasis cases (true positives) and non-metastasis cases (false positives) in the identified subchort using a single test from Algorithm 1. Sensitivity is proportional to the height of the blue bar, while Specificity is proportional to the ratio between the height of blue and red bars. The sparsity parameter (x-axis) is proportional to the number of genes used.

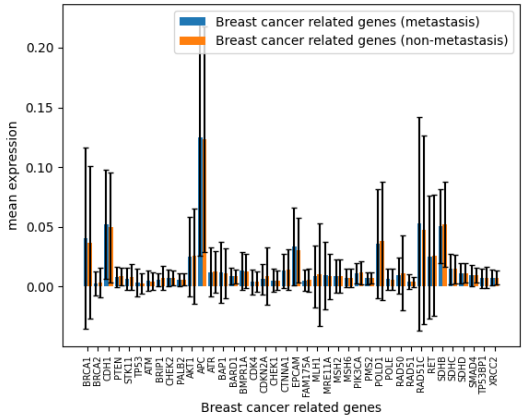


(b) The number of metastasis cases (true positives) and non-metastasis cases (false positives) in the identified subchort using two tests (vs one in L figure), generated using two runs of Algorithm 1. Sensitivity is proportional to the height of the blue bar, while Specificity is proportional to the ratio between the height of blue and red bars. The sparsity parameter is proportional to the number of genes used.

Figure 1: Sensitivity vs Specificity trade-off using Algorithm 1, for the dataset Curtis et al. (2012). Left: Algorithm 1 outputs a single test hyperplane that identifies subcohorts with 57% specificity, for sparsity parameter $s = 6$. Right: When we use the intersection of subcohorts using two tests, the specificity increases to about 70%, but identified subcohort size (sensitivity) decreases to a third, with $s = 6$. Note: the single test in the left figure uses about 15 nuclear receptors to identify subcohorts, while the two tests in the right figure use about 45 nuclear receptors.



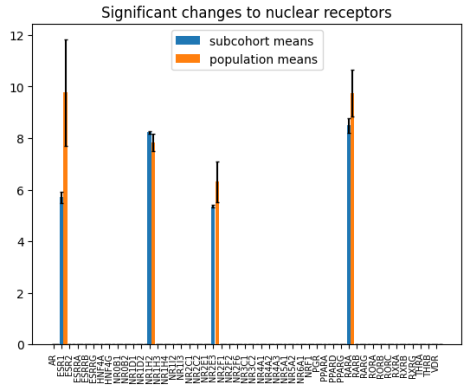
(a) Expressions of nuclear receptors in metastasis and non-metastasis.



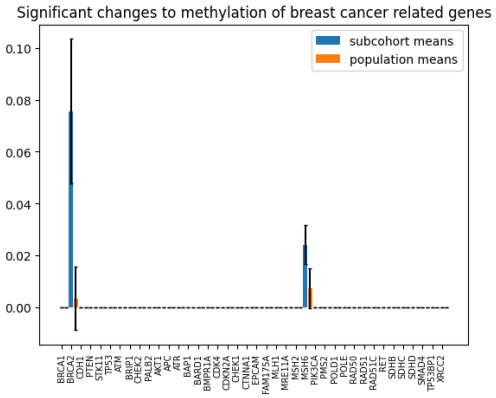
(b) Percentage methylations of breast cancer related genes in metastasis and non-metastasis.

Figure 2: Gene expression and methylation percentages for metastasis and non-metastasis patients computed over the full dataset. Observe that averaging over the entire dataset, consisting of non-homogeneous patient population, leads to large confidence intervals. The latter make it impossible to distinguish between metastasis and non-metastasis using z-scores based on the nuclear receptors and breast cancer related genes used above.

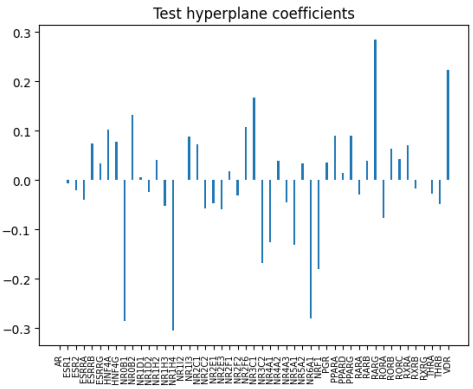
methylation and over-expression. Unfortunately, the situation is not so simple – there are no such sets of tumor suppressors and nuclear receptors where significant change in one is associated with significant change in the other. Figure 2 shows that the confidence intervals are just too large, when we average over the entire dataset (because of inhomogeneity in the disease). However, it is possible that the disease state amongst the patients in some subcohorts is homogeneous enough to allow for narrow confidence intervals, and thus identify any simultaneous significant deviation in tumor suppressor methylation and nuclear



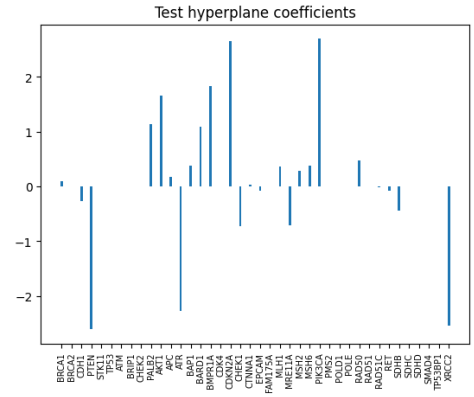
(a) The nuclear receptors in the identified subcohort, consisting mainly of metastasis cases, that are predicted to differ significantly in expression from their mean population values. In particular, for patients in this subcohort, NR1H2 is over-expressed. Statistically insignificant bars are not shown.



(b) The methylation percentages in the identified subcohort, consisting mainly of metastasis cases, that are predicted to differ significantly in expression from their mean population values. In particular, for patients in this subcohort, BRCA2 and MSH6 were methylated more, both are related to DNA repair, and errors can lead to breast cancer progression Song et al. (2019). Statistically insignificant bars are not shown.



(c) The coefficient weights of the genes in the test output by Algorithm 1 to identify the subcohort above.



(d) The coefficient weights of the genes in the test output by Algorithm 1 to identify the subcohort above.

Figure 3: Associating changes in methylation percentages with changes in nuclear receptors using Algorithm 1. Compared to averaging over the entire population (see Figure 2), averaging over patients in a subcohort can lead to statistically significant differences between metastasis and non-metastasis breast cancers. In particular, in the above subcohort of patients, BRCA2 and MSH6 have significantly higher methylation percentages than the entire cohort of the patients in METABRIC dataset; and at the same time they have a significantly higher expression of NR1H2 receptors responsible for lipid metabolism.

receptor expression, amongst patients restricted to such a subcohort. Indeed that is the case below.

Algorithm 2 for relating epigenetics with nuclear receptors: We use a variant of Algorithm 2, with about 1000 different random subsamples of inputs (of 100 patients each), from the original dataset (of size 2000 patients) to generate patient subcohorts of breast cancer patients (and corresponding tests to identify the subcohorts) in two ways: (1) based on methylation percentage levels, and (2) based on nuclear receptor expression levels.

Next, for every pair of subcohorts generated (one from methylation levels and another from nuclear receptor expressions), if the two subcohorts have more than a certain percentage of patients in common (we used a threshold of 10% due to limited dataset size) i.e., they share a relatively large subset of the patient population, then *they are likely to be homogeneous as two different sets of markers – genetic and epigenetic – are identifying a similar subset of patients.*

Next, we compute the mean methylation and expression together with confidence intervals, for each gene using the common patients. Finally, we iterate until we find a patient subcohort with genes that have large deviations from the population means. One such subcohort is shown in Figure 3.

Interpretation of Figure 3: Patients in the subcohort are identified by sparse linear tests, where the coordinate weights (coefficients) for the test hyperplane are given in Figure 3(c,d). Interestingly, this particular subcohort of patients shows significant methylation increase in BRCA2 and MSH6. Moreover, nuclear receptor NR1H2 (equivalently LXRB) is significantly over-expressed when compared to the average non-metastatic patient in the dataset. Therefore, LXR-inverse agonists (see for example Carpenter et al. (2019)) may prove to be a useful therapeutic target for this subcohort of patients, if indicated by further clinical evidence.

5 CONCLUSION

In conclusion, we have demonstrated a method to design tests that identify homogeneous subgroups (subcohorts) of patients. This was done by translating the problem to a MAX-CUT like formulation (as opposed to a rule-based decision tree like formulation), which admits a nearly optimal efficient approximation algorithm via an extension of the Goemans-Williamson algorithm (Goemans and Williamson, 1995). The latter may be independently interesting. Furthermore, the identified subcohorts and accompanying tests could help broaden the set of patients for existing treatments (for e.g. breast cancer patients and LXR inverse agnosts). However, we do not have the resources to verify our clinical predictions and we hope our algorithm (or similar idea) can be adapted by oncologists to pursue it further.

References

Noga Alon and Assaf Naor. Approximating the cut-norm via grothendieck's inequality. *SIAM Journal on Computing*, 35(4):787–803, 2006.

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, page 144–152, New York, NY, USA, 1992. Association for Computing Machinery. doi: 10.1145/130385.130401. URL `https://doi.org/10.1145/130385.130401`.

Joshua Brakensiek, Neng Huang, Aaron Potechin, and Uri Zwick. On the mysteries of MAX NAE-SAT. *CoRR*, abs/2009.10677, 2020. URL `https://arxiv.org/abs/2009.10677`.

Thomas P Burris, Laura A Solt, Yongjun Wang, Christine Crumbley, Subhashis Banerjee, Kristine Griffett, Thomas Lundasen, Travis Hughes, and Douglas J Kojetin. Nuclear receptors and their selective pharmacologic modulators. *Pharmacological Reviews*, 65:710–778, 2013.

Katherine J Carpenter, Aurore-Cecile Valfort, Nick Steinauer, Arindam Chatterjee, Suomia Abuirqeba, Shabnam Majidi, Monideepa Sengupta, Richard J Di Paolo, Laurie P Shornick, Jinsong Zhang, and Colin A Flaveny. LXR-inverse agonism stimulates immune-mediated tumor destruction by enhancing CD8 T-cell activity in triple negative breast cancer. *Scientific Reports*, 9, 2019.

Giacomo Cavalli and Edith Heard. Advances in epigenetics link genetics to the environment and disease . *Nature*, 571, 2019.

Christina Curtis, Sohrab P. Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, Andy G. Lynch, Shamith Samarajiwa, Yinyin Yuan, Stefan Gr af, Gavin Ha, Gholamreza Haffari, Ali Bashashati, Roslin Russell, Steven McKinney, METABRIC Group, Anita Langerød, Andrew Green, Elena Provenzano, Gordon Wishart, Sarah Pinder, Peter Watson, Florian Markowetz, Leigh Murphy, Ian Ellis, Arnie Purushotham, Anne-Lise Børresen-Dale, James D. Brenton, Simon Tavaré, Carlos Caldas, and Samuel Aparicio. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486:346–352, 2012.

Uriel Feige and Michel Goemans. Approximating the value of two power proof systems, with applications to max 2sat and max dicut. In *Proceedings Third Israel Symposium on the Theory of Computing and Systems*, pages 182–189, 1995.

Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Birkhäuser New York, NY, 2013.

Jerome H. Friedman and Nicholas I. Fisher. Bump hunting in high-dimensional data. *Statistics and computing*, 1999.

Johannes Fürnkranz, Dragan Gamberger, and Nada Lavrač. *Foundations of Rule Learning*. Springer Berlin, Heidelberg, 2012.

Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM*, 42(6):1115–1145, November 1995. ISSN 0004-5411. doi: 10.1145/227683.227684. URL `https://doi.org/10.1145/227683.227684`.

Anupam Gupta, Viswanath Nagarajan, and R. Ravi. Approximation algorithms for optimal decision trees and adaptive tsp problems. In Samson Abramsky, Cyril Gavoille, Claude Kirchner, Friedhelm Meyer auf der Heide, and Paul G. Spirakis, editors, *Automata, Languages and Programming*, pages 690–701, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

Anupam Gupta, Madhusudhan Reddy Pittu, Ola Svensson, and Rachel Yuan. The Price of Explainability for Clustering. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1131–1148, Los Alamitos, CA, USA, 2023. IEEE Computer Society.

Samantha A Hutchinson, Priscilia Lianto, Hanne Roberg-Larsen, Sebastiano Battaglia, Thomas A Hughes, , and James L Thorne. ER-Negative Breast Cancer Is Highly Responsive to Cholesterol Metabolite Signalling. *Nutrients*, 11, 2019.

W Ray Kim, Ajitha Mannalithara, Julie K Heimbach, Patrick S Kamath, Sumeet K Asrani, Scott W Biggins, Nicholas L Wood, Sommer E Gentry, and Allison J Kwong. MELD 3.0: The Model for End-stage Liver Disease Updated for the Modern Era. *Gastroenterology*, 2021.

Caleb Koch, Daniel Strassle, and Zhixian Tan. Superconstant Inapproximability of Decision Tree Learning. *arXiv preprint arXiv:2407.01402*, 2024.

Michael Lewin, Dror Livnat, and Uri Zwick. mproved rounding techniques for the MAX 2-SAT and MAX DI-CUT problems. *International Conference on Integer Programming and Combinatorial Optimization*, 2002.

J. A. Mestres, A. B. iMolins, L. C. Martínez, J. I. C. López-Muñiz, E. C. Gil, A. de Juan Ferré, S. del Barco Berrón, Y. F. Pérez, J. G. Mata, A. G. Palomo, J. G. Gregori, P. G. Pardo, J. J. I. Mañas,

A. L. Hernández, E. M. de Dueñas, N. M. Jáñez, S. M. Murillo, J. S. Bofill, P. Z. Auñón, and P. Sanchez-Rovira. Defining the optimal sequence for the systemic treatment of metastatic breast cancer. *Clinical and Translational Oncology*, 2016.

Michal Moshkovitz, Sanjoy Dasgupta, Cyrus Rashtchian, and Nave Frost. Explainable k-means and k-medians clustering. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7055–7065. PMLR, 13–18 Jul 2020.

Armin Ott and Alexander Hapfelmeier. Nonparametric Subgroup Identification by PRIM and CART: A Simulation and Application Study. *Computational and Mathematical Methods in Medicine*, 2017.

John P Overington, Bissan Al-Lazikani, and Andrew L Hopkins. How many drug targets are there? *Nature Reviews Drug discovery*, 5, 2006.

Bernard Pereira, Suet-Feung Chin, Oscar M. Rueda, Hans-Kristian Moen Vollan, Elena Provenzano, Helen A. Bardwell, Michelle Pugh, Roslin Russell Linda Jones, Stephen-John Sammut, Dana W. Y. Tsui, Bin Liu, Sarah-Jane Dawson, Jean Abraham, Helen Northen, John F. Peden, Abhik Mukherjee, Gulisa Turashvili, Andrew R. Green, Steve McKinney, Arusha Oloumi, Sohrab Shah, Nitzan Rosenfeld, Leigh Murphy, David R. Bentley, Ian O. Ellis, Arnie Purushotham, Sarah E. Pinder, Anne-Lise Børresen-Dale, Helena M. Earl, Paul D. Pharoah, Mark T. Ross, Samuel Aparicio, and Carlos Caldas. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nature communications*, 7(11479), 2016.

Massard J Ratziu, Charlotte F, Messous D, Imbert-Bismut F, Bonyhay L, Tahiri M, Munteanu M, Thabut D, Cadranel J, Le Bail B, De Ledinghen V, and Poynard T. Diagnostic value of biochemical markers (FibroTest-FibroSURE) for the prediction of liver fibrosis in patients with non-alcoholic fatty liver disease. *BMC Gastroenterology*, 2006.

Oscar M. Rueda, Stephen-John Sammut, Jose A. Seoane, Suet-Feung Chin, Jennifer L. Caswell-Jin, Maurizio Callari, Rajbir Batra, Bernard Pereira, Alejandra Bruna, H. Raza Ali, Elena Provenzano, Bin Liu, Michelle Parisien, Cheryl Gillett, Steven McKinney, Andrew R. Green, Leigh Murphy, Arnie Purushotham, Ian O. Ellis, Paul D. Pharoah, Cristina Rueda, Samuel Aparicio, Carlos Caldas, and Christina Curtis. Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. *Nature*, 567, 2019.

S. Shiovitz and L.A. Korde. Genetics of breast cancer: A topic in evolution. *Annals of Oncology*, 26, 2015.

Detlef Sieling. Minimization of decision trees is hard to approximate. *Journal of Computer and System Sciences*, 74(3):394–403, 2008. ISSN 0022-0000. Computational Complexity 2003.

Yun Song, William T. Barry, Davinia S. Seah, Nadine M. Tung, Judy E. Garber, and Nancy U. Lin. Patterns of recurrence and metastasis in BRCA1/BRCA2-associated breast cancers. *Cancer*, 2019.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

A Proof of Theorem 3.2

Proof: We need to show that if the SDP solution has value OPT then the rounded solution has value at least αOPT, for some constant α (to be determined). Hence it suffices to upper-bound the ratio of each term in the in the SDP solution objective by the corresponding term in the objective for the expectation of the rounded solution. In other words, we need to upper-bound:

$$r := \frac{2 - 2\langle v_u, v_{u'}\rangle - 2\langle v_{u_0}, v_{u'}\rangle + 2\langle v_u, v_{u_0}\rangle}{\mathbb{E}\left[(z_u - z_{u'})^2 + (z_{u_0} - z_{u'})^2 - (z_u - z_{u_0})^2\right]}, \tag{9}$$

where $z_u, z_{u'}$ and z_{u_0} are obtained upon rounding $v_u, v_{u'}$ and v_{u_0} using the Gaussian rounding procedure defined in Algorithm 1. Now imagine the three unit vectors $v_u, v_{u'}$ and v_{u_0} as lying on the unit circle, and let $x \in [0, \pi]$ be the angle between u_0, u and $x' \in [0, \pi]$ be the angle between u_0, u'.

Next, consider each term in the denominator, $\mathbb{E}\left[(z_{u_0} - z_{u'})^2\right]$ equals the probability that the random chosen Gaussian – that is equivalent to choosing a uniformly random point on the unit circle (with high probability) – is at an acute angle with u_0 and an obtuse angle with u' or vice versa. Thus,

$$\mathbb{E}\left[(z_{u_0} - z_{u'})^2\right] = 4 \cdot \frac{x'}{\pi}. \tag{10}$$

Similarly, one has:

$$\mathbb{E}\left[(z_{u_0} - z_u)^2\right] \quad = \quad 4 \cdot \frac{x}{\pi} \tag{11}$$

Two cases arise:

$$x + x' \leq \pi: \ \mathbb{E}\left[(z_{u'} - z_u)^2\right] \quad = \quad 4 \cdot \frac{x + x'}{\pi} \tag{12}$$

$$x + x' \geq \pi: \ \mathbb{E}\left[(z_{u'} - z_u)^2\right] \quad = \quad 4 \cdot \frac{2\pi - x - x'}{\pi}. \tag{13}$$

Furthermore, the numerator equals:

$$2 - 2\langle v_u, v_{u'}\rangle - 2\langle v_{u_0}, v_{u'}\rangle + 2\langle v_u, v_{u_0}\rangle = 2 - 2\cos(\theta) - 2\cos(x') + 2\cos(x), \tag{14}$$

where we have used $\cos(\pi - x) = \cos(x)$ and θ is either $x + x'$ (for $x + x' \leq \pi$), or θ is either $2\pi - x - x'$ (for $x + x' \geq \pi$). Thus the ratio r in Equation 9 equals:[6]

$$x + x' \leq \pi: \ r(x, x') = \frac{\pi}{4} \cdot \frac{2 - 2\cos(x + x') - 2\cos(x') + 2\cos(x)}{2x'} \tag{15}$$

$$x + x' \geq \pi: \ r(x, x') = \frac{\pi}{4} \cdot \frac{2 - 2\cos(2\pi - x - x') - 2\cos(x') + 2\cos(x)}{2\pi - 2x}. \tag{16}$$

Note that whether $x' \to 0$ in Equation 15 or $x \to \pi$ in Equation 16, the denominator goes to 0 but the numerator goes to 0 as well. Moreover, since $\cos(x)$ is even, the numerator goes to 0 as a quadratic in both cases, so the limit is 0. Thus 0 or π are likely not the maximum points.

[6] Just to compare, for the well-known rounding in Goemans and Williamson (1995), one finds the corresponding ratio to be: $r_{GW} = \frac{1-\cos(x)}{x}$, and the maximum is well-known to be at $x \simeq 134$deg.

For $x + x' \leq \pi$: Taking the partial derivative with respect to x, one gets: $\sin(x + x') - \sin(x) = 0$. Thus either (1) $x + x' = x$, so $x' = 0$; or (2) $x + x' = \pi - x$, so $2x = \pi - x'$. However, we have ruled out (1). Substituting (2) into Equation 15, we get (at the extremum):

$$r(x) = \frac{\pi}{4} \cdot \frac{1 - \cos(x + x') - \cos(x') + \cos(x)}{x'} \tag{17}$$

$$= \frac{\pi}{4} \cdot \frac{1 - \cos(\pi/2 + x'/2) + \cos(\pi/2 - x'/2) - \cos(x')}{2x'}. \tag{18}$$

Plotting the single variable function above, shows that the maximum value of r is 1.22 and is reached at $x = 1.86$rads.

For $x + x' \geq \pi$: Taking the partial derivative with respect to x' and equating with 0 shows: $\sin(x') - \sin(x + x' - \pi) = 0$ at the extremum. Thus either (1) $x' = x + x' - \pi$, so $x = \pi$; or (2) $\pi - x' = x + x' - \pi$, so $x' = \pi - x/2$. However, we have ruled out (1). Substituting with (2) into Equation 16, we get (at the extremum):

$$r(x) = \frac{\pi}{4} \cdot \frac{1 - \cos(\pi - x/2) - \cos(\pi - x/2) + \cos(x)}{\pi - x} = \frac{\pi}{4} \cdot \frac{1 - 2\cos(\pi - x/2) + \cos(x)}{\pi - x}. \tag{19}$$

Plotting the single variable function above, shows that the maximum value of r is 1.22 and is reached at $x = 1.28$rads.

Thus the maximum value of r is 1.22, and the rounded solution is within $\frac{1}{1.22} = 0.82$ of the optimal SDP solution (and hence the optimal solution). Finally, note that the ratio of vertex set sizes, i.e., $\frac{|S_U|}{|S_U^*|}$, equals the ratio of cut sizes as long as $|S_U^*| \ll |U|$. Hence the proof follows. □

B Proof of Theorem 3.3

Proof: Without loss of generality assume that the variable z_{u_0} in Equation 1 is $+1$ (if not one can flip all the signs). Next, note that each term of the objective of the optimization problem in Equation 1 is either $+1$ or -1 (for the ± 1 solution). The total value of the cut is assumed to be $\kappa \cdot |U||V|$. Therefore, if x and y denote the fraction of U and V vertices on the subcohort side of the cut, then we have by comparing the leading term in the objective with the cut value above:

$$2x(1 - y) \cdot |U||V| \geq \kappa \cdot |U||V| \tag{20}$$

However, the cut we compute using SDP rounding is not optimal but only close to it, i.e.,

$$2x(1 - y) \cdot |U||V| \geq \alpha\kappa \cdot |U||V|, \tag{21}$$

with $\alpha \simeq 0.82$ (cf. Theorem 3.2).

Finally, the specificity is the ratio of U points on selected side of the cut with all U and V points on the subcohort side of the cut, i.e., specificity is given by the ratio of U vertices over all vertices on the subcohort side of the cut, i.e.,

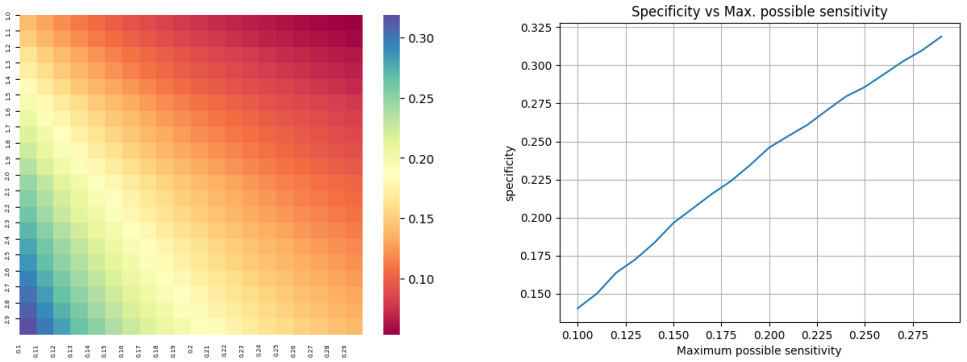
$$\frac{x|U|}{y|V| + x|U|} \tag{22}$$

In order to lower bound the ratio, we simply need to compute its minimum over $x, y \in (0, 1)$ under the constraint in Equation 21. Substituting y with $1 - y$ proves the result. □

One may think of κ as the maximum fraction of edges that can be in the cut given the sparsity constraints. Thus κ is the upper bound on sensitivity, while the value of the optimum in Equation 5 is a lower bound on the sensitivity. Thus we plot the objective in Equation 5 in Figure 4 for various values of κ and $\frac{|V|}{|U|}$, to illustrate the lower bound on specificity.

C Proof of Theorem 3.4

Proof: A straight forward application of Lemma C.1 (also known as Grothendieck's identity) implies that the RHS of the constraints in Equations 5 and 5 equals $-\frac{2\theta}{\pi}$ and $\left(1 - \frac{2}{\pi}(z_{u'} z_{u_0})\right)$ respectively, in expectation. Thus the constraints of our SDP relaxation are not violated excessively. Hence the proof follows. □



(a) Heatmap depicting lower bound on specificity as a function of κ (x-axis) and $\frac{|V|}{|U|}$ (y-axis)

(b) Lowerbound on specificity vs κ (max. possible sensitivity)

Figure 4: Lower bound on specificity as a function of κ and $\frac{|V|}{|U|}$, based on Theorem 3.3.

Lemma C.1 *Given a 2-dimensional Gaussian (X, Y) with mean zero and covariance matrix $\begin{bmatrix} \delta_x & \rho \\ \rho & \delta_y \end{bmatrix}$, we have:*

$$\mathbb{E}[\text{sign}(X)\text{sign}(Y)] = 1 - 2\mathbb{P}\left(\frac{Z_1}{Z_2} \leq t\right) = \frac{2}{\pi}\arcsin\left(\frac{\rho}{\sqrt{\delta_x\delta_y}}\right), \tag{23}$$

where Z_1 and Z_2 are independent mean zero Gaussians with covariance ρ and variance $1 + \frac{\rho^2}{\delta_x\delta_y}$ and δ_y respectively.

Proof: Note that, $Z := \frac{X}{\sqrt{\delta_x}} - \frac{\rho Y}{\sqrt{\delta_x\delta_y}}$ is a mean zero Gaussian independent of Y, and has variance $1 + \frac{\rho^2}{\delta_x\delta_y}$.

$$\begin{aligned}
\mathbb{P}(XY \geq 0) &= \mathbb{P}\left(ZY \geq \frac{\rho}{\sqrt{\delta_x\delta_y}}Y^2\right) && (24)\\
&= \mathbb{P}\left(\frac{Z}{\sqrt{1+\frac{\rho^2}{\delta_x\delta_y}}} \geq \frac{\rho Y}{\sqrt{\delta_x}\delta_y\sqrt{1+\frac{\rho^2}{\delta_x\delta_y}}}\right) && (25)\\
&= \mathbb{P}\left(\frac{\tilde{Z}}{\tilde{Y}} \geq \frac{\rho}{\sqrt{\delta_x\delta_y}\sqrt{1+\frac{\rho^2}{\delta_x\delta_y}}}\right), && (26)
\end{aligned}$$

where $\tilde{Y}$ and $\tilde{Z}$ are independent standard Gaussians, and thus their ratio is the Cauchy distribution. Therefore, the last probability above can be calculated as:

$$\begin{aligned}
\mathbb{P}(XY \geq 0) &= \frac{1}{\pi}\arctan\left(\frac{\rho}{\sqrt{\delta_x\delta_y}\sqrt{1+\frac{\rho^2}{\delta_x\delta_y}}}\right) && (27)\\
&= \frac{1}{\pi}\arcsin\left(\frac{\rho}{\sqrt{\delta_x\delta_y}}\right). && (28)
\end{aligned}$$

This completes the proof of Lemma C.1. □

D Comparison with PRIM (continued from Subsection 4.1)

In this section, we compare our algorithm's performance against local search based algorithms (PRIM based subgroup discovery).

As a baseline, we implemented a version of the PRIM subgroup discovery algorithm Friedman and Fisher (1999). PRIM is a local search algorithm, where the subgroups are defined by axis-parallel hyper-rectangles. The algorithm work as follows: One starts with a hyper-rectangle covering the entire dataset, and peels off a small portion of the rectangle from one or more dimensions (sides) in each iteration of the algorithm. In each iteration, one computes and stores the specificity for the subgroup defined by the points within the rectangle and after trying a sufficient number of rectangles one returns the optimal subgroup (with highest specificity) seen thus far. In this way the algorithm gradually trades-off sensitivity for increased specificity.

In our case, we additionally want to ensure a sparsity constraint, and then compute the specificity for a fixed sensitivity in order for comparison with our algorithm. So we make the following modifications:

- We enforce our sparsity constraint by using rectangles of dimension at most k that are chosen as random subsets of size k from $\{1, ..., 50\}$. The reason for using 50 is that in our case the subsets correspond to subsets of the 50 possible nuclear receptors.
- We require that the generated rectangle have at least σ fraction of points that correspond to patients with metastatic disease (σ was chosen to be either 0.1 or 0.3 for comparison). Thus ensuring a minimal sensitivity for the generated subgroups.

Finally, we capped our search by generating 400K axis-parallel rectangles. It required about 5× the computational time needed compared to Algorithm 1.

Baseline (PRIM) vs Experiment (MAX-CUT) specificity: The results for the rule-based local search PRIM algorithm are shown in Figure 5.

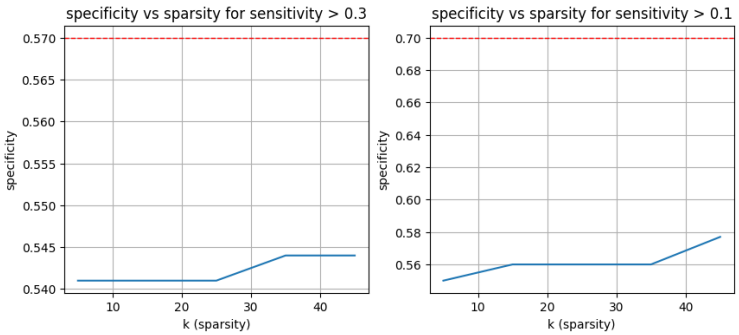


Figure 5: Plot specificity vs k (sparsity) as sensitivity is held above a lower bound for the variant of PRIM algorithm used for comparison. The dashed Red line (in both figures) denotes the specificity for the MAX-CUT algorithm, from Figure 1, for the given sensitivities (0.3 and 0.1), when the number of genes used is 15 (L figure) and 45 (R figure). In both cases, the PRIM local search heuristic fails to approach the specificity of Algorithm 1.

Therefore, the MAX-CUT based linear separator approach shows better specificity than the rule-based PRIM algorithm for a given sensitivity. This is not unexpected as Algorithm 1 is not a simple local search confined to an axis-parallel hyper-rectangle, but a more sophisticated SDP rounding based algorithm that computes a general (non-axis-aligned) sparse linear separator.