Finding the Bad Apples in the Bunch: Using Robust Optimization to Optimally Structure A Cancer Megafund

Ivan Paskov, Fransisca Susan

March 2019

1 Introduction

While promising breakthroughs in biomedicine such as molecular-targeted drugs and gene-therapy techniques offer new hope for curing cancer, they have also made biomedical innovation riskier and more expensive. These breakthroughs generate many new potential methods for investigation, each of which requires years of research and clinical testing that costs hundreds of millions of dollars while facing a high likelihood of failure [1]. Indeed, the productivity of big pharmaceutical companies has been declining steadily over the past 70 years, as measured by the number of drugs approved per billions USD spent [2]. Stock-price performance and venture capital investment returns in the pharmaceutical sector have also been disappointing, shown by a loss of roughly $850 billion in shareholder capital among the 15 biggest pharmaceutical companies from 2000 to 2008 [2].

As biomedical innovation became riskier, more expensive, and more difficult to finance with traditional funding sources, a new financial structure, *Megafund*, was proposed by Fernandez, Stein, and Lo [1]. The cancer megafund invests in a large number of drug development projects at various stages - phases in drug development which include preclinical, phase 1, phase 2, phase 3, and the NDA stage - to substantially reduce the portfolio's risk. Here, project selection is a crucial aspect in constructing a portfolio which maximizes the return over risk ratio, as we want to allocate capital to high-potential opportunities and avoid the non-performing ones. In determining which projects are worth investing in, we need to consider many factors including the cost, timeline, and risk of each project. The idea of canceling out the risks among projects only works if we consider a large number of projects, yet project selection easily becomes a highly combinatorially complex problem.

2 A Robust Optimization Approach

As discussed above, the success of such a megafund completely hinges upon our ability to select good projects and avoid bad ones. Motivated by this, we elect to take on a robust optimization approach, but with an interesting twist. Typically, with robust optimization, we desire to be inoculated against a variety of scenarios. In our case, we are not so concerned with being inoculated against the worst case, but rather in understanding what that worst case in fact is - and what essential features make it "bad." In our cancer megafund example, that would correspond to considering the space of all possible subsets of projects, and from those, identifying the worst possible subset, i.e., the bad apples in the bunch.

In the field of statistics, there is great interest in having such methodologies that are robust to all possible sub-populations of the larger population, but unfortunately such problems are often thought to be NP-Hard, due to the exponential nature of considering all possible sub-populations. Here we show that by employing a particular representation of that space, in conjunction with some tricks from duality theory and robust optimization, that certain classes of problems are in fact very tractable, often resulting even in linear programs.

2.1 Intuition

It is well known that for a set with cardinality n, that one can derive 2^n possible subsets from it. The number 2^n arises by realizing that any subset is completely defined by visiting the elements in the original set, and for each one, asking a binary question: "are you in this subset or not?" Since each element independently has two choices, to participate or not participate, and there are n such elements, we get that there are 2^n possible subsets.

Taking this line of reasoning further, it becomes natural to associate to each of our projects a binary variable z_i that indicates whether or not that specifc project will participate in a given subset of projects. This can then

be woven into either a regression or a classification task where the goal will be to have the model do well even over this worst case scenario. We do this in the next section.

2.2 Subpopulation Robust Regression and Classification

2.2.1 Formulation

Formalizing the intuition developed in the last section into a model, consider the following problem:

$$\min_{\beta} \max_{z_i} \sum_{i=1}^{n} z_i f_i(\beta; x, y) \quad \text{subject to} \quad \sum_{i=1}^{n} z_i = k, \quad z_i \in \{0, 1\}$$

where $f_i(\beta; x, y)$ can either be a regression loss such as $f_i(\beta; x, y) = |y_i - x_i^T \beta|$ or a classification loss such as $f_i(\beta; x, y) = \max\{0, 1 - y_i \beta^T x_i\}$. It is clear why the above formulation is a faithful translation of our earlier intuition (i.e. it says find a β that does the best against the hardest subset of size k in the data), but unfortunately at present it is quite a difficult problem as we have the multiplication of binary variables with a nonlinear function of another decision variable.

2.2.2 Efficient Algorithm

We solve this difficulty by making the following observation: for a fixed β, note that $f_i(\beta; x, y)$ is a constant. Thus first we may relax the $z_i \in \{0, 1\}$ constraint into $0 \leq z_i \leq 1$ as we are now maximizing a linear (and hence convex) function over an interval, $[0, 1]$, and thus we know the optimal z_i will occur at one of the two endpoints. Second, since the inner (max) optimization problem is linear in z_i, and all the constraints are now linear, we can take its dual, and are guaranteed that strong duality obtains. Doing so, we observe a significant simplification, and arrive at the following problem:

$$\min_{\beta, \theta, u_i} k\theta + \sum_{i=1}^{n} u_i \quad \text{subject to} \quad \theta + u_i \geq f_i(\beta; x, y), \quad u_i \geq 0$$

In the case of both the regression and classification losses given earlier, the above problem reduces (after a simple transformation) into a linear program.

3 Application to Cancer Megafund

Once such a model is trained on the cancer megafund data, there are several interesting directions to consider:

- All of the projects for whom the corresponding $z_i's$ came back equal to 1 represent those projects that are most risky. With this information, a portfolio manager could choose to simply remove them from consideration, use them to form a secondary "junk-bond portfolio" independent of the main portfolio, or separate projects into junior and senior tranches appropriately.
- Since the $\beta's$ were learned from this hardest subset, they contain information, i.e. features, that predict risk. Thus it would be helpful to examine them and identify which features are most predictive of risk so they could in the future be used in an advisory capacity when designing new projects. A principled way that this could be done is by formulating this analysis as a classification problem, where we use the learned z vector as our labels, and our original data as our feature matrix, and then run classification trees.
- We can also do an iterative process by optimizing the value of β on a subset of the original project pool, without the projects whose corresponding $z_i's$ came back 1, doing this iteratively gives us the final "least risky" project that we can compare with the result of solely minimizing the original objective function with respect to both z and β.

Knowing a group of projects with the highest probability of failure is beneficial for the megafund to make an informed and well-calculated investment decision. It will help them identify which factors contribute to a higher risk in drug development, giving them insights about the investment without requiring a deep understanding of the biomedical field itself.

References

[1] Jose-Maria Fernandez, Roger M Stein, and Andrew W Lo. Commercializing biomedical research through securitization techniques. *Nature biotechnology*, 30(10):964, 2012.

[2] Jack W Scannell, Alex Blanckley, Helen Boldon, and Brian Warrington. Diagnosing the decline in pharmaceutical r&d efficiency. *Nature reviews Drug discovery*, 11(3):191, 2012.