BIORANKINGS®
TECHNICAL REPORT SERIES

Dealing with High-Dimensional Data

When deciding how to analyze ***big data*** there are two ways of thinking about it. Are the data organized in a few columns and lots of rows (*tall and narrow*), or are the data organized in lots of columns and few rows (*short and wide*)? An example of *tall and narrow* is credit card transactions. According to the Federal Reserve Bank, in 2012 there were 73.9 billion transactions on credit, debit, and prepaid cards – each representing a different row in the bank's database. However, each transaction contains very little information – card number, transaction amount, date, location, what was purchased, etc. While it is big data because of the number of rows, the amount of analyses that can be done is limited because of the small number of variables.

In biomedical research using high-throughput technology (i.e., -omics), *short and wide* occurs -- lots of columns (*wide*) corresponding to the biological measurements, and few rows (*short*) corresponding to patients or samples. While this may not be big data in terms of storage, it presents huge problems from having a very large number of ways of analyzing the many biological measurements.

In this ***Technical Report*** we focus on *wide and short* data.

Why Dimensionality is Cursed

When statisticians think about –omics or other big data in medical research, one important property of the data are the dimensions. In this way, the data become points on a graph where each axis represents a different covariate (e.g., taxon in microbiome data, an edge in fMRI connectome graphs, a clinical variables in electronic health records). Each sample can be plotted on the graph (if not on paper then on computers and in mathematical notation) based on the values of its covariates (we call the position where the sample is plotted as its *coordinates*). In this framework, classical statistical methods work if the number of dimensions is relatively low. However, when the number of dimensions gets large, very weird things begin happening to the data, and applying statistical methods used for low dimensional data will give wrong results. This is known as the ***curse of dimensionality*** or ***large P, small N*** problem. In other words, you have a lot more covariates than samples (*short and wide* data).

Figure 1 shows 1000 data points plotted on two axes – or two dimensions – randomly generated from Uniform distributions between -0.5 and 0.5. This means that every number between -0.5 and 0.5 has an equal chance of being selected. Overlaying the points are the solid reference lines at the center of the data corresponding to the (0, 0) coordinate. The square indicated by the dashed lines is the 10th percentile region which means that 10% of the data is expected to be contained in it, and 90% outside of it.

There are expectations that data should behave which occurs when there are few dimensions, but everything falls apart as the number of dimensions gets larger. ***This is where the curse of dimensionality gets weird and why trying to make predictions from high-dimensional data is a mistake.***

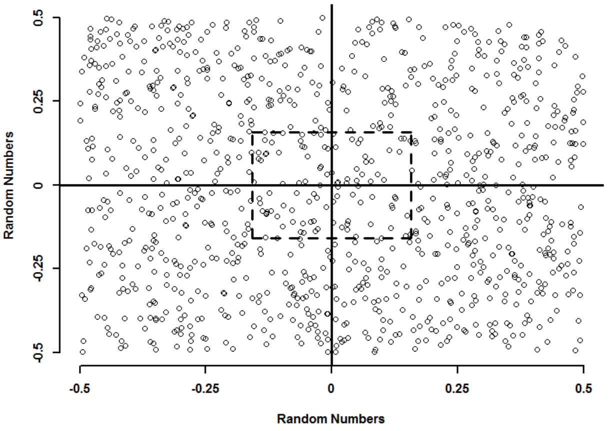

Figure 1

A simulation was run with 1,000 samples and the number dimensions (covariates) ranging from 1 to 1,000. The data for each sample was generated from a Uniform distribution ranging from -0.5 to 0.5 along each dimension. Half of the samples were randomly assigned to one group and half to a second group. Since all the data is randomly generated there are no associations or significant variables for predicting the group assignment. The simulation was run 100 times with the results averaged across them.

Figure 2 shows the weird things that happen in the vast space that is generated in high dimensions. In each of the four plots the X axis corresponds to the number of dimensions. The Y axes vary depending on the measure. Focus on the

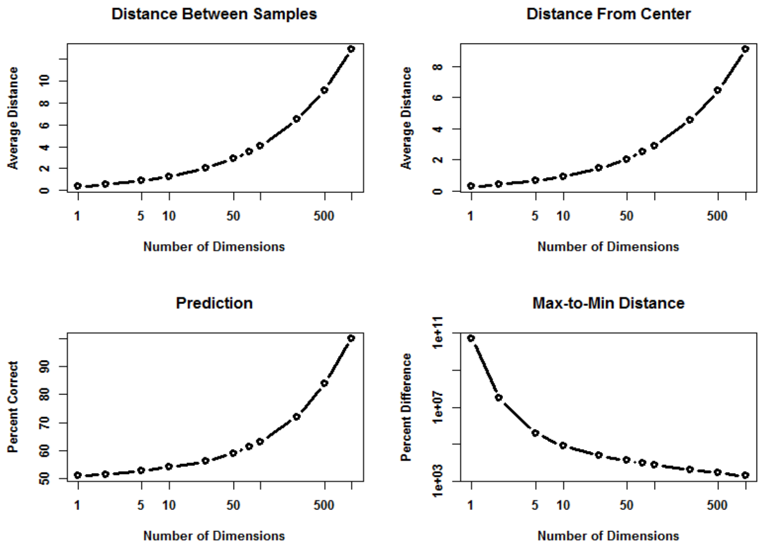

Figure 2

trends of the curves since the magnitudes of the Y axes are irrelevant and these trends are the same for any sample size, distribution of the data, distance measures, or prediction models.

The **Distance Between Samples** plot shows that the average distance between pairs of samples get larger with increasing dimensions. This means that the data points are getting further away from each other.

The **Distance From Center** plot shows that the average distance between each sample and the center point at (0, 0) get larger with increasing dimensions. This means that the data points are moving away from the center of the data and towards the outer edges.

The **Prediction** graph shows that as the number of dimensions increases models can be found that predict group membership perfectly, even when the data are random and unrelated to group membership. This results in wrong prediction models.

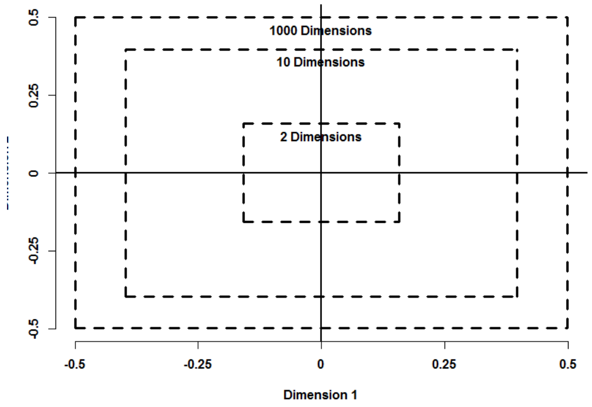

Figure 3

The **Max-to-Min Distance** graph is subtle in terms of how it is calculated and what it means. The Y axis is the percent difference between the maximum and minimum pairwise distances with large values indicating the two samples furthest apart from each other are much further apart than the two closest samples. This is seen for low dimensions. For small values this means the two samples furthest apart are about the same distance apart as the two nearest samples in the data. In other words all the samples become equally distant from every other sample making any cluster analysis meaningless.

Figure 3 shows that as the dimensions increase the data moves towards the outer edges of the coordinates. This is another way to view the **Distance From Center** plot. In Figure 3 the 10% boundary squares (i.e., 10% of the data will fall within and 90% outside of the dashed lines) for 2, 10, and 1,000 dimensions are shown. This shows that as dimensions increase 90% of the data is at the outer ring of the data space – data disappear from the center!

Multiple Testing Adjustment

Almost everyone with high-throughput data ask about multiple testing adjustment and say they aren't happy with it because none of the variables remain significant. The ***multiple testing problem*** and the ***curse of dimensionality*** go hand in hand with each other – if you have one you have the other since they both rear their ugly heads when you have measured a lot of covariates.

Most people have somewhat of an understanding of what P < 0.05 means. In simple terms, if you test a *non-significant covariate* between 2 groups, say using a t-test, there is a 5% chance its P value will be < 0.05. When this happens you won't know for certain if it is significant or non-significant, but can say that you only have a 5% chance it is not.

Multiple testing becomes a problem when you are testing more than 1 covariate. Consider testing 2 covariates. It may happen the first covariate is significant, the second covariate is significant, or both covariates are significant. This impacts the probability calculations. If you find one covariate significant with P < 0.05, the probability it is not significant is no longer 5% as in the case with 1 statistical test, but rather 9.75% chance it is not significant. If there are 3 covariates the chance of a significant covariate not really being significant is 14.3%, and for 10 covariates it becomes 40.1%. In other words, multiple testing really increases your chance of being wrong by deciding a non-significant covariate is significant, and you waste time and money following up on false positive results.

Figure 4 shows P values from a t-test 1,000 random variables comparing two random groups of 50 samples. (Again, this holds for any distribution, sample size, and statistical test – we

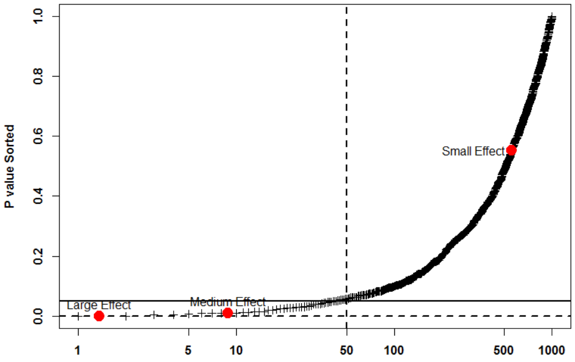

Figure 4

are interested in the trend). Since this was all random data no covariates are predictive of group membership except by chance. Along the X axis is the ordering of the P values from smallest to largest (on a log scale to spread them out a bit), and along the Y axis the P values. The black '+'s' correspond to the 1,000 non-significant covariates. Three *significant* covariates with standard small, medium, and large effect sizes were simulated and their P values shown as red dots.

Accepting all covariates with P < 0.05 (points below solid dashed line at P = 0.05) would capture the medium and large effect size significant covariates, as well as about 50 false positives (+'s to the left of the vertical dashed line at 50) – and there is no way to distinguish the true positives from the false positives.

Many analysts want to use multiple testing adjustment to eliminate false positives and enrich for true positives. However, this approach is only moving the goal post and does not ensure you end up with the significant covariates. The horizontal dashed line at P = 0.05 / 1000 = 0.00005 is the *Bonferroni* adjusted P value. If we only declare covariates significant with P less than this value we capture the large effect but lose the medium effect covariate. Since significant covariates do not necessarily have large effect sizes (in fact most are probably small), you will have no idea how many significant covariates you have lost, or what the false positive and false negative rates are.

Multivariate versus Univariate

Testing each covariate or dimension by itself is called ***univariate*** testing, and looking at 2 or more simultaneously is called ***multivariate*** testing. While moving the P value goalpost with multiple testing adjustment seems appealing, it does nothing to help find the dimensions or covariates that by themselves have no signal, but in combination with other covariates become very important. In genetics this is known as epistasis and can result in missing important genetic variants. The same problem exists for any high-throughput technology producing a lot of dimensions.

There are three seemingly obvious approaches to solve this. The first is a dimension reduction approach such as PCA. However, the curse of dimensionality results in samples being equidistant from each other so the mathematics won't work (the mathematics is so unstable in l*arge P, small N* problems that the addition or deletion of a single sample, or the slight change to some of the numbers can result in completely different conclusions – imagine a regression model where the slope is positive until you add one more sample and the slope becomes negative).

The second is to select dimensions based on biology and an understanding that they share a common function (e.g., SNPs in a pathway). This can reduce the dimensions to where standard multivariate methods apply and is something we strongly encourage.

When biology does not indicate which dimensions (covariates) to keep, another approach is to explore subsets of covariates to find which in combination are significant. This is a multivariate approach and avoids problems like epistasis. However, the ability to compute all the combinations is difficult (it is called combinatorial explosion) because there are too many ways to test all two-way combinations, three-way combinations, etc. To solve this problem BioRankings has developed a genetic algorithm to find subsets of the covariates in high-throughput data that in combination are significant. This is discussed in BioRankings Technical Report #3.

Conclusion

The ***curse of dimensionality***, or ***large P small N*** problem, causes weird behavior in data that make classical statistics lead to the wrong conclusions. This includes distances which destroy any cluster structures that might exist in lower dimensions, and the ability to be able to predict a random outcome with perfect precision. It must be remembered that any results you get may be completely wrong and due simply to the dimensions. Avoid being overconfident.

High-dimensional data should be and will be analyzed, but should be thought of as an exploratory analysis. Conclusions from these first analyses should be confirmed in well designed controlled follow-up experiments.

BioRankings' Technical Report Series

BioRankings' mission is to help biomedical researchers move their technology from the lab to clinical applications using statistically valid analytical tools for efficient study designs, correct data analyses and conclusions, and rigorous and objective decision making for designing follow-up studies and eventual FDA approval.

To help achieve its mission, BioRankings publishes a Technical Report series focused on applying various statistical methods to real data analyses. Written for *understanding* by scientists and administrators, these reports will provide an intuitive understanding of the analyses leaving the statistical details to other publications.

For more information, contact BioRankings at 314-704-8725 or bill@biorankings.com.