Robbie Eberhart-Garah
Thesis Draft
1/15/14

A Novel Statistical Method for Identifying Monotonic Relationships in Noisy Plots

Introduction

As scientific technologies and techniques have improved in past decades, it has become possible to quickly collect unprecedented quantities of experimental data. In the study of gene expression, for instance, datasets comprising tens of thousands of variables and hundreds of treatments can be produced in a matter of months. A common purpose of experiments is to investigate relationships between variable pairs, and various statistical methods exist for identifying trends in data. However, many of these methods are not suited to handling very large datasets for several reasons.

First, large datasets are expected to contain some low quality data. When an analysis consists of three or four treatments, the robustness of each treatment must be verified, and poorly produced or nonsensical treatments must be pruned. When analyzing three or four hundred treatments, manual verification becomes implausible. For some statistical measures, such as correlation, a single particularly bad treatment in a set of hundreds can be confounding. A bad datapoint can get washed out in a large dataset, but theoretically, using the statistical method described in this paper, one bad data point cannot disrupt the analysis. Because low quality data cannot be manually pruned, a useful statistical method must be robust to various types of random and semi-random noise.

Even if no low quality data is introduced into a dataset, strength of signal is frequently low. Even if two variables are closely related, such as the expression of two genes involved in a specific pathway, it is unlikely that they are related under all conditions. Unless hundreds of treatments and replicates have been performed to study a single aspect of the relationship between one variable pair out of thousands or millions, many can be expected to be non-informative for each individual variable pair. Under these conditions culling bad data is not only impractical but impossible: bad data, properly considered, is just good data in the wrong context. A useful statistical method should, therefore, be well suited to identify subsets of data in which a relationship is expressed while ignoring non-informative data.

This paper presents a novel statistical method for identifying monotonic relationships between all variable pairs in large data sets. Specifically, it is designed to perform two functions: 1) to determine the probability that a plot is random, and 2) if the plot does not appear random, to select the subplot which is most likely to contain the signal. While this method does not attempt to identify the function which describes a relationship, it can perform the useful task of selecting the specific treatments or conditions under which the relationship manifests. The selected subplot can also be easily passed to other methods, such as least squares regression, allowing them to identify the function with substantially reduced interference.

Method

The method presented in this paper uses the longest monotonic path of a plot (the largest set of points which can be connected in a single monotonic path) as an indicator of signal strength, and will appropriately be referred to as Longest Path Analysis (LPA) throughout this paper. Put simply, LPA finds the longest monotonic path (LMP) of a plot, compares it the expected LMP of a random plot of that size, and gives the probability that the plot is random based on the difference. If the p-value is low, the null hypothesis that the plot is random is rejected and all data which is included in a longest path is selected for further analysis.

LPA rests on several key assumptions. First, the LMP through a plot containing a monotonic signal must be longer than the average LMP through a random plot of the same size. Another way of saying this is that since signal should represent an unusually dense section of a plot, the LMP through dense regions of a plot should, all things being equal, be longer than the LMP through sparse regions. This follows from the assertion that the LMP grows with plot size (which represents a universal increase in plot density), and was clearly demonstrated in testing.

There must be a relationship between the average LMP length in a random plot and the size of the plot itself. The method requires a function which accurately describes the relationship. There must also be a relationship between the standard deviation of LMP length in a random plot and size of the plot, and this too must be described accurately by a function. Both of these assumptions prove quite robust. A third, and less intuitive, assumption is that all random plots have the same geometric conformation, a problem which is more difficult to resolve but may be irrelevant in practice. In any case, this third assumption will be discussed later in the paper.

Before any of these assumptions can be tested, it is necessary to be able to identify the LMP of a plot. This problem is difficult one, and a longtime favorite of computer scientists and mathematicians. In the worst cases it is functionally impossible to calculate in large plots; in the best cases it is practically instantaneous even in substantial plots. After several amusing false starts using worst-case algorithms (O($3^{n/3}$)), it was determined that the conditions of the plot which LPA operates in are ideal for LMP finding algorithms, and can be solved with complexity O(nlog(n)) using a dynamic programming approach, making LPA of quite large plots practical.

Results

An equation describing average LMP of a random plot as a function of plot size was developed using random plots of 2 to 125 points. An average LMP was found for 3,500 sample random plots of each size and plotted, and least squares regression was used to find a model function. The function found in this manner was $f(n)=1.315n^{0.553}$, with an R^2 of 0.9992. This fit could be improved by increasing plot size range and sample number (both were limited in the original tests by the method that was used), but the R^2 value does suggest that the relationship is real and that the equation found is quite close to the actual function.

An equation describing standard deviation of LMP of a random plot as a function of plot size was developed from the same data set. The standard deviation was found amongst all 3,500 plots of each size and plotted, and least squares regression was again used to fit an equation to the data. The function found for standard deviation was $f(n)=0.508n^{0.255}$, with an R^2 of 0.9948 (fig. 2). The larger variability in this plot may be due to the substantially smaller range and some slight inconsistencies in how plots were grouped when standard deviation was calculated. Nevertheless, the R^2 value is certainly high enough to suggest a strong relationship between standard deviation of the LMP and plot size, and the derived equation is again quite close to the actual function.

Both of these functions describe characteristics of a random plot whose data is distributed uniformly in a square. They are not accurate when the distribution of background noise is different. However, average LMP and standard deviation of LMP can be calculated for a specific experiment. Assuming that most variable pairs in an experiment are not expected to be related, the LMP and standard deviation of LMP can be found for a random sampling of variable pairs and used as the expected values for that analysis.

Using these two equations, an expected LMP, actual LMP, and standard deviation can be found for any plot. This information can be readily converted into a p-value using a simple Z-test. Insomuch as the purpose of LPA is to determine whether a plot follows a truly random distribution, this is enough information to validate the method. The p-value produced is the absolute probability that a given random plot would have a particular LMP: false discovery rate and statistical power do not apply to the core of the method itself. This answer is not satisfying for practical concerns, however, and while developing meaningful metrics of a statistical method's behavior is a thorough and exhaustive process which is beyond the scope of this project, some cursory tests have been performed to test the capacity and sensitivity of LPA in discovering monotonic trends.

A brief examination of LPA's ability to detect signal in a noisy environment was performed by planting a signal function, f(x)=x, in a plot and adding various levels of noise. In a plot of 125 points (Fig. 3), 80% noise levels were no obstacle to the detection of the 25 signal points. This demonstrates LPA's affinity for detecting a strong signal subplot in an environment of random noise, especially as compared to Pearson's correlation, which showed no relationship between the two variables.

It is worth mentioning that real data, of course, rarely forms a perfect line, because perfect signal is rarely found in experimental data. Signal and noise are usually integrated in each data point. This method was developed to find subplots with a low noise content, and the decay of signal detection as integrated noise levels rise in a plot with no purely random data has not been tested widely yet. A function probably exists and could be discovered relating LMP to noise level in an integrated noise/signal plot, and this could be used to predict the statistical power of LPA in environments with truly random data and integrated noise.

Comparison was made to Pearson's Correlation Coefficient (PCC), as it is a widely used statistic which searches for similar patterns. LPA performs significantly better than PCC at detecting linear relationships in noisy environments, as should be expected: PCC assumes that all data is informative, while LPA assumes that most is not. In data sets with no random data but some integrated noise, PCC performs admirably but not any better than LPA.

In Fig. 4, PCC finds an R of ~1, and LPA finds a p-value of 10^{-64}, somewhat below the .05 significance threshold often employed. LPA can find any monotonic trend while PCC is limited to straight lines, but Spearman's correlation also identifies monotonic trends generally so this is not a revolutionary feature. LPA has far better outlier resistance than either measure of correlation, but automatic outlier culling methods such as Cook's distance (which I know only the name and description of) may make this quality relatively irrelevant.

Conclusion

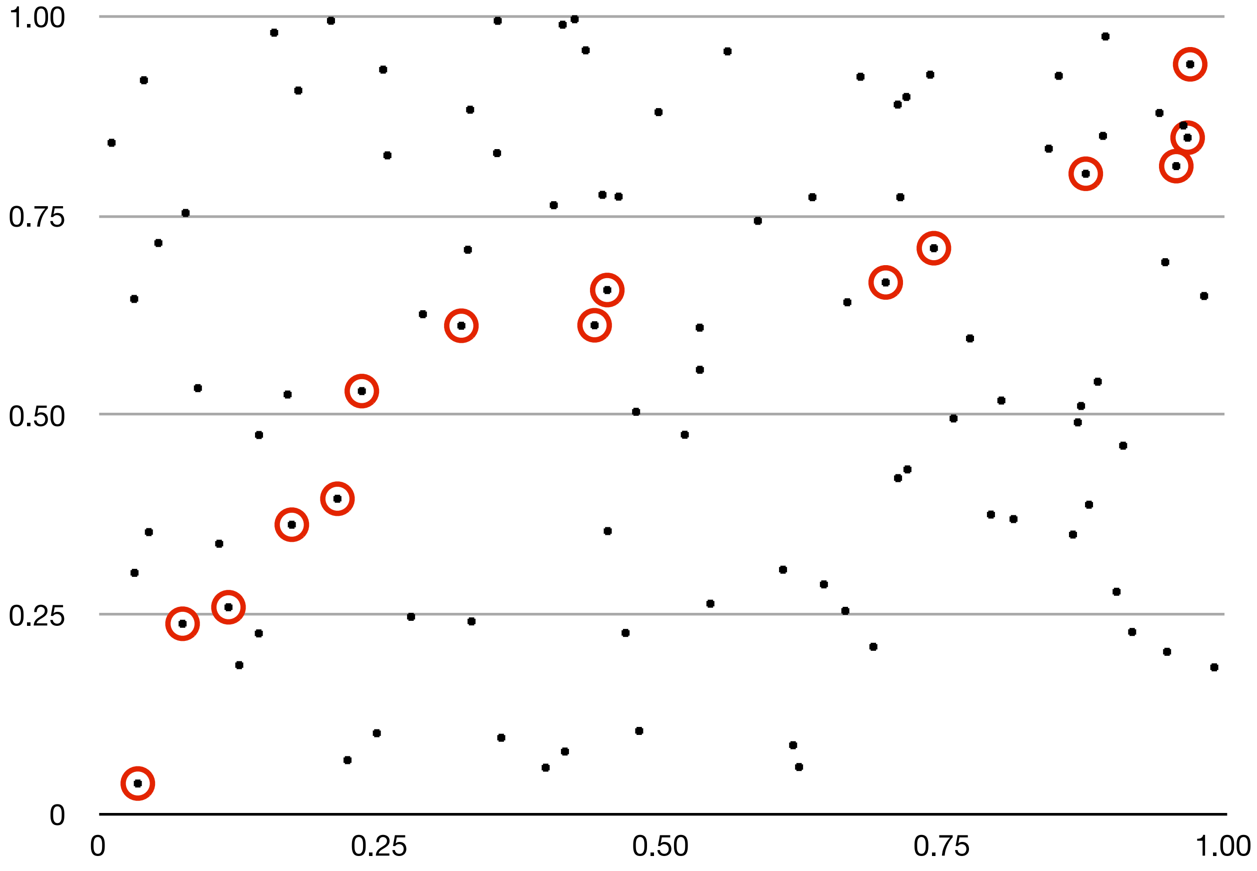
Longest Path Analysis fills a scantly occupied niche in statistics of identifying functional subsets of data for analysis instead of looking for trends in the behavior of an entire plot. Whether any methods already exist which perform this function is uncertain (I am not a statistician, nor do I know where any live), but it is at least outside the main focus of statistical analysis. It is likely that such methods will become more desirable as datasets increase beyond the capacity of traditional manual examination and analysis, and as low quality data becomes more inextricably included in analysis.

Beyond the simple task of removing low quality data, LPA makes it possible to search for relationships between variables under conditions which were never tested for. Components of large datasets are often gathered from different labs by different people who may not be working under uniform conditions: traditionally this is an obstacle to analysis, because two treatments which are supposed to vary in only one parameter may have undocumented variation in confounding parameters. In the case of genetics, organisms in different labs may be raised with different temperatures, lighting, nutrition, mediums, etc. which make comparison of response to variation in a single parameter impossible. However, when data from enough experiments performed under essentially random conditions is considered as a whole, the net effect may be that treatments which were never intended to be tested nevertheless appear. Analysis of a large number of control treatments conducted in different labs could produce a wide range of data containing detectable relationships between genes. Effectively, searching through low quality data for significant subplots may be at least as useful as traditional variation of treatments because, while each signal may be weaker, there may be so many more signals.

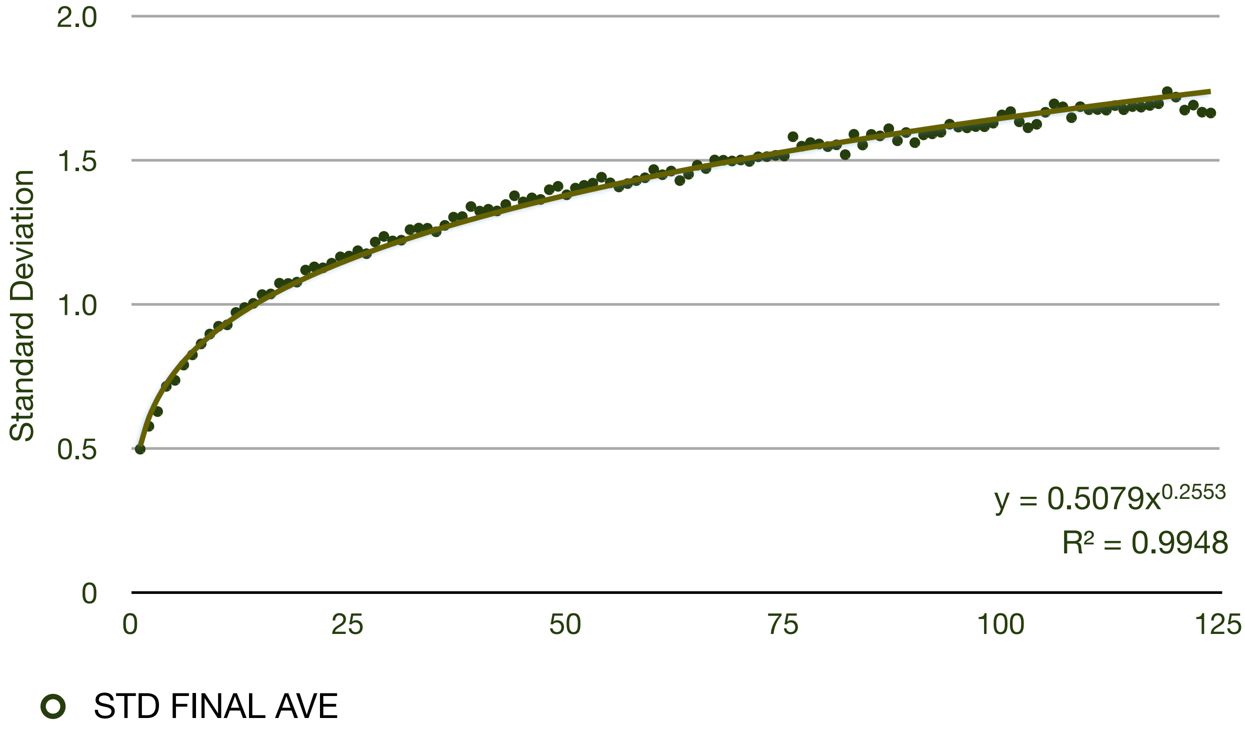
As such, LPA is a useful tool for identifying predicted signals in experimental data, but may be more instrumental in detecting novel, unexpected relationships where traditional noise reduction techniques are harder to employ. It reduces the need to manually select relevant treatment pairs, a crucial characteristic when the total of number of treatment pairs in an experiment can easily be in the thousands. And it makes it possible to simultaneously search for relationships between every single variable pair in an experiment, a process which is pointless when considering variation in a single parameter or small set of parameters.

In the limited extent to which it has been tested, LPA is quite robust: expected LMP changes slowly in relation to plot size, so adding noise to a plot with good signal has little effect, regardless of the distribution of that noise. Its greatest obvious limitation is that it cannot identify non-monotonic trends and is thus unsuitable for identifying common patterns such as sinusoidal waves and parabolas. Its further limitations I lack the expertise to anticipate, but will likely be revealed in future testing or examination by people who do have expertise.

Longest Monotonic Path of a Random Plot (Fig. 1)



Standard Deviation of LMP Size as a Function of Random Plot Size (Fig. 2)



Signal f(x)=x with Added Interference (n=125, signal=25) (Fig. 3)

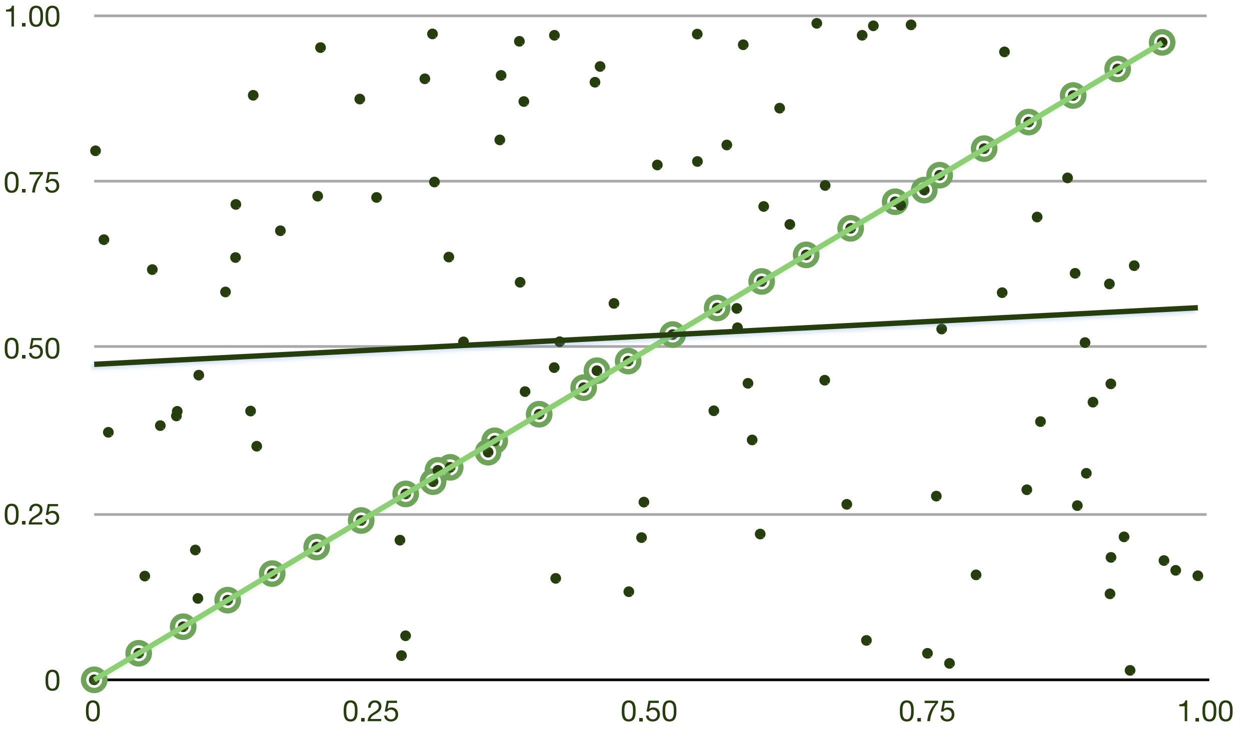
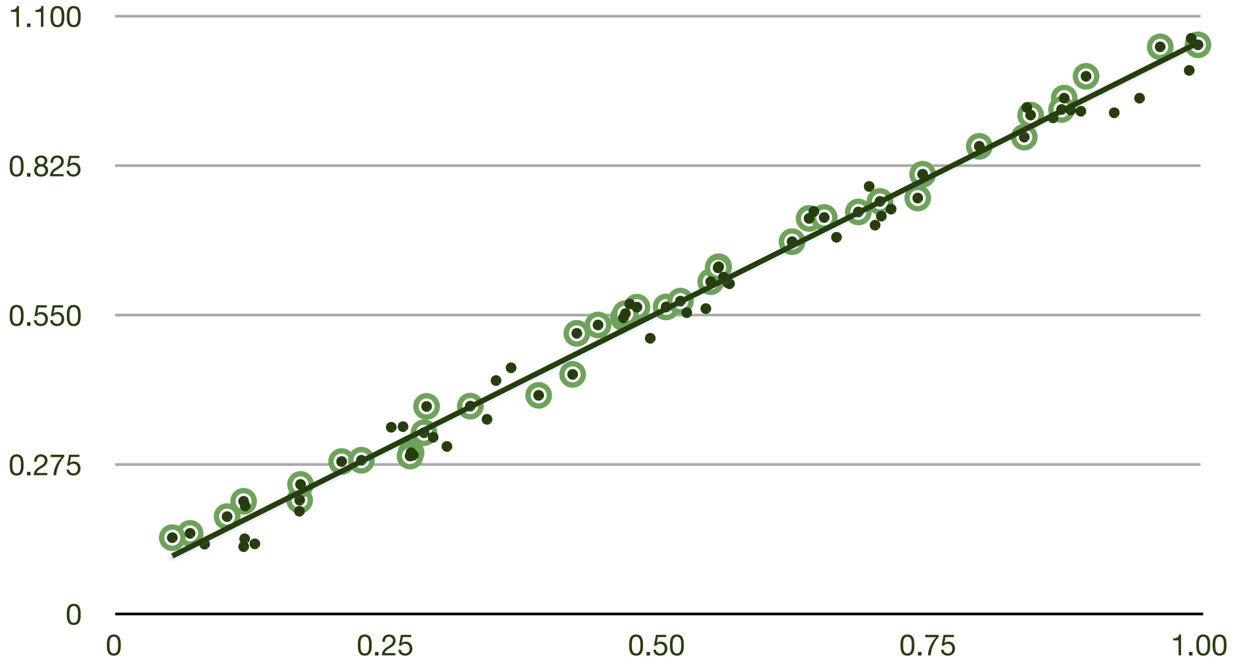


Fig. 4


1.100
0.825
0.550
0.275
0
0
0.25
0.50
0.75
1.00