Summer 2013

The Applications of Compressed Sensing in High Frequency Trading

Behrooz Ghorbani

Supervisor: Ozgur Yilmaz

Given a system of linear equations, $A = \Phi X$, where Φ is a $n \times N$ dimensional matrix and X is the N dimensional vector of unknowns, in general we need $n \geq N$ so that we can solve for X. Unfortunately, if N is a large number, producing $n \geq N$ independent equations from the physical phenomenon that we are modeling might be very costly or even impossible. Therefore, we would like to be able to impose some assumptions on the system and solve the system with a much smaller n. Compressed sensing is a new technique that allows the user to solve the system with $n \ll N$ provided that most of the entries of X are zero and Φ satisfies certain conditions. In order to understand this technique better, we introduce the following basic concepts:

Definition I. If Y is a vector, with at most *s* non-zero entries, Y is called *s* sparse.

Theorem I. If Φ is satisfied certain conditions[1], then

$$\arg min_y \, ||y||_1 subject\ to\ A = \Phi y$$

can recover the vector X.

Theorem II. If Φ is a matrix where each entry is drawn from N(0, 1/*n*) then Φ satisfies the above conditions with overwhelming probability if $s \sim \frac{n}{\log(\frac{N}{n})}$

The theorems above state that the system can be efficiently solved with at most $s\log(N)$ equations. In this project, we introduced a novel usage of this technique in statistics. Let's assume that we have a multivariate regression model, $Y = \sum_{i=1}^{N} \beta_i f_i + u$, where u is the unobserved error and we are trying to estimate β_i. In order to estimate the β_i, we take n measurements and we will have to solve the following system of equations:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ Y_n \end{pmatrix} = n \updownarrow \underbrace{\begin{bmatrix} f_{1,1} & \cdots\cdots\cdots\cdots\cdots\cdots & f_{1,N} \\ . \quad . & . \quad . & . \\ . \quad . & . \quad . & . \\ f_{n,1} & \cdots\cdots\cdots\cdots\cdots\cdots & f_{n,N} \end{bmatrix}}_{\Phi\ \text{Matrix}} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ . \\ . \\ . \\ . \end{pmatrix} N \updownarrow$$

[1] The conditions are usually stated as conditions on Restricted Isometry Constants. Due to limited space, they cannot be introduced in the framework of this report but interested readers can access the full theorem in the paper by Candes, Romberg, and Tao (2004)

Traditionally, ordinary least squares method is used to find the β_i that provide least l_2 error. However, this methodology suffers from two setbacks. First, the number of observations needed, n, has to be at least equal to N. Secondly, because of the problems of multicollinearity, if the correlation between f_i is high, the standard errors would be high. Usually when N is large, principal component analysis approach (PCA) is used in order to overcome this problem. However, PCA involves heavy calculations and still does not provide any remedy for the necessity of having a large number of observations.

Compressed sensing techniques, on the other hand, behave better with respect to correlations and the number of observations that they need is less than N by one or two orders. In this case, if the number of regressors is large and only a few of them are independent, in principle, we could use compressed sensing to approximate the β_i. However, the problem that one faces in using this method is that only special types of Φ matrices, e.g. Gaussian matrices, are suitable for compressed sensing.

In this project, our attempt was to devise various techniques that can be applied to different matrices that are not suitable for compressed sensing and transform them to suitable matrices without changing the solution vector. The special focus of the project has been on the factors that arise in stocks valuation. We have a list of more than 1500 of these factors and we use them as regressors while knowing that only a few of these factors are independent. Our goal is to sample the past daily data (or alternatively the high frequency data) of returns of a particular stock and predict the future returns of this stock. We have:

$$R_{specific_t} = \sum_{i=1}^{1500} \beta_i f_{i_t} \rightarrow R_{specific} = \Phi\beta$$

where Φ is an n-by-1500 dimensional matrix and β is a 1500 dimensional vector to be recovered. We have designed a function T such that if $R_{specific} = \Phi\beta$ then $T(R_{specific}) = T(\Phi)\,\beta$ such that $T(\Phi)$ is a matrix that is an appropriate (at least empirically) compressed sensing matrix. After designing such a procedure, we examined its predictive power on the real daily data. By using only 180 data points, our system was able to almost perfectly predict the daily stock returns of most stocks for the relevant time period. Since we are using only 180 observations, unlike the traditional method, the assumption that β is stable is in the time period of sampling is much more valid. Therefore, the resulting accuracy is much higher than in the normal case. Moreover, executing this regression barely takes a few seconds and this can give the practitioner the ability to run the regression more frequently and predict even the local variation of the target stocks.

Having obtained such results, we tried to move beyond financial signals and generalize the function T for other set of matrices that are not suitable for compressed sensing. Although we have reached some results, the scope of this goal is beyond a summer project and more time has to be dedicated to this problem. The expansion of this technique from a cross section regression to a time-series regression is another topic of interest that can be followed beyond this summer.

A quick survey of modern businesses shows that many pioneer firms use “big data” in order to achieve maximum productivity and profit[2]. Working with such large data sets needs revisions in the traditional approach to data analysis. This project demonstrates that to increase accuracy, speed, and

[2] http://en.wikipedia.org/wiki/Big_data

feasibility of widely used statistical instruments they can be reviewed with respect to the sparsity constraints that are present in the data. Instead of spending hours to run cumbersome calculations to estimate the regression coefficients, by using the new techniques concerning sparsity of the data, one can get faster, more accurate, and less resource intensive predictions.

This project draws various concepts from different fields, such as applied mathematics, finance, econometrics, and electrical engineering. This interdisciplinary nature presented me with the opportunity to learn various concepts in each one of these fields. In the beginning of the project in May, I familiarized myself in with the mathematics behind compressed sensing as well as its relation to signal processing. After this introduction, I started to search for a way to apply compressed sensing methodology to economic applications. This task required knowledge of the economic fundamentals and modern trading strategies. In order to find out more about with these subjects, I went through the finance literature. It was with this insight that we were able to exactly define the question that we were interested in.

Familiarizing myself with the statistical tools that are used to work with these large cross-sections gave me new ideas to implement in compressed sensing. Beside the standard de-correlation practices, I got to learn the principle component analysis (PCA), which is a widely used tool in modern econometrics. Moreover, I got to combine what I had learned in econometrics courses with matrix algebra, which was extremely helpful in looking at the data geometrically.

In this project, I used MATLAB and its various packages as the primary software for conducting analysis. I also started using Bloomberg to extract market data from time to time. As most of the research assistants in the department, I started using LaTex as the primary method of documenting my process.

In conclusion, I believe that for me, the most important effect of this project was to help me choose my future research interests. Through this project, I got a brief introduction to each one of the field that I was interested in, and I saw the kind of questions that the modern research in each field tries to answer and how does working in each of these field like. I should especially thank Dr. Yilmaz for supporting me in my exploration in different fields and not limiting me to only one narrow specific arena of studies.