Research on high frequency data sets

Nicholas Humphreys

Abstract

The task of this project is to take a data set that has several data points and using statistical analysis make the set into about 4 or 5 data points that still contain between ninety-eight and ninety-nine percent of the original information. This makes it so the data as a whole is easier to analyze and draw conclusions from. The data set I used for this project was a set of the average monthly temperatures in Prague ranging from 1775-1989. The goal is to take this 215 by 12 matrix and break it down to a function that we can analyze in order to determine if the temperature in Prague has changed over the last 215 years. Each year of this data set can be represented as a stepwise function by saying that each month represents 1/12, so the entire year is set along the x axis on an interval from 0 to 1. The points along the y-axis are just the temperatures for each of the twelve months. We can then use a Fourier expansion to make it a smooth continuous function. This new function is easier analyze. We can then see if the function follows Brownian Motion.

My research had four major parts this semester. The first was to understand L_2 spaces, the second was to find the eigenvalue and eigenfunction so we could use the Fourier expansion, The third part was to apply the eigenfunction and the stepwise function to the Fourier expansion, and finally I did some research to learn about Brownian Motion.

1 An introduction to L_2 spaces

My first assignment was to do some research on L_2 spaces to learn about all the different properties of this space. This was neccesary to do because I needed to find a way of representing the stepwise function as a continuous function for each year of data. The space that these functions are then analyzed in is the L_2 space. This is a summary of what I found out about L_2 spaces.

An L_2 space is an example of a metric space with the points in the space being continuous functions defined on an interval I = [a,b]. The metric of this space is defined by

$$d(f,g) = ||f-g|| = \left(\int_a^b |f(t)-g(t)|^2 dt \right)^{1/2} \tag{1.1}$$

and the limit as k goes to infinity of $f_k = g$ can be written, after squaring the results, as

$$\lim_{k\to\infty} \int_a^b |f_k(t) - g(t)|^2 dt = 0 \tag{1.2}$$

(1.1) is called the mean square convergence and is used for working with Fourier series. Even though mean convergence and uniform convergence are being used on the same functions they are in fact different. This is also true for the metrics of the L_2 space compared to that of the l_2 space which is defined by

$$d(p,q) = ||p - q|| = \left(\sum_{k=1}^{\infty} |a_k - b_k|^2 \right)^{1/2} \tag{1.3}$$

Mean convergence and uniform convergence are different, just as the metrics of the two different spaces are different, even though they are being used on the same collection of functions.

The inner product in L_2 is written as,

$$< f, g >= \int_a^b f(t)g(t)dt \tag{1.4}$$

and the same Schwartz inequality holds for this space, that is,

$$| < f, g > | \leq ||f||||g|| \tag{1.5}$$

Two functions f and g in L_2 are orthogonal when neither of them are 0, and $< f, g >= 0$. An infinite set of functions noted by $\{\rho_k\}$ is an orthogonal set if every distinct pair is orthogonal. This is stated as

$$< \rho_k, \rho_\ell >= 0 \quad k \neq \ell \tag{1.6}$$

and in integral form as

$$\int_a^b \rho_k(t)\rho_\ell(t)dt = 0 \quad k \neq \ell \tag{1.7}$$

If we have an infinite set of functions in L_2 such that the norm of these functions equals one, then the set of functions is said to be orthonormal on the interval I. These orthonormal

sets can be constructed from any infinite linearly independent set of functions by taking appropriate linear combinations to achieve the orthogonality.

The space of continuous functions is not complete in L_2. This means that there are functions in L_2 that may obey the Cauchy criterion, $\lim_{k\to\infty} f_k(t) = 0$, but may not be a continuous function.

A Fourier series may be used to enlarge the function space to include more discontinuous functions. Now, any sequence that is Cauchy in L_2 space will now converge in that L_2 space to one of the functions in the new space. In order to make the space larger, all piecewise continuous functions on the interval $I = [a, b]$ are included along with functions having an infinite number of discontinuities.

Orthonormal basis also exist in L_2 space. These basis are called Fourier basis. A orthonormal set $\{\rho_k\}$ in L_2 is a Fourier basis if every function in L_2 has the unique expansion.

$$f = \sum_{k=1}^{\infty} c_k \rho_k \tag{1.8}$$

Here $c_k =< f, \rho_k >$ for $k = 1, 2, 3 \ldots$

The sequence $\{c_k\}$ is usually referred to as the sequence of Fourier coefficients of f with respect to the orthonormal set $\{\rho_k\}$. $\{c_k\}$ can be constructed for any $\{\rho_k\}$ whether or not they form a basis. It is always possible to form the series

$$\sum_{k=1}^{\infty} c_k \rho_k = c_1\rho_1 + c_2\rho_2 + \ldots$$

Assuming that $\sum_{k=1}^{\infty} c_i^2 \leq \infty$.
This series is called a Fourier series for f with respect to the orthonormal set $\{\rho_k\}$. The integral form of c_k is

$$c_k = \int_a^b f(t)\rho_k(t)dt \tag{1.9}$$

When f is assosciated with its Fourier series, it is often written as

$$f(t) \sim \sum_{k=1}^{\infty} c_k \rho_k(t) \tag{1.10}$$

The set $\{\rho_k\}$ is a basis if and only if

$$\lim_{N\to\infty} \int_a^b |f(t) - S_N(t)|^2 dt = 0 \tag{1.11}$$

Where $S_N(t) = \sum_{k=1}^{\infty} c_k \rho_k(t)$ is the sequence of partial sums.

After completing the research on this subject I was given the opportunity to see how quickly functions converge in L_2 space. I was shown the Brownian Motion equation

$$W(t) = Y_0 t + \sqrt{(2)} \sum_{k=1}^{\infty} Y_k \frac{sin k\pi t}{k\pi} \tag{1.12}$$

I used the statistical package R to generate standard normal random variables to use in (1.12). I then made a plot of the equation with $k = 1$ then $k = 2$ and so on all the way up to $k = 20$. I found that the function is converging relatively quickly after about $k = 12$ This was useful because I knew I would be plotting the functions from the Prague data, so this gave me a good idea of how many terms I should use to construct the plot of each function.

2 Derivation of the equation for eigenvalues and eigenfunction

Now that I knew some propeties of L_2 spaces I could use (1.8) to create a nice smooth function of each year in the data set. c_k are the Fourier coefficients defined in (1.9). The variable f in (1.9) is the stepwise function created for each year of the data set as explained in the abstact of this report. I was able to find the orthonormal basis $\rho_k(t)$ by solving the equation $\lambda_k \rho_k(t) = \int_0^1 c(t,s)\rho_k(s)ds$ for $c(t,s) = \min(t,s)$.

We can rewrite the equation as

$$\lambda_k \rho_k(t) = \int_0^1 \min(t,s)\rho_k(s)ds$$

Now, the integral on the right hand side of the equation needs to be split up for the cases when $0 < s < t$ in which case $\min(t,s) = s$, and when $t < s < 1$ in
which case $\min(t,s) = t$. The equation is then rewritten as

$$\lambda_k \rho_k(t) = \int_0^t s\rho_k(s)ds \; + \; \int_t^1 t\rho_k(s)ds \tag{2.1}$$

To solve the integrals, we introduce two new functions. They are $f(s) = s\rho_k(s)$ and $g(s) = \rho_k(s)$. We call $F(s)$ and $G(s)$ the antiderivatives of $f(s)$ and $g(s)$ respectively, so that $F'(s) = f(s)$ and $G'(s) = g(s)$. (2.1) then becomes

$$\lambda_k \rho_k(t) = \int_0^t f(s)ds \; + \; \int_t^1 tg(s)ds$$

Recalling the fundamental theorem of calculus: $\int_a^b f(x) = F(b) - F(a)$ Where F is the antiderivative of f. By solving the integrals using the fundamental theorem of calculus we get

$$\lambda_k \rho_k(t) = F(t) \; - \; F(0) \; + \; t[G(1) \; - \; G(t)] \tag{2.2}$$

We can now take the derivative with respect to t of both sides of the equation. We can do this because the functions on the right hand side of the equation are antiderivatives, so therefore they can also be differentiated. Differentiating both sides we get

$$\lambda_k \rho_k'(t) = F'(t) \; - \; F'(0) \; + \; tG'(1) \; + \; G(1) \; + \; -tG'(t) \; + \; -G(t) \tag{2.3}$$

We can cancel out like terms and use the fact that the derivative of a function of a constant is 0 to rewrite (2.3) as

$$\lambda_k \rho_k'(t) = F'(t) \; - \; G(1) \; + \; -tG'(t) \; + \; -G(t) \tag{2.4}$$

Plugging in our known equations for $F'(t)$ and $G'(t)$, (2.4) is rewritten as

$$\lambda_k \rho_k'(t) = t\rho_k(t) \; - \; G(1) \; - \; t\rho_k(t) \; - \; G(t) \tag{2.5}$$

We can take the derivative of both sides again since the functions on the right hand side are antiderivatives they are also differentiable. We get

$$\lambda_k \rho_k''(t) = -G'(t) \tag{2.6}$$

Plugging in for G'(t) we get:

$$\lambda_k \rho_k''(t) = -\rho_k(t) \tag{2.7}$$

We now want to find an orthonormal system for this equation. The series of functions need to satisfy two conditions. They are:

(i) $\int_0^1 \rho_k^2(t)\, dt = 1$.

(ii) $\int_0^1 \rho_k(t)\rho_\ell(t)\, dt = 0$ where $k \neq \ell$

I found two sets of functions that meet these conditions. They are

(i) $\rho_k(t) = \sqrt{2}\cos(k\pi t)$ for all $k = 1, 2, \ldots$

and

(ii) $\rho_k(t) = \sqrt{2}\sin(k\pi t)$ for all $k = 1, 2, \ldots$

We can now take the second derivative of both of these equations and plug either one of them into (2.7) to get:

$$\lambda_k(-\sqrt{2}k^2\pi^2\sin(k\pi t)) = -\sqrt{2}\sin(k\pi t)$$

Which can be simplified to:

$$\lambda_k = \frac{1}{k^2\pi^2}$$

λ_k is an eigenvalue for for our orthonormal system. (2.7) can now be rewritten as

$$\frac{1}{k^2\pi^2}\rho_k''(t) = -\rho_k(t) \tag{2.8}$$

This shows that $\{\rho_k\}$ does in fact form an orthonormal system.

Now that we have the orthonormal basis, we can apply one of the eigenfuntions to solve the equation for the Fourier coefficients c_k. The stepwise functions tend to look like a sin function so I choose to use the eigenfunction $\sqrt{2}\sin(k\pi t)$. I calculated these coefficients in the programming package Maple. Recalling that the functions tend to converge to a smooth good representation of the data after only about 12 steps, I only needed to calculate the first 12 Fourier coefficients for each year of data. After calculating these Fourier coefficients, I was able to plug them into (1.8) to get a function that we could really start to work with. A graph of this function revealed that it was a good fit of the data because it had the same approximate shape as the stepwise function. So 12 Fourier coefficients was adequate for what we wanted to accomplish.

3 Brownian Motion

The next and final part of the research project was to do a little research on Brownian Motion.

Brownian motion is a broad topic with many applications in math, physics, finance, and economics. In physics, Brownian motion is used to describe the random movement of particles suspended in a fluid, the mathematical process used to describe the random movements is called the wiener process. It is the wiener process that we are most interested in studying.

Before we can study the Wiener process we must first know what continuous time stochastic processes and Levy processes are. A stochastic process is the opposite of a deterministic system where the outcome is known. In a stochastic or random process there is some indeterminacy in its future evolution described by the probability distributions. This means

that even if the initial condition or starting point is known, there are many possibilities the process might go to, but some paths are more probable than others.

A continuous-time stochastic process assigns a random variable X_t to each point $t \geq 0$ in time. It can also be called a random function of t. The Increments of such a process are the differences $X_s - X_t$ between its values at different times $t \leq s$.These increments can be called independent if we say that the increments $X_s - X_t$ and $X_u - X_v$ are independent random variables whenever the two time intervals do not overlap and any finite number of increments assigned to pairwise non-overlapping time intervals are mutually independent. The increments can be referred to as stationary if the probability distribution of any increment $X_s - X_t$ depends only on the length $s - t$ of the time interval; if the increments have equally long time intervals they are identically distributed.

A levy process is simply a continuous time stochastic process that starts at 0, is everywhere right continuous and has left limits everywhere, and has "stationary independent increments" as described in the previous paragraph.

The wiener process is one of the best known Levy processes and plays an important role in both pure and applied mathematics. There are three important characteristics of the Wiener process W_t. They are:

1. $W_0 = 0$
2. W_t is almost surely continuous
3. W_t has independent increments with distribution $W_t - W_s \sim N(0, t - s)$

The 3rd characteristic needs a little explaining. $N(0, t - s)$ denotes the fact that W_t is a normal distribution with mean 0 and variance $t - s$. The condition that it has independent increments means that if $0 \leq s_1 \leq t_1 \leq s_2 \leq t_2$ then $W_{t1} - W_{s1}$ and $W_{t2} - W_{s2}$ are independent random variables.
One such Wiener process is defined by

$$W_t = \sum_{k=1}^{n} \lambda_k N_k \rho_k(t) \tag{3.1}$$

Where $\lambda_k = \frac{1}{k^2\pi^2}$, N_k are independent identically distributed standard normal random variables, and $\rho_k(t) = \sqrt{2}\sin(k\pi t)$ which is the orthonormal basis found in section 2.

4 Results and Future Plans

I was unable to finish this project in one semester, so hopefully I can finish it next semester. This semester was mostly devoted to researching the theory behind the project. Because of this, I did not have a concrete answer to our original question of whether the

temperature is increasing in Prague. I did feel my research set me up really well to finish the project next semester. My major result by the end of this semester was finding the Fourier coefficients. This will be a great starting points for next semester.
I really enjoyed the project and would like to continue in spring semester. If I am able to continue, I will continue where I left off with the Fourier coefficients. I will use Maple to calculate 12 Fourier coefficients for each of the 215 years of data that I have. Once I have all of the coefficients i can plug them into the equation

$$\left| \sum_{i=1}^{k} c_i \; - \; \frac{k}{n} \sum_{i=1}^{n} c_i \right|$$

Where, k is the number of Fourier coefficients and n is the number of years of data. The resuting equation will give us a smooth equation that holds all the information of the 215 years of data points. I will then be able to run some statistical analysis in order to determine if the temperature in Prague really has been changing over the last 215 years.

References

[1] Buck, R. Creighton (1978). Advanced Calculus Third edition, 307–312.

[2] Breiman, Leo Probability, chapter 12, Wiener and some Related Gaussian Processes, 55.

[3] Brown, Robert, "A brief account of microscopical observations made in the months of June, July and August, 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies." Phil. Mag. 4, 161-173, 1828.

[4] Kleinert, Hagen, Path Integrals in Quantum Mechanics, Statistics, Polymer Physics, and Financial Markets, 4th edition, World Scientific (Singapore, 2004);

[5] Applebaum, David (December 2004), "Lvy ProcessesFrom Probability to Finance and Quantum Groups", Notices of the American Mathematical Society (Providence, RI: American Mathematical Society) 51 (11): 1336-1347,