An Exploration of Barley Reflectance Spectra Using Functional Principal Component Analysis

Logan Q. Sullivan

May 1, 2015

Abstract

Reflectance spectroscopy has been successfully used in plant science to investigate the internal certain internal states of a plant from pigment production to water status. Dr. Kevin Smith of the Agronomy and Plant Genetics Department was interested in using hyperspectral data gathered from barley leaves to make inferences about how efficiently those plants use nitrogen. Through the UMN Statistical Consulting Center, we collaborated with Dr. Smith and his lab to explore how reflectance spectra differ between plants and whether those differences correspond to differences in nitrogen use. Analysis focused on identifying and attempting to eliminate systematic variation in reflectance measurements. Using techniques including functional principal component analysis, we determined that the time of day the measurement was taken strongly influenced the reflectance and this source of variation likely masked any signal in the data. Although we did not answer the initial research question, our work indicated potential problems in hyperspectral reflectance data and allowed us to suggest improvements to this technique for future studies.

Introduction

Plant scientists like those in the Barley Breeding Program work to identify genes that influence specific traits, then breed plants with those different genes together to produce offspring with desirable characteristics. In order to identify which genes influence a trait, geneticists need to analyze many plants with positive and negative examples of this trait in order to see the signal through the genetic noise. For some traits like plant height or color, finding plants to serve as these examples is easy, but for agriculturally useful traits like nutrient use, it can be more difficult, sometimes involving destructive chemical processing.
Nitrogen use is a key component in determining the viability and growth characteristics of plants like malting barley, where protein content is a carefully monitored feature. One goal of plant geneticists is to identify genetic factors that influence a trait in order to favorably manipulate them through breeding. Unfortunately, understanding certain plant processes like this can be difficult, often requiring destructive testing of the plant[3]. It is well known in agronomy that reflected light from leaf canopies can provide insight into the internal state of a plant[5], however, but it is not known whether these measurements can be used to identify plant performance under nitrogen stress. In a previous experiment, hyperspectral reflectance data were gathered from several hundred varieties of barley under different conditions.

Our main problem was determining how to use the high-dimensional data they had gathered to identify differences in nitrogen use without knowing what those differences might look like. We were particularly concerned about how external sources of variation that might pollute whatever trace of a signal was there.

**Consulting Situation**

This project came to us through the Summer Statistical Collaboration Program of the UMN Statistical Consulting Center. In this program, teams of students carry out a consultation with researchers from another department on campus. Dootika Vats, a PhD student in the statistics department, and myself were paired with Dr. Kevin Smith of the Barley Breeding Program and two members of his lab, postdoctoral researcher Mohsen Mohammadi and PhD candidate Celeste Falcon. Dootika and I acted as the lead consultants, taking charge of everything from initial client contact to the final debriefing under the watch of consulting director Dr. Aaron Rendahl.

After receiving the initial proposal and content information from Dr. Rendahl, we arranged an initial meeting with Dr. Smith, Mohsen, and Celeste at the consulting program’s headquarters in Saint Paul. Prior to this meeting, we studied the proposal and pulled some relevant papers from plant science journals to familiarize ourselves with terms and techniques used in this field. The meeting was also video recorded so we could both reference what was said in the meeting and critique our consulting skills.

In the first session, Dr. Smith outlined the goals of the Barley Breeding Program and placed this particular experiment into a broader context of the lab’s research before discussing the specific goals of our analysis. Celeste, who carried out the experiment, then explained the design and data collection. We encountered some difficulties in pinning down the research question we wanted to answer, in part because this analysis was open and exploratory. Understanding how to operationally define nitrogen use and evaluate the success of the
experiment was a significant hurdle.

One tool that served us well in structuring the initial meeting was the POWER process. This meeting organization scheme focuses on setting goals for the meeting before the work is done and summarizing them at the end. The acronym POWER represents the steps of collaboration meetings—prepare, open, work, end, reflect. The ‘open’ and ‘end’ steps were a departure from meeting structures I was familiar with, but I also found them to be the most useful. Explicitly discussing expectations for the meeting felt a little awkward, since the open step essentially is meeting about the meeting and both myself and the client were eager to jump into meatier statistical questions. To keep the discussion on track, I found it useful to introduce the client to the POWER process briefly at the beginning of the meeting so they understand and cordon off that part of the conversation. Setting these meta-meeting goals was extremely helpful because it forced me to have a rigid definition of what would make this client contact successful so I could make sure the conversation covered all the necessary topics. Revisiting and summarizing these goals during the ‘end’ step complemented that very well and helped me determine what needed more attention in future contact.

We decided that the most practical way to collaborate was to work in Dr. Smith’s lab on the Saint Paul campus. Dootika and I travelled there two to three days each week and worked closely with Mohsen and Celeste when they were available. This not only allowed instant feedback on our work and this continuous updating helped us ensure our progress was in the right direction.

At several points during the summer, we compiled progress reports and met with Dr. Smith to discuss the analysis up to that point. These sessions were critical to the success of our work because the openness of the problem and the nature of the data led to quite a few roadblocks, and with Dr. Smith’s guidance we could refocus on issues that were more productive and interesting to his group. For example, in one meeting he encouraged us to explore commonly used vegetation indices in the context of this data—a task that was not part of the original plan, but yielded some interesting and helpful results.

At the conclusion of the session, Dootika and I compiled a final report that catalogued our analysis and findings along with the cleaned, reformatted data and the R code generated during our work. We made these documents available via Dropbox to the entire lab. As a final debriefing, we presented our findings to a group of barley and wheat researchers at a lab meeting.

Overall, Dr. Smith, Mohsen, and Celeste were pleased with the work we did. Although we hit several roadblocks, working so closely with the clients allowed us to redirect our efforts into the most useful, achievable goals. After the project, Dootika and I agreed that we were fortunate to work with a group of scientists who were so understanding of the difficulties
this data posed and who put so much trust into our analysis, even when it didn’t allow us to answer questions in the way we had planned. I remain disappointed that that we could not answer the initial questions posed, but I feel like we respected the data and did with it what we could.

Scientific Background

The University of Minnesota Department of Agronomy and Plant Genetics puts a great deal of effort into understanding how plants grow and, more importantly, how researchers can manipulate them to do better. The Barley Breeding Program is a research group in this department devoted to identifying the genetics of traits in barley plants and manipulating those traits to produce better plants.

Some properties of plants can be difficult to measure. Quantities like total biomass or nitrogen content require harvesting and chemically testing the plant. In some situations this isn’t a problem, but it makes these measurements impossible to take repeatedly. It would be helpful to have a different measurement that could be used as a proxy for the measurement of interest so plants could be easily monitored over time[3].

One idea for such a proxy measurement is spectroscopy. In this technique, light of known wavelength is bounced off the plant and measured when it returns. Using the difference between the input and reflected light, we can understand how the light interacted with matter, which hopefully reveals something about the internal state of the plant[1]. It is well known that spectroscopy can be used to monitor certain aspects of a plant physiology, but these relationships are usually determined by one or two wavelengths. The development of cheaper, more portable hyperspectral cameras, which measure hundreds or thousands of wavelengths simultaneously, could open the doors to more complex and revealing relationships between light and plant biochemistry.

Dr. Smith and his group were interested in using hyperspectral leaf canopy images to identify differences between groups of plants in fields with low and normal soil nitrogen. The research question was rather broad—can we use hyperspectral images of leaf canopies to pick barley plants for further genetic analysis of nitrogen use? Genetic techniques like association mapping that are used to locate a genes corresponding to traits require many different examples of that trait, so to make this selection for nitrogen use, we should identify not only plants that use nitrogen well, but also those that use nitrogen poorly.

During the initial meeting, we identified a few possible approaches to this question. First, we have to remember that we are interested in some abstract knowledge of how well a plant uses nitrogen, but we have no idea how that could translate into a change in the spectrum,
so any technique we use will probably rely on detecting differences between the low and normal nitrogen fields. We could use techniques that reduce the dimension of the data so that the 2048 measured wavelengths can be distilled down to a few important components. Alternatively, we expect some plants to use nitrogen in similar ways to each other, so it could be helpful to identify groups of plants that behave similarly based on their spectra. One of the most difficult parts of this project was deciding how to approach the research question because both of our ideas felt valid, but led us down very different analytical paths.

**Data**

**Design**

Celeste Falcon carried out the data generating experiment in June 2012. Figure 1 shows a visual layout of the two experimental fields—one with normal soil nitrogen and one with low soil nitrogen. Each field was divided into 21 subdivisions each containing 15 plots, and a single genetic lineage, or line, of barley was planted in each of the 315 plots. One line, Tradition, was designated the primary check line and planted in each of the 21 field subdivisions. Four other lines were designated as secondary checks and planted in 6 of the 21 subdivisions. These 6 were selected to have at most one in each sub-column and exactly 2 in each sub-row. The primary check line was always planted in the middle plot of each subdivision, while the secondary checks were randomized. All other barley lines were planted in random locations across the remainder of the field and the randomizations were done independently for each field. The unique lines, primary check, and secondary checks account for 298 plots in each field, the other 17 were not covered by the experimental design. Celeste used these plots to plant 8 commonly studied varieties of barley that she could repeatedly measure at different points during the experiment. This extra information provided critical information on how the spectra changed over time that was not present elsewhere.

**Collection Procedure**

Spectral reflectance measurements were taken on barley leaf canopies using an Ocean Optics Jaz spectrometer with readings at 2048 different wavelengths ranging from roughly 340 nm to 1020 nm by, spaced by roughly one-third of a nanometer. Before certain groups of measurements, the instrument was ‘calibrated’ to a white reference (perfect reflectance) and a dark reference (no reflectance). Reflectance values given are a percentile of incident light.
Figure 1: Experimental field layouts. Colored squares are ‘check lines’ that were repeatedly planted in the field with the exception of orange plots which were left over after the random assignments.

at each wavelength, adjusted to the white and dark references by the formula

$$\frac{R_{raw} - R_{dark}}{R_{white} - R_{dark}}$$

Our first priority was to become familiar with the data, understand its structure, and recognize deviations from that. Celeste had previously aggregated the hundreds of JAZ machine files into two Excel spreadsheets, one for the normal nitrogen field and one for the low, using a Perl script provided by the team that designed the study. We imported these .csv files into R and began to disassemble them into the variables of interest—barley line, collection date, JAZ file number, plot number, wavelength, and reflectance—in order to make the format more regular and workable. First, we removed the 17 columns corresponding to the repeatedly measured plants that were not part of the formal experiment. Those were set aside and processed similarly to the other data. Next, we converted line names to vectors of
lowercase strings to standardize naming, and we briefly considered converting collection date 
to an R-recognized date format, but because date only took two values, we didn’t find it 
necessary. The other four variables were simply numeric and cast as such. This left us with 
one matrix of reflectance measurements with rows and columns indexing wavelengths and 
barley lines respectively, along with five vectors of line names, dates, file, and plot numbers. 
Both the low and normal nitrogen data sets had the same initial structure and were treated 
identically.

Thinking forward, we had some idea that we might want to combine the low and normal 
nitrogen data in some way later in the analysis. We knew that the order of the barley 
lines in the two fields were randomized independently, so it might be worth re-ordering the 
columns so identical lines were in identical positions. This required us to first standardize 
names, uniquely number duplicates, and sort the names, then reorder all the data for one 
field condition based on the other. This required a few attempts, but it made the rest of our 
analysis more straightforward and also confirmed that the correct lines were planted in both 
fields the required number of times.

Later in the analysis, we became interested in not just the dates, but exact times at 
which these measurements were taken. Either the JAZ machine file does not make this 
readily accessible or the Perl script did not output it into the final spreadsheet. Fortunately, 
Celeste had the original JAZ files that were timestamped at creation, so the information was 
available. We manually transcribed the 630 times and added this variable to the analysis. 
This turned out to be a slightly inconvenient, but critical piece of information.

We then checked for missing values, duplicate JAZ or plot numbers, and other unex-
pected data features. The reflectance measurements were all automatically taken by the 
JAZ spectrometer and the other variables were included by the aggregating script, so we 
were not surprised to find that the data was complete. We did, however, find two anomalies. 
One plot/barley line pair was mismatched from the experimental design we were shown and 
one measurement appeared to be taken twice in the original machine files. Celeste again 
referenced her impressively comprehensive field notes, located the transcription error in the 
plot number, and explained that the duplicated file was a result of operator error; the spec-
troscope was moved during the reading, so it was taken again. We identified which file was 
in error and removed it.

Analysis

This left us with clean, orderly data structures that were ready to be explored further. The 
obvious first step was to plot the individual spectra. From the background research we had
previously done, we had some expectation that the spectra would be smooth and vary slowly over the range of wavelengths. We also knew that because the measurements were a percent of total reflectance, they should take values from 0 to 100. The naively plotted data, shown in Figure 2, violated several of those expectations. Most noticeably, at the lowest and highest wavelengths, the reflectance measurements varied wildly and had values well outside the expected range. If we truncated the wavelength measurements and only used a range of around 300-800 nm, these problems resolved. We consulted the spectrometer’s documentation and found that each instrument is fitted with a diffraction grating that determines a ‘best efficiency’ range of wavelengths. This instrument had grating #3, which is intended for capturing visible light from 350-850 nm. This matched with our observations of where the measurements looked well-behaved, but it was comforting to have a more scientific rationale for throwing away portions of the data.

Figure 2: Initial spectra plotted against wavelength for each variety of barley, colored by measurement date.

Figure 3 shows the plots after we removed wavelengths outside of this range. Inside the ‘best efficiency’ range, the spectra are mostly well-behaved. There are still some negative reflectance values around 460 nm, but they are small and could be due to the fact that white and dark references were not taken before every measurement. The more interesting pattern is in the measurement date, which are plotted in different colors. The spectra behave differently for the two days. Near the first large peak at 550 nm, the 6/28/2012 reflectance values are consistently lower in the normal nitrogen field. There is also a double-peaked feature around 750 nm that is stronger in measurements taken that day, especially in the low nitrogen field. These differences directed our attention to consider what factors other
than barley variety might influence the reflectance spectrum.

![Figure 3: Spectral plots after trimming the range to 350-850 nm.](image)

Although we believed many factors like weather conditions, dust particles in the air, and even the temperature of the sensor could influence the spectra, the most obvious sources of variation were measurement date and time. Because the JAZ spectrometer uses sunlight as its light source, we were concerned that this reference spectrum would change throughout the day and cause systematic changes in the input light source which would carry over to the reflectance. The spectrometer was calibrated with a dark and light reference, which would account for changes in the light source, but this was only done a few times during the data gathering. To investigate this, we first needed timestamps for each measurement. As discussed in the analysis section, this information was not initially given to us, but we tracked it down and integrated it into the existing data structures.

We were concerned that both day and time might affect the spectra, but we were unsure if there were enough measurements on different dates overlapping in time to distinguish the two. Unfortunately, as Figure 4 indicates, the measurements on the 28th were taken after 12:49 pm, while those on the 29th were taken before the same time. There was not enough overlapping information to distinguish date effects from time effects, but we wanted to explore one of those effects in more detail. We had a greater range of observations for time and, from a scientific perspective, it seems likely that time would have a stronger effect, so we decided to pursue that further.

At this point we took a step back to think about the research question. We want to make a statement about the factors that affect the reflectance spectra of barley plants. We believe there are some internal factors specific to the variety of barley and some external factors
like measurement time, as well variation due to the individual barley plant. To understand the effect of measurement time, we should hold the other possible factors constant and see how the reflectance changes. To do this, we used data from the 17 plots Celeste measured several times over the 2 days. In Figure 5, we plotted the spectra of Karl and Lacey, two common varieties of barley, at their four different measurement times. The curves show mostly similar patterns with some separation in the exact percentage of reflectance, but there are some worrisome features in the 750 nm range. Early time points, specifically 9:14 am in the low nitrogen field and 10:47 and 12:20 in the normal field, show a double peak that the other times don’t. These are not just the same variety, but the exact same plants measured at different points in the day, so there are definite time effects on reflectance.

We can see a similar trend in Figure 6 by looking at the primary check line, which is measured at more time points. This does introduce variation due to the individual plant, but since this is essentially error, it should be small compared to any signal, or else we can’t do meaningful analysis. These plots also show large differences in the overall shape of the spectra, but they don’t seem to change smoothly over time, so the interpretation is difficult. However, this does support our suspicion that time plays an important role in the reflectance spectra of a plant. This complicates the analysis since, as Figure 4 shows, there are no overlapping time points between the low and normal fields that we could match.

After identifying this problem, we devoted much of our remaining time to attempting to fix it. These efforts were ultimately unsuccessful, but revealed some other interesting patterns in the data. We first took a different perspective on the data—instead of thinking of a spectrum for one barley wavelength (taken at one time) varying over a range wavelengths,
we considered the measurements at an individual wavelength varying over time. We noticed that the times in which measurements were taken could be grouped into a few ‘chunks’ between which Celeste saved the data, let the spectrometer cool down, and did both white and dark balancing before continuing more data collection. Some of these ‘chunks’ of time appeared to systematically change over time. Because these time points represent different varieties of barley, we can’t definitively say the change is due to time, but the trend certainly suggests it. The changes inside of each ‘chunk’ appear
to be a drift in the reflectance measurement over time. We decided to fit a polynomial and capture the residuals of each ‘chunk’ at each wavelength to adjust for this trend, as shown in Figure 7. Even though this seemed to correct for some measurement drift, especially in the low nitrogen field measurements on 6/29/2012, the problematic differences resulting

![Figure 7: Quadratic fits on each ‘chunk’ of measurements taken at the same time.](image)

Even though we weren’t able to answer the research questions as we had initially planned, during our time learning about this data, we became interested in some other spectral techniques used to evaluate plants. Vegetation indices are functions of one or more wavelengths that claim to indicate something about the internal state of a plant. Many different vegetation indices with specific applications are used in plant science, and seeing the effect of measurement time on whole spectra made us curious about whether these faire better.

Since we were interested in seeing these indices over many time points as possible, we used the primary check line, which was the “Tradition” variety of barley planted in 21 locations in each field. We then computed several common vegetation indices on this data, including one that is intended to measure nitrogen use. The plots in Figure 8 demonstrate how two indices change over time for this particular variety of barley. The two indices shown are Green-Normalized Difference Vegetation Index (GNDVI), which is often used for general vegetation monitoring from aerial or satellite images, and Normalized Difference Red Edge (NDRE), which claims to report nitrogen status. Neither index appears to perform well over the range of measurement times. The GNDVI ranges from nearly 0, which normally indicates bare soil or rock, to 0.3, which is consistent with green vegetation. More concerning, a one
hour difference in measurement time can mean the difference between some of the lowest values or the highest.

![Figure 8: GNDVI and NDRE vegetation indices for the Tradition primary check line plotted over time.](image)

We briefly considered trying to use the spectral data from these hundreds of barley plants to create a new vegetation index that would help identify nitrogen stressed plants. However, the techniques we hoped to apply to this problem would have returned at best linear combinations of wavelengths, while indices in the literature are usually ratios of linear combinations. The space of ratios of linear combinations of wavelengths is much larger and more difficult to search and even then there is no guarantee that format is the best way to construct an index. It would be interesting to use hyperspectral data to find new vegetation indices, but that hinges on efficiently searching very large function spaces.

**Functional Data Analysis**

Data like this that vary over a continuum suggest a different analytical approach. We could think of each observation from a plant as a set of 1449 unrelated responses, one for each wavelength. But, far from being unrelated, the reflectance values for nearby wavelengths are so correlated they look more like a smooth curve than distinct points. Functional data analysis is equipped to handle this situation where observations can be described as curves or surfaces.

**When is data functional?**

It may be helpful to treat discrete data as smooth, but when is it ok to make that leap? If we assume there is a smooth generating function for our data, we would expect the data to be
nearly smooth. But, formal definitions of smooth functions that require derivatives of every order make little sense when applied to discrete data. We can approximate derivatives by sequential differences, but because of noise, these plots become uninterpretable very quickly. Figure 9 shows the first two approximate derivatives for one barley line. The plots aren’t very helpful in demonstrating that the data is functional, but the combination of the visually smooth initial reflectance plot and the intuition that reflectance should change smoothly over a range of wavelengths suggest that it’s close enough.

Figure 9: Approximate derivatives for one low-nitrogen Tradition line.

**Functional Principal Component Analysis**

In traditional principal component analysis (PCA), the goal is find a set of linear combinations of the initial variables that are mutually orthogonal and maximize the variance of the projection onto those components. Similarly, functional principal component analysis (FPCA) attempts to find orthogonal component functions that capture as much variation as
possible. FPCA is preferable to standard PCA in high-dimensional settings, especially when the number of variables is much larger than the number of samples. In those situations the sample covariance matrix isn’t a good estimate of the population covariance matrix because individual instances are more spread out in $p$-dimensional space[6]. But by treating the data as functional, the assumption of smoothness tells us something more about how the dimensions are related so we can use fewer of them to represent the data.

In FPCA the data and principal components are now functions, so we need to extend the concept of projecting the data onto a component so we apply our usual definition of variance and do some maximization. If we have a data function $x(s)$ and a component function $\xi(x)$, we define the principal component score $f$ as the continuous inner product of those functions, $\langle \xi(s), x(s) \rangle = \int \xi(s) \cdot x(s) \cdot ds$.

Applying this to the standard PCA framework, the first functional principal component is a function $\xi_1(s)$ that maximizes the variance of the principal component scores for the data functions $x_i$

$$\frac{1}{N} \sum_{i=1}^{N} f^2 = \frac{1}{N} \sum_{i=1}^{N} \langle \xi_1(s), x_i(s) \rangle^2 = \frac{1}{N} \sum_{i=1}^{N} \left( \int \xi_1(s) \cdot x_i(s) \cdot ds \right)^2$$

with the constraint $\|\xi_1\|^2 = \int \xi_1(s)^2 \cdot ds = 1$.

Higher components involve a similar maximization

$$\frac{1}{N} \sum_{i=1}^{N} \langle \xi_k(s), f_i^{k-1}(s) \rangle$$

where the new data functions accounting for previous components are recursively defined by

$$f_i^k(x) = x_i(s) - \sum_{j=1}^{k-1} \langle \xi_j(s), x_j(s) \rangle \cdot \xi_j(s)$$

These components also have the added orthogonality condition $\int \xi_j \xi_k = 0, j < k$.

Computing these maximizations is not as straightforward as in the discrete case. Ramsay and Silverman outline three common methods. The earliest approach was to treat the data function $x(s)$ as a grid of evenly spaced discrete values, $x_i, i = 1, \ldots, N$ and use standard PCA, then perform some interpolation to go back from discrete principal components to smooth functions. Next, we could use numerical integration to approximate the solution, but this also requires interpolation to get a smooth function. Most FPCA today is done by representing the data with basis functions before doing standard PCA on the coefficients for those bases. This still gives discrete principal components, but along with the basis functions, we can reconstruct basis-approximations of the functional principal components. Even bet-
ter, as long as the basis functions are smooth, so are the resulting functional components[4]. Because this technique is so common and basis functions are so understandable, we will use this technique to explore the barley data.

Basis Representations

Basis representations express one function in terms of a number of other functions that, as the name implies, form a basis for some function space. Every function in that function space can be represented as a linear combination of those basis functions much like analogous basis vectors that span a vector space. However, function spaces (at least spaces complex enough to be useful on real data) are considered to be infinite dimensional, so there is often no finite set of bases that will reconstruct the true function.

We don’t have a smooth data function, just discrete observations of it, but we can reconstruct those points perfectly with at most \( N \) basis functions, the number of observations. If we use fewer bases, \( k < N \), then our reconstruction of the data is an imperfect approximating function, often called a smoother. The choice of \( k \) determines how closely that smoother follows the data. Let \( \phi = (\phi_1, \phi_2, \ldots) \) be a set of functions that form a basis for the function space in which we believe our true data function lies. Then we can represent the smoothed data function as a linear combination of \( k \) of those basis functions

\[
x(t) = \sum_{i=1}^{k} c_i \phi_i(t)
\]

But to do this, we first need to pick a basis system. Constructing a basis can be a very complicated task because the properties of the basis are very important in determining the theoretical and computational properties of FPCA. There are two commonly used bases—Fourier and B-spline—that fill most needs. The Fourier basis follows from Fourier analysis’ ability to represent functions as a linear combination of trigonometric functions. The well-studied Fourier transformation algorithms are efficient at computing the basis coefficients when the data has certain properties, but the bases are periodic, and so are the data reconstructions and principal components. In some situations this is a desirable, natural property for the model, but our reflectance spectra are clearly not periodic over our range of wavelengths, so we will use B-splines.

B-Spline Basis  Splines are piecewise functions that are polynomial on each piecewise interval, but have continuous derivatives of several orders where those intervals meet. Splines are defined by an order \( m \) and a sequence \( \tau \) of knots, also called breakpoints, that divide the domain into intervals. Each interval contains a polynomial of order \( m \) and, to ensure
the spline function is sufficiently smooth at the knots, polynomials in adjacent intervals have matching derivatives up to order \( m - 2 \) at these knots.

B-splines are a way to generate a basis using splines. The individual basis functions are splines defined by a common order \( m \) and knot sequence \( \tau \), with the added property, called the compact support property, that each basis spline is nonzero on at most \( m \) adjacent subintervals. This makes B-splines an efficient basis system because for most computation only care about those non-zero intervals. Figure 10 shows an example B-spline basis\[^4\].

Unlike the Fourier basis, B-splines aren’t orthogonal, but this isn’t a problem for FPCA because we only need the basis functions to turn a functional problem into a standard multidimensional one. Orthogonality does have computational benefits, but B-splines are competitive because of the compact support property.

![Figure 10: Example B-spline basis with 10 B-splines of order 4.](image)

**Basis Reconstructions of the Data**

Now that we have a basis system selected, we need to compute the coefficients of those bases for each of the data functions. Again, there are many ways to do this, but the simplest way is to use ordinary least squares regression to minimize the squared distance from the basis expansion to the true values at those points.

Let \( x(t) = \sum_{k=1}^{K} c_k \phi_k(t_j) = c^T \phi \) be the basis expansion and \( \Phi \) be a \( n \) by \( K \) matrix containing \( \phi_k(t_j) \), the values of the basis functions evaluated at the data points. Then we want to find

\[ 17 \]
the $c$ that minimizes
\[ \sum_{j=1}^{n} \left[ y_j - \sum_{k=1}^{K} c_k \phi_k(t_j) \right]^2 = \| y - \Phi c \|^2 = (y - \Phi c)^T (y - \Phi c) \]

Taking the derivative with respect to $c$ and setting it equal to zero, we find the solution
\[ \hat{c} = (\Phi^T \Phi)^{-1} \Phi^T y, \]
just as we’d expect for this problem. Recalling the assumptions of OLS regression, this smoothing assumes that the errors in the model $y_j = x(t_j) + \varepsilon_j$ relating the data to the basis function values are independent and identically distributed, have mean zero and some constant variance $\sigma^2$. Assumptions are always important and should be checked, but keep in mind that this is just a means to the end of obtaining coefficients for the basis representation, and it would be easy to get lost in a sub-problem that might not drastically affect the end goal.

This simple example suggests that many of the extensions of regression would extend to smoothing. Indeed, things like weighted least squares and regularization are used to improve fits.

Computing Functional Principal Components from the Basis Expansion

Now that we have defined a basis system and computed the coefficients to reconstruct the data functions, we can return to the actual problem of finding the functional principal components. The traditional PCA problem can be stated as repeatedly finding a vector that maximizes the variance of a projection or equivalently solving an eigenequation involving the covariance matrix of the data. Similarly, the FPCA problem discussed previously is equivalent to solving an eigenequation involving the variance-covariance operator
\[ \int v(s, t) \cdot \xi(t) \cdot dt = \rho \xi(s) \]

[2] The integral makes this difficult, but discrete eigenproblems are better-understood, and the basis expansion allows us to turn one into the other. Suppose the data function $x_i(t)$ has an expansion on basis functions $\phi = (\phi_1, \ldots, \phi_k)$ with coefficients $c_i = (c_{i1}, \ldots, c_{ik})$, then
\[ x_i(t) = \sum_{k=1}^{K} c_{ik} \phi_k(t) \]
If we write the coefficients as an $N$ by $K$ matrix, we can express all $N$ data functions at once as

$$x = C\phi$$

this allows us to write the variance-covariance function in terms of the basis functions as well

$$v(s, t) = \frac{1}{N} (C\phi(s))^T (C\phi(t)) = \frac{1}{N} \phi(s)^T C^T C\phi(t)$$

Further suppose the eigenfunction $\xi$ has its own basis representation

$$\xi(t) = \sum_{k=1}^{K} b_k \phi_k(t) = \phi(t)^T b$$

now we can rewrite the eigenequation we want to solve in terms of the basis functions

$$\int v(s, t) \cdot \xi(t) \cdot dt = \rho \xi(s)$$

$$\int \left( \frac{1}{N} \phi(s)^T C^T C\phi(t) \right) (\phi(t)^T b) dt = \frac{1}{N} \phi(s)^T C^T C \left( \int \phi(t) \phi(t)^T dt \right) b = \rho \phi(s)^T b$$

Call $\int \phi \phi^T = W$, a symmetric matrix of order $k$ whose entries are the inner product of pairs of basis functions $w_{k_1, k_2} = \int \phi_{k_1} \phi_{k_2}$. This supports using an orthonormal basis, because then $\int \phi_{k_i} \phi_{k_i} = 1$ and $\int \phi_{k_i} \phi_{k_j} = 0$, $i \neq j$, so $W = I_k$, the order $k$ identity matrix. B-splines also have a benefit here because, if we have to resort to numeric integration, at least the basis functions are only nonzero on small intervals.

This simplifies the equation to

$$\frac{1}{N} \phi(s)^T C^T C W b = \rho \phi(s)^T b$$

which must hold true for any basis function $\phi(s)$, so the continuous eigenproblem becomes the discrete

$$\frac{1}{N} C^T C W b = \rho b$$
let \( u = W^{\frac{1}{2}} b \), so \( W^{\frac{1}{2}} u = W b \)

\[
\frac{1}{N} C^T C W b = \rho b \\
W^{\frac{1}{2}} \left( \frac{1}{N} C^T C W b \right) = W^{\frac{1}{2}} \rho b \\
\frac{1}{N} W^{\frac{1}{2}} C^T C W^{\frac{1}{2}} b = \rho W^{\frac{1}{2}} b \\
\frac{1}{N} W^{\frac{1}{2}} C^T C W^{\frac{1}{2}} u = \rho u
\]

Working backwards from an eigenvector \( u \), the associated coefficient vector \( b = W^{-\frac{1}{2}} u \) corresponds to the chosen basis representation of that functional principal component[4].

**Applying FPCA to the Barley Data**

The derivation shows that actually doing FPCA is much less intimidating than it appears, but, thankfully, the authors of *Functional Data Analysis* wrote the associated R package *fda* to handle many of the techniques covered in their text. All we need is a matrix of data where columns are observed data functions, in this case reflectance spectra for individual barley varieties, and rows are argument values inside the data functions, here, reflectance values at each wavelength.

First, we decide on a suitable set of basis functions to represent our data and create it with `create.bspline.basis`. Then, we can center our data functions around zero by subtracting the mean for each barley variety and compute the matrix of coefficients that reconstruct our data functions with the basis system. The function `smooth.basis` handles the previously discussed OLS method, as well as weighted least squares, penalization. There are also functions to handle constraints discussed in the text for monotone or non-negative data[2]. Figure 11 compares the spectral plots two observed barley lines with their reconstructions on 50 basis functions.

These basis functions and coefficients are used to create a functional data object that the `pca.fd` function can use to generate the eigenvectors corresponding to the eigenfunctions in basis form, the eigenvalues, and the component scores of each observation for those eigenfunctions. Figure 12 shows a plot of the ordered eigenvalues and a Scree plot of the proportion of variance explained that can help us determine how many eigenvalues to use. The eigenvalue plot indicates the first 7 eigenfunctions might be important, but a check of the proportion of variation shows that the first 3 account for 99% of the variability; that is good enough.

Figure 13 shows those first three eigenfunctions plotted as perturbations of the mean-
small multiples of the component functions added and subtracted to the mean function. The strongest component shows variation over most of the spectrum, but more near the end of the measured spectrum, which agrees with our visual analysis that the range 750-850 nm had the most variability in reflectance. Interestingly, the first component seems to pick up the 750 nm double-peaked feature that we initially recognized as resulting from measurement time. Even though they seem to back up our inspection of the data by showing us where variation is highest among the reflectance curves, these plots don’t seem to provide much insight into the research question.

Recall the principal component score \( f = \langle \xi(s), x(s) \rangle = \int \xi(s) \cdot x(s) \cdot ds \). This scalar gives us a way to numerically evaluate how a data function is behaving for a given eigenfunction.
Figure 13: First three eigenfunctions for the low nitrogen barley reflectance.

without trying to visualize the functions themselves. Figure 14 shows all 298 barley varieties plotted by the first two component scores. There barley varieties definitely separate into rough groups, and this clustering might be useful in identifying barley varieties that are different enough to warrant further genetic analysis. However, we suspected that measurement
time had such a large effect on the reflectance spectrum that it could render those statements meaningless.

The varieties in the factor score plot in Figure 14 are grouped by the day on which the measurement was taken and colored by time. Plants measured at similar times have similar factor scores. Furthermore, the progression from one color to the next appears to follow a continuous path on the plot counterclockwise from the upper left quadrant, suggesting that the factor causing this pattern works continuously over time. There is also a strong separation by measurement day, suggesting that both of these factors could be important, even they are still indistinguishable in this experiment.

Figure 14: Factor scores for the first two eigenfunctions grouped by measurement day colored by measurement time.

Results

After Figures 5 and 6, we hypothesized that time was having some effect on the reflectance, but as strongly as we felt, resorting to pointing at differences in the data functions felt
anecdotal at best. Functional principal components provided us with a more formal way to examine the differences between our spectra and the resulting factor score plot in Figure 14 is more convincing evidence in favor of our hypothesis.

Unfortunately, this discovery means that the research question we were trying to answer about finding differences between barley lines based on nitrogen use is probably obscured by other sources of variation in this experiment. But the success of FPCA in identifying time as a problem makes me optimistic that, with cleaner data and some careful combination of the low and high nitrogen data sets, it might also be able to identify plants that use nitrogen differently.

Although we weren’t able to discover wavelengths or indices that can help identify nitrogen use, we made progress towards understanding the application and limitations of these methods, and from that we proposed a few changes. We recognize measuring every plant simultaneously is impossible in large experiments with hundreds of varieties of barley in multiple fields, so we set 5 goals for future experiments.

1. Take measurements in a short time span around solar noon. No more than one hour before and after if possible.

2. Perform white and dark balancing as often as time permits, but especially when there is a break taken to save data or let the machine cool down.

3. Pay careful to the conditions on different measurement days. Ideally, gather data in one day, but if it must be split up, then cloud cover, airborne dust, and other environmental changes affecting sunlight should be minimized. There was insufficient data to fully explore this, but we would expect an effect.

4. Match measurements on the same barley varieties for time as closely as possible so time effects are minimized between identical varieties of barley in different conditions. We understand that in current experiments, planting order is randomized in every field. While randomization is a key aspect of many experiments, we believe that one random plan is sufficient and it should be duplicated between the field conditions to facilitate matching measurement times.

5. Finally, if at all possible, collect at least twice with an overlap of time for each line. This would help account for variation due to collection day, and also provide repeated measures on all lines.

In the near future, it may become common to gather this type of hyperspectral data with drones. This could give researchers the ability to quickly survey an entire field worth of
plants and possibly eliminate concerns about measurement time. However, drone collection probably has its own set of problems including atmospheric conditions, hover height, exposure time, and selection of regions of interest. This is definitely an interesting avenue to explore for future analysis, but we don’t consider it a solution, just a different, hopefully easier, problem.

Figure 15: Example drone-collected field image.

References


