Finding Structure in Multivariate Time Series

by

Jeremy L. Praissman

(Under the direction of Andrew Sornborger)

Abstract

The scope of scientific data collection in modern projects such as the human genome project has made it effectively impossible for careful by-hand analyses of such data to be carried out. Simultaneously, the increase in computer power raises the possibility of replacing human scrutiny with computer systems that could effectively sort and filter copious data, presenting only the most salient features to researchers. This thesis details a method for combining a generalized version of the classical statistical method known as canonical correlation analysis, that possesses good computational properties, with the more recently developed multitaper spectral estimators. The developed method allows researchers to combine data from multiple experiments to generate more accurate spectral decompositions of the underlying processes involved while also giving researchers a sensitive method for finding the links between variables in the data sets. The only limitation is that the data to be analyzed must be homogeneous in certain specific ways (for example, it must contain no pronounced trends).

Index words: Generalized Canonical Correlation Analysis, Multitaper Spectral Estimation, Singular Value Decomposition

Finding Structure in Multivariate Time Series

by

Jeremy L. Praissman

B.S., Carnegie Mellon University, 2001

A Thesis Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Master of Arts

Athens, Georgia

2007

Finding Structure in Multivariate Time Series

by

Jeremy L. Praissman

Approved:

Major Professor: Andrew Sornborger

Committee: Malcolm Adams
Edward Azoff

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2007

Acknowledgments

I would like to acknowledge the contributions made by many of the faculty at the University of Georgia to my education. In particular, I would like to thank Malcolm Adams for rekindling my interest in mathematics and (in memory) David Galewski, for his encouragement and wonderful introductory course on abstract algebra (especially his notes). More recently, I have had the pleasure of working with Andrew Sornborger, whom I would like to thank for stimulating discussions, his patience and his dedication. I would like to thank Ed Azoff for always being available to discuss mathematics and life, and Robert Varley for his encouragement, openness to all mathematical ideas, willingness to discuss those ideas and for the wonderful job he did as the graduate coordinator. Finally, I would like to thank Ming-Jun Lai for his encouragement, for teaching a very good course on numerical analysis, and for helping me to obtain a very enjoyable NASA internship.

Table of Contents

Chapter 1

Introduction

The widespread availability of computers as well as the nature of many forms of scientific inquiry today presents new challenges as well as possibilities. It is essential in this setting for techniques to be developed and implemented for dealing with the onslaught of data demanded by the types of analyses being undertaken. Numerous techniques are available although many techniques are presented in specific settings that, if we confine ourselves to, place unnecessary limitations on the analyses that may be performed. Combining powerful techniques from across the mathematical spectrum in order to help researchers cope with the pace of the accumulation of data and, at the same time, to improve the fineness of their results, is of some importance and it is from this point of view that the following work derives.

The mathematical approach taken specifically within this work is the combination of the statistical technique known as "canonical correlation analysis" with another statistical technique called "multitaper spectral estimation" to form a technique we call "spectral canonical correlation analysis." The types of experimental data we wish to be able to handle are described, mathematically, by "stationary stochastic processes" and it is the case that a certain type of spectrum is a good description of these processes. Further, we will see that multitaper spectral estimation allows us to obtain good estimates of the actual spectrum of any given stationary stochastic process from realizations of that process (data). The estimated spectra of any given set of variables (say from data set A) may then be compared to the estimated spectra of the variables from other data sets (data sets B,C,...) using canonical correlation analysis. This process both picks out variables with similar spectra across

the data sets and also gives a sort of averaged estimate of what the common spectra are across the data sets[1]. In particular, if the same experiment is performed multiple times, this averaging should result in better estimates of the spectra of specific underlying processes.

[1]These "canonical" spectra are different from normal spectra and may include negative values: they should really be understood in a vertical translation invariant manner as indicating which frequencies are being accentuated and which removed in the corresponding transformed data sets, see below.

Chapter 2

Statistical Background

2.1 Introduction

One of the primary ways mathematics interfaces with the physical world is through measurements. Given an appropriate numerical scale and a method for assigning a number from that scale to a specific attribute of a physical object, we may summarize (physical) attributes concisely and in a form amenable to treatment by the methods of mathematics. For example, wealth may be measured in dollars, distance in miles and electrical potential in volts. This is particularly crucial for the concise specification of relationships between attributes (in the form of formulas, ranging from the fact that one particular skyscraper may be taller than another to the specification of the electrical potential between two points in a circuit as it depends on the amount of electricity flowing through the circuit and the resistance of the circuit). For various reasons, however, this approach does not work out as cleanly as the description above might suggest. Measurement error, the fact that even reproducible experiments may give somewhat different results each time they are carried out, the desire to have mathematically tractable models and the importance of human interpretability of the results leads to quite a lot of complication on top of the simple initial idea. The methods of statistics may be used to deal with the questions of measurement error, differences in reproducible experiment outcomes and human interpretability. We will primarily deal with the last two here without much consideration of the issue of measurement error.

There are two primary issues that arise in handling measurements taken from the physical world. The first issue is due to the sheer quantity of measurements of even a single attribute that may be taken. A data set containing the carbon dioxide concentrations of

the atmosphere recorded across the United States on any given day might contain thousands of data points[1]. This leads naturally to the further summarizing of the data itself by "measuring" attributes of the data such as "central tendency" often clarifying the relationships between different sets of data both objectively and subjectively (i.e., from a human interpretability standpoint). Summarizing data in this manner is the purview of *descriptive statistics*. The relevant descriptive statistics for this work will be means, covariances, correlations and spectra. The second issue arises from the variability of measurements we may record due to differences in the state of the system we are drawing measurements from over time as well as measurement error. Using the carbon dioxide concentration example, it may be that if the measurement recorded on a given day had been recorded an hour later, the concentration would have been found to be nearly double that which was recorded at the time the sample was taken (for example if there were a fire nearby and the wind changed direction). This is an example of uncertainty arising in the form of *sampling error* and, in order to robustly characterize and analyze measurements, some accounting of this uncertainty must take place. This naturally falls into the realm of probability theory which, when applied to problems of this nature, forms the backbone of *inductive statistics*. Methods from inductive statistics will be used to help ensure that our estimates of the various descriptive statistics we are interested in are sound[2].

By way of example, consider the problem of determining the height of a young tree. If twenty different people measure the tree using the same measuring tape and record their measurements in a table, there may well be twenty different numbers recorded in the table. One way of looking at the numbers in the table is as a relative frequency distribution, an essentially probabilistic point of view. We can then ask, for example, for methods of combining the measurements in the table to get an estimate that is more likely (than any one of the individual measurements) to be close to the actual height of the tree. The process

[1]Information from the Data Assimilation for the Carbon Cycle workshop held at MSRI in 2006.

[2]Confidence bounds, estimates of how good a descriptive statistic is (does it tend to be closer to the "correct value" than others...), etc., provide examples of inductive statistics in action.

of combining the measurements is the process of computing (a "realization" of) a descriptive statistic while the desire that the computed estimate be more likely to be close to the actual height is a statement about the reduction of uncertainty in the computed estimate (an example of inductive statistics). The multitaper spectral estimation technique that will be detailed later works (statistically) in an exactly analogous manner by essentially splitting up signals (the results of which correspond to the distinct measurements recorded by different people in the tree example) and combining estimates based on these statistically distinct measurements to reduce uncertainty and variability in the overall estimate.

The remainder of this chapter will alternate between more casual descriptions of the important concepts necessary for the rest of this work and the introduction of mathematical language and notation implementing these concepts precisely. Specifically, we begin with the basic mathematical development following from the concepts already mentioned which is needed for the following chapter. In the second half of this chapter, we outline and fill in the development needed for the spectral analysis chapter. The following section assumes knowledge of the mathematical definitions of random variable, sample space, etc. from probability theory.

2.2 Mathematical Setting (Probability Theory)

2.2.1 General Notation

The greek letter Ω will generally be used to represent a (usually abstract) probability space and ω a sample point in such a space. A (real) *random vector* $\boldsymbol{X} : \Omega \to \mathbb{R}^n$ is a real vector-valued function on some sample space Ω, the coordinate functions of which are simply (real) random variables on Ω. A (real) *random matrix* is defined analogously. Random vectors and matrices will be distinguished from random variables in part by the convention of bold-facing letters that name vectors. Random vectors/matrices and random variables are distinguished from fixed matrices and constants by the convention of using only upper-case italicized

symbols for probabilistic elements (fixed matrices and vectors will be denoted by upper-case, bold-faced but non-italicized symbols). Of course, any statement made about random vectors will usually apply in the case when $n = 1$, that is, to a real-valued random variable, and it will often be the case that we will not explicitly distinguish random variables, random vectors and random matrices when defining terms and introducing notation (preferring to use just the term random vector). In particular, a random vector may contain a mixture of coordinate functions including an arbitrary selection of the coordinates of, say, a random matrix previously defined, along with explicitly named random variables etc. The notation $P(\cdot)$ will be used for probability measure and $E\{\boldsymbol{X}\}$ will be used for the expected value of $\boldsymbol{X}$ which for vectors would be a vector consisting of the expected values of each coordinate random variable. Realizations of random variables and vectors may be written using a lower-case version of the letter naming the random vector or variable, e.g., $\boldsymbol{x}$ is a realization of the random vector $\boldsymbol{X}$ which is simply $\boldsymbol{X}(\omega)$ for some $\omega \in \Omega$. Finally, covariances may be written $\sigma_{X,Y}$ and correlations $\rho_{X,Y}$ (and the shorthand $\sigma_X = \sigma_{X,X}$ for the variance of X). Subscripts may be dropped when there is no ambiguity.

2.3 Statistics

As touched upon in the introduction, the process of defining and computing a (descriptive) statistic is a method of data reduction. Mathematically, given a set of random vectors $\{\boldsymbol{X}_t | t \in 1, \ldots, n\}$, a *statistic* is a random vector defined by $\boldsymbol{T}_n = g_n(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ where g_n is an appropriate (deterministic) vector valued function (in particular, there must be one mathematical function for each n). A given statistic is often defined in such a way as to give information about a specific parameter of the underlying distribution of its input random vectors. For example, the mean $E\{X\}$ of a random variable is a parameter of its distribution. The sample mean is a statistic where $T_n = g_n(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} X_i$ and

$X_1, X_2, \ldots, X_n$ are independent identically distributed random variables[3]. The greek letter $\boldsymbol{\theta}$ will be typically be used to represent a fixed (vector) parameter of a probability distribution. In this case, the data reduction role of a statistic comes from the idea that estimates and inferences about a specific parameter $\boldsymbol{\theta}$ will (then) be based solely on realizations of $\boldsymbol{T}_n$ rather than the entire set of observations. In the case where $\boldsymbol{T}_n$ is defined specifically to be used to provide a single fixed numerical estimate of a given parameter $\boldsymbol{\theta}$ of the underlying distribution, $\boldsymbol{T}_n$ is called a *point estimator* of that parameter. As $\boldsymbol{T}_n$ is a random variable, in general, realizations of it will not be equal to the parameter of the distribution it is an estimator of. So the term *sampling error* is used to refer to the difference between $\boldsymbol{t}_n$ (some realization of $\boldsymbol{T}_n$) and $\boldsymbol{\theta}$. In the univariate case, this is usually quantified by $T_n - \theta$, in the multivariate case it may be quantified by matrix or vector norms.

With these notions in hand, we can now set out to specify properties of a point estimator that are desirable. These properties are generally concerned with keeping sampling error low both on average and minimizing its effect for a given realization. In terms of keeping sampling error low on average, we will typically work with estimators that are *unbiased* and *consistent*. An *unbiased* point estimator of a parameter $\boldsymbol{\theta}$ has the property that for fixed n, $E\{\boldsymbol{T}_n\} = \boldsymbol{\theta}$. This means in a strict sense that if we average the estimates of $\boldsymbol{\theta}$ provided by $\boldsymbol{T}_n$ over all possible samples of size n (with n fixed, each taken once), we will obtain the correct result $\boldsymbol{\theta}$ and in a looser sense that, under random sampling, as the number of samples of size n taken increases, the average of the $\boldsymbol{T}_n$'s calculated from those samples tends to $\boldsymbol{\theta}$. A *consistent* point estimator of a parameter $\boldsymbol{\theta}$ has the related property that as the size of a given single sample increases, the estimates of $\boldsymbol{\theta}$ given by $\boldsymbol{T}_n$ tend toward $\boldsymbol{\theta}$ in probability. Minimization of the effect of sampling error due to use of any specific given realization is generally characterized by the variance of the estimator in question. It is desirable to have an estimator that is both unbiased and has minimum variance. Such an estimator is said to be *efficient* and in practical terms, may be thought of as typically giving an estimate close to the true value for

[3]We will typically omit n in the names of statistics, except when necessary to indicate properties of statistics in the abstract.

any given individual sample (for example, it is possible to give precise confidence bounds on the closeness of an unbiased estimator in general based on its variance using Chebyshev's Theorem, this may be sharpened considerably when information is available as to the type of the true distribution). Demonstrating that an estimator has minimum variance over all possible estimators of a given parameter is somewhat less straightforward than checking whether or not the estimator is unbiased and consistent. Interestingly, it turns out to be possible to demonstrate an exact lower bound for the variance of a given estimator, without reference to a specific distribution, under certain mild regularity conditions[4]. We will be content to indicate that the estimators we are interested in are fairly efficient with respect to this lower bound.

2.4 Background for Canonical Correlation Analysis

In this section, we define the estimators underlying canonical correlation analysis and justify their use based on the preceding discussion. We then provide some discussion to aid in understanding and recalling the properties and meaning of correlation(s).

2.4.1 Estimators for Canonical Correlation Analysis

To fix notation for this section, let $\boldsymbol{X}$ be a random vector and let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ denote realizations of this random vector. Let $\boldsymbol{\mu} = E\{\boldsymbol{X}\}$ be the mean vector of $\boldsymbol{X}$ and $\boldsymbol{\Sigma} = E\{\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})'\}$ be the covariance matrix of $\boldsymbol{X}$ (the notation $'$ indicates transpose). Define the *sample mean vector*[5] for such a set by $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x_i}$ and the *sample covariance matrix* for such a set by $\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x_i} - \bar{\boldsymbol{x}})(\boldsymbol{x_i} - \bar{\boldsymbol{x}})'$. The following section will detail the desirable properties that these estimators possess.

[4]The Cramér-Rao bound, see Anderson [2], p. 86 and Panik [14], p. 387

[5]Note that the term sample mean, for example, refers both to the statistic $\bar{\boldsymbol{X}} = \sum_{i=1}^{n} \boldsymbol{X}_i$ where the $\boldsymbol{X}_i$ are independent identically distributed random vectors and also to any realization of this statistic, $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x_i}$.

The sample mean vector and sample covariance matrix are each unbiased and consistent estimators of their respective parameters. Further, the (relative) efficiency of $\bar{\boldsymbol{X}}$ and $\boldsymbol{S}$ when $\boldsymbol{\theta}$ includes both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is $[(n-1)/n]^{p(p+1)/2}$ (Anderson [2], p. 85). These properties generally justify the use of the sample covariance matrix in the classical canonical correlation analysis computation (it is also possible to justify precisely the use of a slightly scaled version of S, $[n/(n-1)]S$, when normally distributed random vectors are used, see Anderson [2], p. 103 and p. 500).

2.4.2 Correlation

We provide here a little background on correlation to aid in understanding the operation and effect of canonical correlation analysis. Two viewpoints helpful in understanding most classical statistical techniques and notions (including sample correlation and canonical correlation analysis) are the linear algebraic and/or geometric viewpoint, and the probability theoretic or measure theoretic viewpoint (involving the form and analysis of functions). The more geometric side will be covered later although we will point out now that saying that random variables are uncorrelated is equivalent to saying that they are orthogonal in $L^2(\Omega)$. We focus here on a useful tool of measure theory that may be used to obtain additional intuition as to how sample correlation ranges as a (data) sequence ranges with respect to another and dually what we can expect of sequences depending on their correlation. This tool is Chebyshev's inequality. Chebyshev's inequality is a generalization of the intuitive statement that if n positive numbers sum to s, then for any number $c > 0$ we choose, only $\lfloor s/c \rfloor$ of the numbers may be greater than or equal to c (or equivalently the proportion $\lfloor s/(cn) \rfloor$ of the numbers). For our purposes, Chebyshev's inequality constrains the number of measurements that may be any given distance from the mean, most usefully measured in terms of the standard deviation, and can be viewed as providing a characterization of the notion of standard deviation. Specifically, standard deviation is "average" deviation in the sense that if a random variable (uniformly) deviates from its mean by an amount $+d$ with

probability 1/2 and deviates by $-d$ with probability 1/2, then its standard deviation is d. Chebyshev's inequality constrains the possible distributions of deviations in the case when these are not uniformly distributed and not limited to two values. The characterization thus provided may also be applied to a deterministic sequence using the definitions of sample mean and scaled sample standard deviation[6] yielding precise (though in some sense loose) bounds for the given deterministic sequence[7]. We will see how this constrains sequences based on correlation below.

First, we work out the algebra in the setting of random variables and then follow this by illustrative examples focused specifically on data sequences (or data vectors). Chebyshev's inequality may be stated in the convenient form: given any random variable Z with $\mu = E\{Z\}$, and $\sigma = (E\{(Z-\mu)^2\})^{1/2}$, $P(|Z-\mu| \geq k\sigma) \leq 1/k^2$. Now, let X and Y be random variables each with mean zero and variance one (so that $\sigma_X = \sigma_Y = 1$ - the situation is slightly more complicated if X and Y have different variances). We wish to look at how different X and Y can be so we consider the expression $X - Y$ and note that this random variable has mean zero. Consequently, $P(|X-Y| \geq k\sigma_{X-Y}) \leq 1/k^2$ by Chebyshev's inequality. Correlation comes into play in calculating σ_{X-Y}. Specifically,

$$\begin{aligned} \sigma_{X-Y}^2 &= E\{(X-Y)^2\} \\ &= E\{X^2\} - 2E\{XY\} + E\{Y^2\} \\ &= 2 - 2E\{XY\} = 2 - 2\rho_{X,Y}. \end{aligned}$$

We see that the higher the correlation, the smaller σ_{X-Y} becomes and the tighter the region of variation (the less X is allowed differ from Y) on the given set. So, for a given fixed k, X must be closer to Y as correlation increases, or alternatively, as correlation increases, the measure of the set on which X is within some fixed quantity of Y becomes larger.

[6] $\sqrt{\sum_{i=1}^{n} \frac{1}{n}(x_i - \bar{x})}$, (the root mean square deviation of the sequence values from the sample mean)

[7] Note that properties such as bias are not important in the deterministic case.

Simply applying the Chebyshev inequality, we see that X and Y cannot be further from each other than 2 standard deviations more than 1/4 of the time and cannot be further than 3 standard deviations more than 1/9 of the time (etc.). So, specifically, if $\rho_{X,Y} = 1/2$, then $\sigma_{X-Y} = 1$ and so X and Y cannot be further than 2 apart more than 1/4 of the time and cannot be further than 3 apart more than 1/9 of the time. Considering movement on the order of 3 standard deviations for each, if X and Y were uncorrelated in this example, they could be 3 apart about 1/4 of the time (25% vs. approximately 10%). If $\rho_{X,Y} = 7/8$ (a correlation of .875), then $\sigma_{X-Y} = 1/2$ and so X and Y in such a case cannot be further than 1 apart more than 1/4 of the time and cannot be further than 1.5 apart more than 1/9 of the time. This is fairly close correspondence as X and Y were assumed to have standard deviations of one each and so they may vary fairly arbitrarily within a range of 2 around 0 (in particular being at say $+1$ half the time and -1 the other half the time).

We now transition to applying these ideas to simulated "time series" (realizations of random variables). The following examples (with figures on the next page) were generated in matlab and are labelled by their computed correlations. The rectangles give a clearer impression of how far the series tend to be from each other progressing across the simulated data.

Denote the time series in 2.1 rendered with asterisks (*) for its points by $\{x_i\}$ and that rendered with circles (o) for its points by $\{y_i\}$. Comparing these simulated time series to the results of applying Chebyshev's inequality, we notice first of all that there aren't any indices at which $|x_i - y_i| \geq 2$. Chebyshev's inequality essentially covers the extreme where all of the difference between x_i and y_i is concentrated at subset of the points with no differences between the series on the complementary subset. The time series $x_i - y_i$, for x_i, y_i from 2.1, has its variation distributed much more evenly (as can be seen from the sizes of the rectangles) so that most of it falls below the one standard deviation mark. Using Chebyshev and correlation at about .5, we can say, however, that $P(|x_i - y_i| \geq 3/2) \leq 4/9$. We see that in this case $|x_i - y_i| \geq 3/2$ only for $i = 6, 10, 15$ and the measure of this set is 3/20. That is,

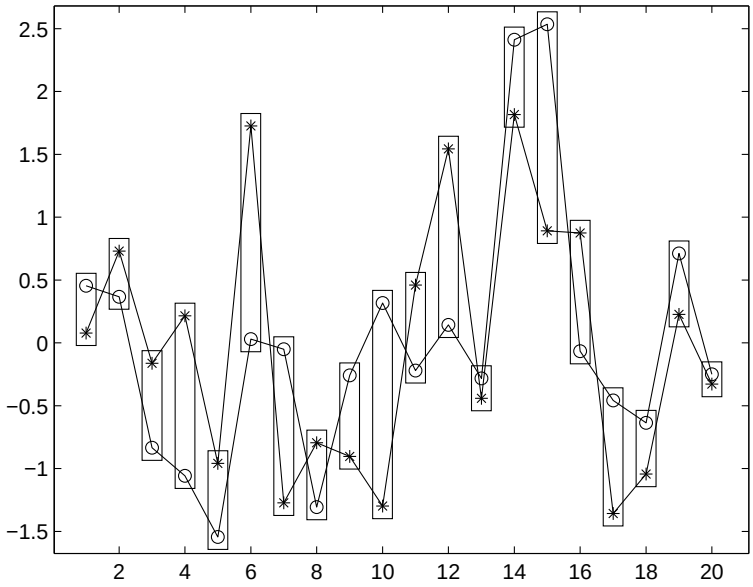


Figure 2.1: Correlation .5511 ($\mu = 0, \sigma_X = 1, \sigma_Y = 1$)

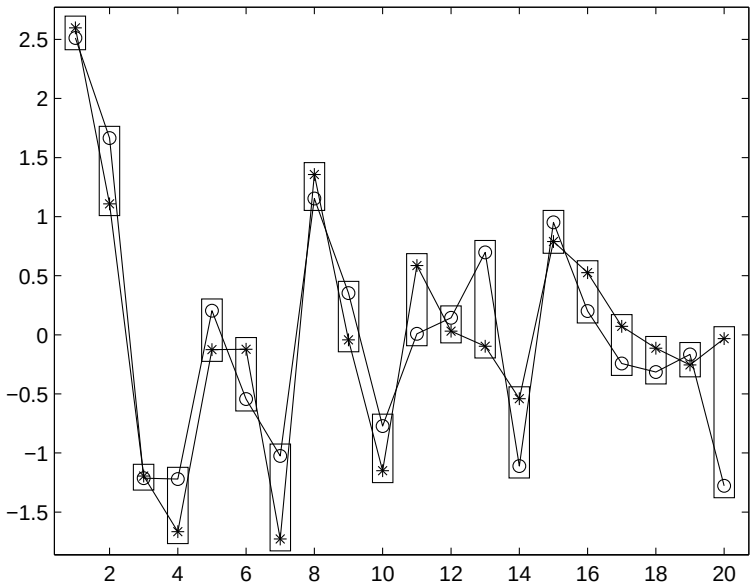


Figure 2.2: Correlation .8806 ($\mu = 0, \sigma_X = 1, \sigma_Y = 1$)

$|x_i - y_i| \geq 3/2$ on only 15% (by measure) of the measure space, far less than the allowable $\sim$44%.

As for the time series' in 2.2, it is immediately apparent, visually, that they follow one another significantly more closely than those in 2.1. Using $\{x_i\}$ now to denote the time series in 2.2 rendered with asterisks and $\{y_i\}$ to denote the time series in 2.2 rendered with circles, we notice first of all that the time series $x_i - y_i$ deviates from its mean of 0 by more than its standard deviation of $\sim 1/2$ only at the indices $i = 2, 7, 11, 13, 14, 20$. Only at $i = 20$ is the deviation greater than two standard deviations (magnitude one with $\sigma \approx 1/2$), in comparison to the bound given by Chebyshev's inequality, which allows this to occur at up to five indices. Also, this example is even more illustrative of the fact that the greater the measure of the set on which deviation is larger than the standard deviation, the smaller deviation tends to be on the set on which deviation is below the standard deviation (notice the smallness of $|x_i - y_i|$ at $i = 1, 3, 12, 13, 19$).

While these types of analyses give a good idea of what to expect for a given sample correlation level, the exact significance of a correlation obtained still depends heavily on the setting in which the correlation was computed. The techniques we are using still depend to a degree on human examination of the features of the data the algorithms deem most important.

2.5 Background for Spectral Analysis

We begin this section with an example relevant to the use of spectral analysis and that further clarifies the difficulties we face when using measurements. Consider a scenario in which we record both an EKG (electrocardiogram) and blood pressure continuously for one person over time, starting at some random time. As there are many other factors that may effect these variables and their relationship to one another, when considering these attributes alone, only probabilistic statements can be made. Again, we may think of the measurements as coming from a certain relative frequency (or probability) distribution (e.g., how frequently

will blood pressure be high given that the heart just beat etc.). If the underlying state of the system we are measuring is changing in some coherent fashion, for example, if the person is a runner and has just started running in a race, each set of measurements we take may better be understood as coming from different probability distributions (in particular, the distributions of these measurements at different levels of exertion). The properties of any one of these exertion-level based distributions would be difficult to determine from the data set collected in such a case as in sampling theory it is important to sample repeatedly from a single given distribution in order to have any reasonable hope of accurately characterizing that distribution. If instead, we consider the same type of measurements taken over a period of time during which the runner has simply been lounging about in a slug-like fashion, then the exertion-level is constant and we are in a good position to determine the parameters of the distribution for this state. The important properties this example possesses, assuming that we are thinking about repeatedly recording some number n of measurements of blood pressure and heart state with a sampling starting time independent of the state of the person's circulatory system[8], are that the distribution of any single sample point is the same as that of any other and that the relationship between a given sample value and one taken at a pre-specified fixed time later is only a probabilistic function of the difference in times. This is basically the notion of *stationarity*. This example is very similar in form to that of a specific type of stationary stochastic process known as a *harmonic process* (more detail to come).

2.5.1 Stochastic Processes Notation

A *stochastic process* is defined as a set of random vectors $\{\boldsymbol{X}_t | t \in T\}$ indexed by a set T (whose elements are often times). The notation $\{\boldsymbol{X}_t\}$ will often be used to indicate that a collection of random vectors is grouped as a stochastic process as opposed to being simply some random vectors distinguished by subscripts. Unless otherwise noted, the index set T

[8]We don't always start measuring exactly after the heart has beat, etc.

is assumed to either be $\mathbb{Z}$ or $0, 1, \ldots, n$ for convenience. A stochastic process $\{\boldsymbol{X}_t | t \in T\}$ is called *stationary* if, for all $k \geq 1$, for any set of indices $t_1 \in T, \ldots, t_k \in T$ and for any τ such that $t_1 + \tau \in T, \ldots, t_k + \tau \in T$, the distribution of the random vector $(\boldsymbol{X}_{t_1}, \ldots, \boldsymbol{X}_{t_k})'$ is the same as the distribution of the random vector $(\boldsymbol{X}_{t_1+\tau}, \ldots, \boldsymbol{X}_{t_k+\tau})'$. As we will only deal with moments of first and second order of the distributions under consideration, we will use the term *stationary* to refer to distributions where the assumption of stationarity applies at least to all moments of first and second order (processes where this slightly weaker assumption is used are often termed *weakly stationary* or *second-order stationary*). Note that the assumption of (second-order) stationarity about a given stochastic process $\{\boldsymbol{X}_t | t \in T\}$ has the following implications (that can, in particular, be used to evaluate the suitability of the assumption of stationarity in a given situation):

1. All of the random vectors in the process have the same mean vector, that is, there is a $\boldsymbol{\mu}$ such that $E\{\boldsymbol{X}_t) = \boldsymbol{\mu}$ for all $t \in T$.

2. The covariance of any two of the random vectors in the process depends only on the difference in their indices (often thought of as the amount of time in the process between the indices), that is, $E\{(\boldsymbol{X}_{t_1} - \boldsymbol{\mu})(\boldsymbol{X}_{t_2} - \boldsymbol{\mu})\} = E\{\boldsymbol{X}_{t_1}\boldsymbol{X}_{t_2}\} - \boldsymbol{\mu} = E\{\mathbf{X}_0\mathbf{X}_{\mathbf{t_2}-\mathbf{t_1}}\} - \mu$, with the last equality following from a direct application of the assumption of second-order stationarity.

Additional detail along with some discussion of the properties of estimators we will use for stationary stochastic processes will be given in the spectral analysis section.

It is important to further point out the difficulties of statistically dealing with the results of experiments. The appropriate mathematical model for such a situation really is that of a stochastic process. In particular, for any subprocess of a fixed length, there are dependence relations between each of the random vectors in that subprocess. Often, however, the measurements from a single realization are used to estimate the distribution of a subprocess $\{\boldsymbol{X}_{t_1}\}$ where t_1 is fixed (essentially, we're down to just a random vector). This is particularly

appealing when the stochastic process is reasonably assumed to be stationary (as the means, variances and covariances between the variables in the vector don't change with index). A simple example illustrates the potentially extremely poor estimates that may arise from this method. Consider recording video of a flashing light. Now, suppose that we start the recording in between flashes and that the flashes occur with the same frequency as frames are recorded by the video camera. Then in every frame recorded, the light may be dark - the recorded sample is clearly significantly skewed and not representative of the distribution of $\{\boldsymbol{X}_{t_1}\}$ at all. In fact, with this setup, it does not matter when the camera is started, using the individual frames of any given sample to estimate the distribution of $\{\boldsymbol{X}_{t_1}\}$ will give very poor results. The problem is a lack of independence between the random vectors (in the process) so that we're not independently sampling one random vector. Therefore, using only a single random vector to represent the mathematical situation is suspect. This problem may be addressed by either introducing some form of independence (for example, by recording multiple samples with the same number of frames with random start times and integrating the data) or through more careful setup of the experimental protocols (to mitigate the effects of "aliasing" which is part of the problem in the example), each of which will be touched upon in the following chapters. Our methods generally address the issue of independence through the use of repeated independent samples, but this has its own perils in the context of stationary processes. We will return to this point many times.

2.5.2 Remaining Notation

We have nearly undoubtedly left out certain notation definitions. Readers are advised to consult either Percival and Walden [15] (who have a good notation section from pages xix to xxvi) or Anderson [2]. The notation used in the remainder is generally consonant with the notation used in these books. Although we have tried to set the notation carefully, it is the case, particularly in the spectral section, that the conventions break down a little. In particular, we use $S(f)$ for the spectrum although it is not random, etc.

Chapter 3

Canonical Correlation Analysis

3.1 Introduction

In this chapter, we consider the first technique we will need: the mathematical technique that will be used to integrate information from, compare variables across, and distinguish variables within, any given class of data sets. The technique we will use is called *canonical correlation analysis*. The idea behind this technique is to find those variables, across the entire class of data sets being compared, that are most similar in the sense of correlation.[1] This will allow for the integration of data to obtain more accurate estimates for the parameters of a given underlying process (such as when one experiment is performed repeatedly) and for the comparison of distinct data sets (from, say, different experiments) to determine important relationships between their underlying processes. The integrative aspect in particular is important as it allows for the use of more clearly independent estimates - those made across data sets - which is advantageous from a statistical standpoint. As to the structure of this chapter, after fleshing out more fully the importance and usefulness of canonical correlation analysis in the remainder of the introduction, we move on (section by section) to the mathematics of "classical" canonical correlation analysis, an example illustrating the results obtainable applying this classical technique to data and a brief discussion of its statistical workings. From this foundation, we go on to detail a generalization of "classical" canonical correlation analysis due to Carroll and a numerically stable computational method

[1] For reference, variables similar in the sense of correlation exhibit a relationship like that of height and weight among people: taller people tend to be heavier (see the statistics section for more detail).

for applying this generalization (justifying its use in the numerical setting, particularly on current computers - see the appendix for more information on stability).

To elaborate on the importance of the integrative and comparative aspects of canonical correlation analysis (CCA), we consider a couple of typical situations starting with one in which an experimenter may run an experiment, collect data, and then at a later date perform essentially the same experiment again. Mathematically, we have two realizations of a stochastic process $\{\mathbf{X}_t | t \in T\}$ where $\mathbf{X}_t$ contains only the variables that are shared between the two experiments. The discussion in the previous chapter makes it clear that there are many reasons to want to combine the data from these two realizations. Combining data can be a challenge, however. As an example, if the experiment involves the recording of video, the setup might be slightly different or the geometry of the contents of the recorded apparatus may have changed even if the underlying setup (the organisms under consideration etc.) is the same. Consequently, a priori we cannot simply concatenate the data sets and use principal component analysis (another potentially applicable technique) even with a nearly identical setup (it isn't clear how the variables match up). Further, a situation in which a few experiments are performed, even without the intent to compare the results rigorously or directly, is fairly common - probably more common than simply taking more data in a given experiment - and so it is quite important to use some technique to maximize useful data extraction from multiple examinations of a given process. Finally, another situation in which CCA can be of significant value occurs when data is recorded of ostensibly distinct processes but there is a desire to see what these processes have in common. For example, there has been some interest recently in looking at the influence of a broad range of demographic factors on aging. One surprising result was a strong positive correlation between length of formal education and length of life. Although CCA was apparently not used for this study, it could have very easily been applied to the various data sets used in a more automated fashion than the analyses actually carried out ([9]).

3.2 Classical CCA, Stationarity

We begin with the "classical" canonical correlation analysis (CCA) problem. Given a random vector with mean vector zero

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}$$

partitioned as shown, determine linear combinations $\boldsymbol{\alpha}'\boldsymbol{X}_1$ and $\boldsymbol{\gamma}'\boldsymbol{X}_2$ each of unit variance ($E\{\boldsymbol{\alpha}'\boldsymbol{X}^{(1)}\}^2 = 1$ etc.) such that their covariance is maximized (i.e., maximizing $E\{(\boldsymbol{\alpha}'\boldsymbol{X}_1)(\boldsymbol{\gamma}'\boldsymbol{X}_2)\}$), and frame the discussion in terms of it. The partitioning of $\boldsymbol{X}$ determines a corresponding partitioning of the covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Note that we may write the variance of any linear combination of the random variables in one of the distinguished sets using a quadratic form $\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha}$ and the covariance of linear combinations of variables from distinct sets using $\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{12}\boldsymbol{\gamma}$ (where $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are column vectors of coefficients for the linear combinations in question).

The solution to the CCA problem may be determined using Lagrange multipliers to maximize $\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{12}\boldsymbol{\gamma}$ subject to $\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha} = 1$ and $\boldsymbol{\gamma}'\boldsymbol{\Sigma}_{22}\boldsymbol{\gamma} = 1$. A variable thus determined, such as $\boldsymbol{\alpha}'\boldsymbol{X}_1$, is known as a *canonical variable* and its correlation with the corresponding variable $\boldsymbol{\gamma}'\boldsymbol{X}_2$ is known as a *canonical correlation*. The canonical variables are usually ordered by canonical correlation so that the first set has the highest canonical correlation, the second set has the second highest etc. The coefficients in $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are called *canonical coefficients*. Let $\boldsymbol{\alpha}_i$, $\boldsymbol{\gamma}_i$ correspond to the ith set of canonical variables and $\boldsymbol{\alpha}_j$, $\boldsymbol{\gamma}_j$ correspond to the jth set. The computations reveal the following important properties of the solutions:

1. The canonical variables derived from a given subvector are all mutually uncorrelated (orthogonal) so canonical correlation analysis produces an orthogonal decomposition of the spaces spanned by the random variables in $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. That is, $\boldsymbol{\alpha}_i'\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha}_j = 0$ when $i \neq j$.

2. Non-corresponding canonical variables derived from $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are uncorrelated (orthogonal). That is, $\boldsymbol{\alpha}_i' \boldsymbol{\Sigma}_{12} \boldsymbol{\gamma}_j = 0$ when $i \neq j$.

These conditions are of significant importance in the usefulness of CCA and we will see that a version of the second property holds for the generalization which will be plenty for our application.

In the practical setting, a straightforward application of the CCA technique just presented would involve the estimation of the covariance matrix $\boldsymbol{\Sigma}$ given two data sets using, for example, the sample covariance matrix estimator. This will not work directly, though, for the types of data we wish to treat. To be precise, we really need to work with stochastic processes and not just random vectors and within this context we will see (later on) that the problem is caused both by the independent sampling of the stochastic processes in question and by the stationarity of those processes. We will return to these issues when we are in a position to transform data from stationary stochastic processes into a form amenable to treatment by CCA (that is, after we cover spectral analysis). For the remainder of this chapter, however, we will work within the general case where there may be arbitrary dependence relations between the subvectors of $\boldsymbol{X}$ and ignore the difficulties more specific to our setting. We begin with an illustrative example.

3.3 Example

Let $\mathbf{X} = [x_{ij}]$ and $\mathbf{Y} = [y_{ij}]$ for $i = 1, \ldots, 128$ and $j = 1, \ldots, 10$ be data matrices with data courses running down their columns (and distinct variables running across their rows). Let the values of the ten variables in each set across the 128 time indices be defined by

$$\begin{aligned}
x_{ij} &= 3\sin(2\pi f_s i) + \cos(2\pi f_c i) + \epsilon \qquad \text{for } j = 1, \ldots, 5 \\
x_{ij} &= \sin(2\pi f_s i) + 4\cos(2\pi f_c i) + \epsilon \qquad \text{for } j = 6, \ldots, 10 \\
y_{ij} &= 3\sin(2\pi f_s i) + \cos(2\pi (f_c + .08) i) + \epsilon \qquad \text{for } j = 1, \ldots, 5 \\
y_{ij} &= \sin(2\pi f_s i) + 3\cos(2\pi (f_c + .05) i) + \epsilon \qquad \text{for } j = 6, \ldots, 10
\end{aligned}$$

where $f_s = 4/128$ and $f_c = 2/128$ are the basic frequencies of the sine waves and cosine waves and ϵ is a Gaussian random variable with mean 0 and standard deviation .3. So the columns of X and Y contain mixtures of sine and cosine "signal" and Gaussian noise. Canonical correlation analysis (generalized and classical are equivalent for two data sets) finds the following:

$$\mathbf{R} = (0.9959 \quad 0.9829 \quad 0.4413 \quad 0.3355 \quad 0.2722 \quad 0.2189 \quad 0.1370 \quad 0.1147 \quad 0.0744 \quad 0.0193)$$

and

$$\mathbf{A} = \begin{pmatrix} 0.1377 & 0.1059 & -2.5346 \\ 0.1197 & -0.0267 & 0.8864 \\ 0.0672 & -0.0943 & 0.4552 \\ 0.0897 & 0.0937 & 1.2267 \\ 0.0994 & 0.0293 & -0.0450 \\ -0.0271 & -0.1378 & 0.0771 \\ -0.0399 & -0.0473 & -0.8972 \\ -0.0348 & 0.0066 & 0.2659 \\ -0.0077 & -0.0506 & -0.9882 \\ -0.0268 & -0.1430 & 1.4938 \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} 0.1306 & 0.0982 & -1.2159 \\ 0.1433 & 0.0397 & -0.8129 \\ 0.1003 & 0.0012 & 1.4480 \\ 0.0798 & -0.0341 & 0.7870 \\ 0.0713 & 0.0447 & -0.1697 \\ -0.0096 & -0.1102 & -1.3591 \\ -0.0619 & -0.3044 & 0.5037 \\ -0.0206 & -0.1781 & 0.9753 \\ -0.0592 & 0.1053 & -1.5334 \\ -0.0404 & -0.0407 & 1.4619 \end{pmatrix}$$

where $\mathbf{R}$ is the vector of canonical correlations and $\mathbf{A}$ and $\mathbf{B}$ are matrices containing the first three sets of canonical coefficients (one set per column). First, we see that CCA has largely identified the variables that are most similar based on their definitions. In particular, the first set of canonical coefficients in both $\mathbf{A}$ and $\mathbf{B}$ clearly distinguish the first 5 variables among each set of variables (for example, in $\mathbf{A}$, the smallest magnitude coefficient for the first five variables is .0672 while the largest magnitude coefficient for the last five variables only has magnitude .0399). These are the variables dominated by the sine signal. The second set of canonical coefficients in both $\mathbf{A}$ and $\mathbf{B}$ distinguish the last 5 variables among each set of variables, although the distinction is less clear (especially in $\mathbf{A}$). This is, in actuality,

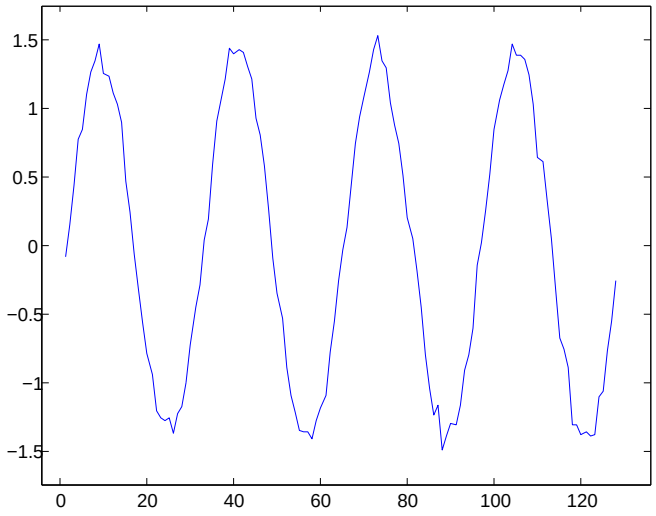


Figure 3.1: Canonical Variable 1 (first data set)

a good illustration of the variable identification aspect of CCA in the presence of noise. In terms of data matrices, these coefficients determine a change of basis for the column spaces of the original matrices. The third set of canonical coefficients and beyond basically determine noise dimensions in the data in this case (note the much lower canonical correlations for three and beyond).

Now, let's look at the canonical variables themselves. The figures are spread across a couple of pages.

We see that CCA has separated the uncorrelated sine and cosine signals out of mixed signals, like that plotted in 3.3, that are also noisy, illustrating the effectiveness of CCA in separating orthogonal coherent signals. Now, let's examine the effect of the averaging. Each input signal was contaminated with mean zero, standard deviation 0.3 gaussian noise. Comparing the plots in 3.4 and 3.5, the effect of averaging in reduced variance (increase smoothness) is apparent.

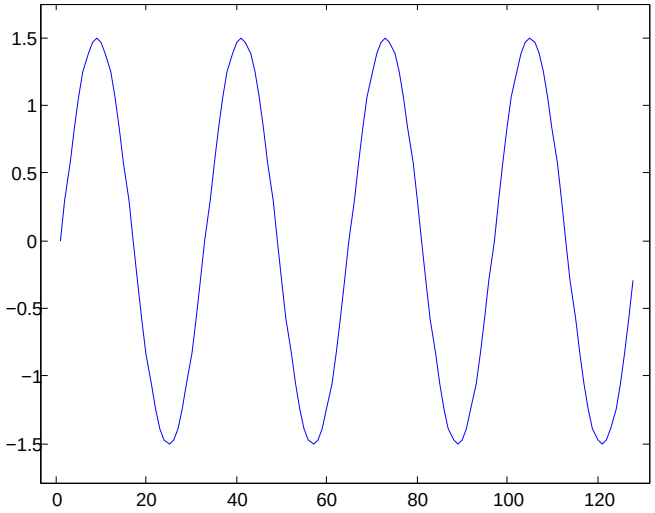


Figure 3.2: Pure Sine Signal

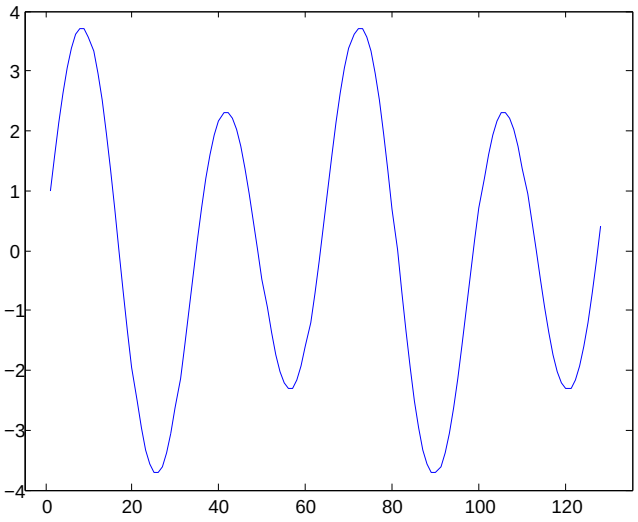


Figure 3.3: 3sin+cos Without Noise

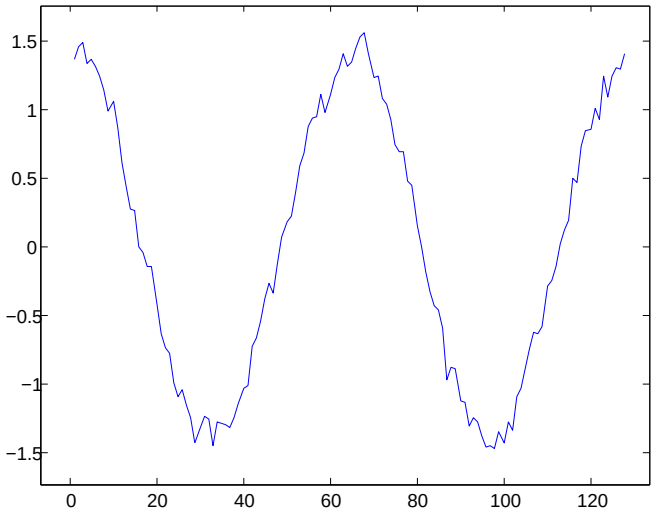


Figure 3.4: Canonical Variable 2 (first data set)

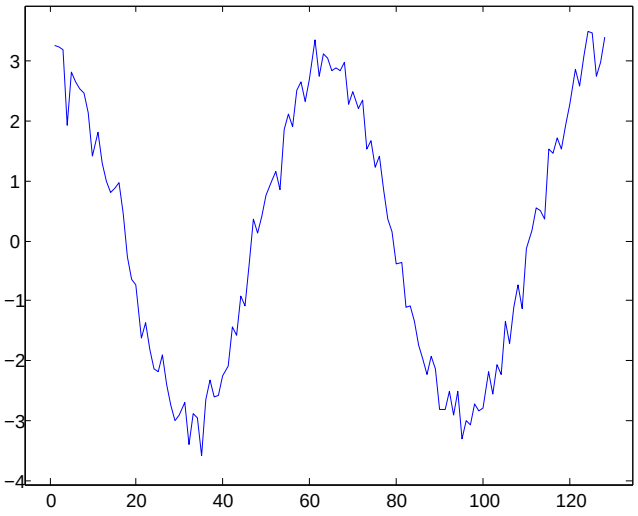


Figure 3.5: cos + noise

This example demonstrates the important properties of CCA - that coherent yet distinct features are separated by orthogonality or uncorrelatedness while noncoherent features are averaged out (noise/variability).

3.3.1 Geometry

It is possible to give a geometric interpretation of CCA that can be quite helpful in understanding its effect. The example just discussed may be represented schematically as in figures 3.6 and 3.7 (the numbers on the axes should be ignored). In these diagrams, the $\mathbf{v}1$ direction corresponds to the sine signal, the $\mathbf{v}2$ direction corresponds to the cosine signal and the complement to the subspace spanned by $\mathbf{v}1$ and $\mathbf{v}2$ is represented by the arrow labelled with $\mathbf{v}3, \mathbf{v}4, \ldots$. These grouped complementary dimensions[2] contain noise and other less important features of the simulated data. In 3.6, the green arrows represent the first five column vectors of the $\mathbf{X}$ matrix (lying primarily along the sine direction and gaussian distributed around the $\mathbf{v}1$,$\mathbf{v}2$ subspace in the noise dimensions), and the red arrows represent the last five column vectors of the $\mathbf{X}$ matrix. The black dotted lines illustrate that the vectors in each set contain the same mixtures of the sine and cosine signals but have differences in the noise dimensions. The same basic comments apply to the figure 3.7. Note that the altered frequency used with the cosine signal in this data set accounts for the fact that the red vectors are not centered on the $\mathbf{v}1, \mathbf{v}2$ subspace.

Geometrically, the classical CCA method consists of a projection onto $x_1 + x_2 + \ldots + x_n = 0$ (removal of the mean) followed by the determination of (specially) ordered orthogonal bases for the subspaces spanned by the two projected collections of variables. The ordered orthogonal bases generated have the important property that the first basis vectors for each projected collection minimize the (projective) distance between the one-dimensional subspaces each spans (or equivalently maximize the cosine of the angle between them), the second set minimize the distance out of the remaining dimensions, etc. The first canonical

[2] In the example there are 126 complementary dimensions.

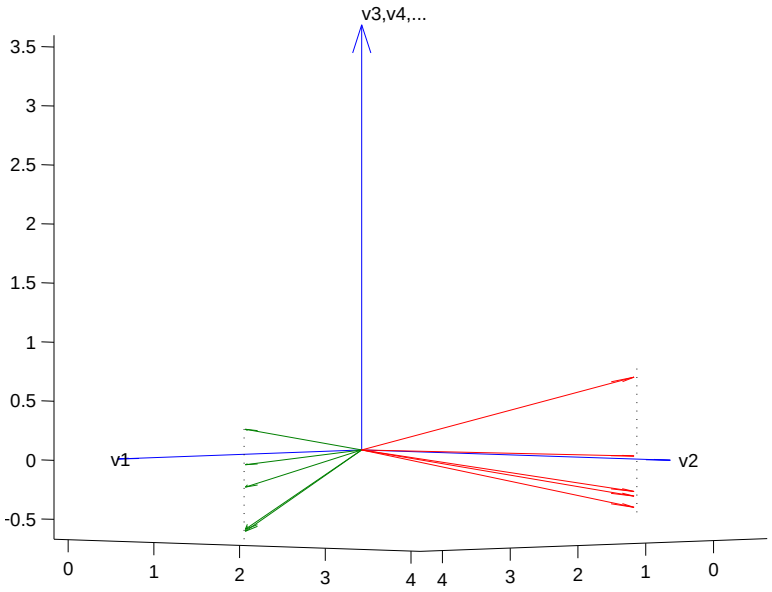


Figure 3.6: First Data Set

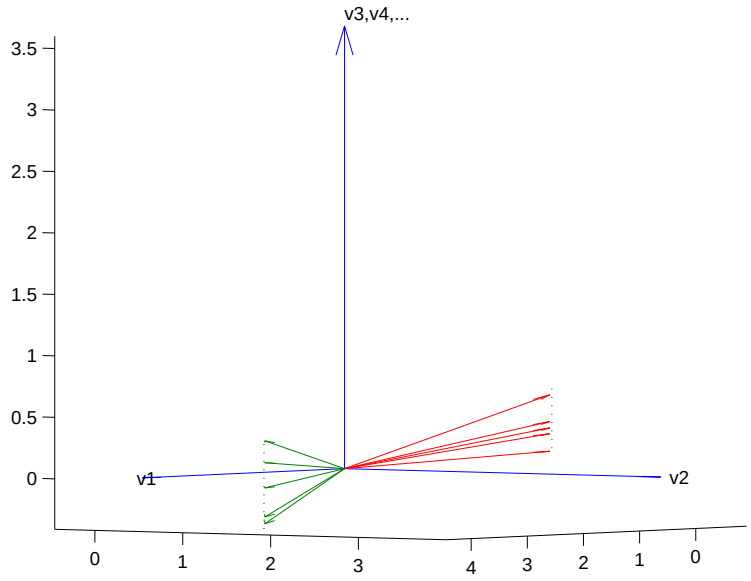


Figure 3.7: Second Data Set

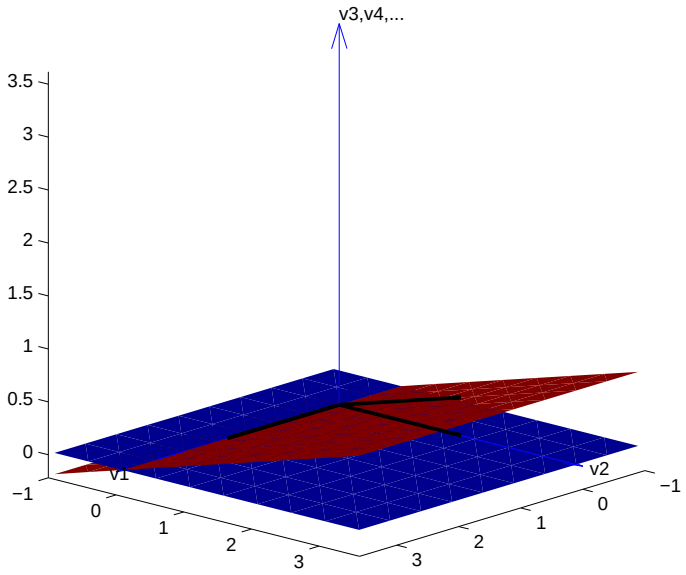


Figure 3.8: After CCA

correlation is, in this interpretation, the cosine of the angle between the subspaces spanned by the first canonical variables, etc. The result of this in the example is illustrated in figure 3.8. The blue plane corresponds to the subspace spanned by the first two canonical vectors (variables) of the first data set with the canonical variables themselves shown as thick black arrows. The canonical variables are nearly exactly $\mathbf{v}1$ and $\mathbf{v}2$ for this data set (the deviation is practically negligible and isn't visible in the rendering). The red plane corresponds to the subspace spanned by the first two canonical variables of the second data set. Note that the first canonical variable is again almost exactly $\mathbf{v}1$ but that the second canonical variable is a more clearly perturbed version of $\mathbf{v}2$.

3.4 Statistics

Briefly, the use of the methods here may be justified by the good statistical properties of the sample covariance matrix. In particular, the sample covariance matrix is usually a good estimate of the actual covariance matrix (as detailed previously) and so the robustness

depends on how much the correlations and coefficients can change due to small perturbations in the matrix. Although we are in the set of polynomials with real coefficients and real roots, and consequently those roots are smooth functions of the coefficients, there may still be some sensitivity of the canonical correlations themselves to perturbations. Specifically, the equations for the classical case above may be written in the form of a generalized eigenvalue problem[3]. In this form, the symmetry of our problem precludes the eigenvalues of the two matrices $\tilde{\mathbf{S}} = \mathbf{A} - \lambda\mathbf{B}$ is split into from being particularly sensitive and we can say that the correlations in general will not be particularly sensitive to perturbations as long as the magnitudes of corresponding ordered eigenvalues of the $\mathbf{A}$ and $\mathbf{B}$ matrices are similar. The only trouble we may have is in the case that $\mathbf{B}$ has a very small eigenvalue and $\mathbf{A}$ has no correspondingly small eigenvalue. References for the sensitivity of the eigenvectors may be found in Golub and Van Loan [10]. In the classical theory, it is also possible to justify the technique using the method of maximum likelihood after assuming that the random vectors are normally distributed. The unbiased estimator of the covariance matrix is simply a scale of the maximum likelihood estimator of the covariance matrix assuming a normal distribution $(n/(n-1))$ and we in fact end up with the same numerical correlations and simply slight scales of the canonical coefficient vectors adding weight to the claim of robustness.

3.5 Generalized Canonical Correlation Analysis

It is desirable to use an extension of the "classical" canonical correlation analysis that may be applied to $n \geq 2$ data sets as opposed to simply 2 data sets (again to incorporate more independence into our estimation). Since we will not be presenting the statistical properties of the method in its general form and due to the elegance and clarity of the approach outlined by Carroll, only the original data matrix presentation of this method will be detailed. Interested readers may refer to Kettenring [11] for a random variable casting of the approach (which for the present discussion would seem only to complicate matters). The statistical theory,

[3]See Anderson [2] for the determinant expression for the equations, p. 490.

which is not the focal point here, will be presented for the two variable classical CCA as this is relatively straightforward and we will simply appeal to the equivalence between the classical and generalized approach in this case.

Carroll's approach (generalized CCA) proceeds from a nice reframing of the canonical correlation problem. Instead of casting the problem in terms of finding linear combinations of random variables maximally correlated with one another, Carroll casts the problem in terms of finding a new variate that allows for a certain sum of squared correlations of linear combinations of the original variables with this new variate to be maximized. Specifically, let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be n matrices with $\mathbf{X}_i$ of dimensions $k \times m_i$ representing k observations of m_i variables and the column means all 0. The goal is then to find a variate $\mathbf{Z}$ (a column vector of length k whose components sum to zero) and linear combinations $\mathbf{A}_i$ ($\mathbf{A}_i$ a column vector with m_i components) with $\mathbf{A}_i'\mathbf{X}_i'\mathbf{X}_i\mathbf{A}_i) = 1$ (variance one) maximizing the expression

$$R^2 = \sum_{i=1}^{n} r(\mathbf{Z}, \mathbf{X}_i\mathbf{A}_i)^2$$

where $r(\mathbf{Z}, \mathbf{X}_i\mathbf{A}_i) = (\mathbf{Z}'\mathbf{X}_i\mathbf{A}_i)/(\mathbf{Z}'\mathbf{Z})$ is the product moment correlation between $\mathbf{Z}$ and $\mathbf{X}_i\mathbf{A}_i$. We now present a quick run through to the solution of this problem (which is also quite elegant), filling in a couple of details alluded to in Carroll's paper.

First, it is possible to determine the maximum possible value of the expression $r(\mathbf{Z}, \mathbf{X}\mathbf{A})^2$ for a given fixed $\mathbf{Z}$ over all possible $\mathbf{A}$'s with the variance of the linear combination defined by $\mathbf{A}$ held equal to 1. It will be shown below that the maximum value is

$$\max_{\mathbf{A}} r(\mathbf{Z}, \mathbf{X}\mathbf{A})^2 = \frac{\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}}{\mathbf{Z}'\mathbf{Z}}$$

subject to $\mathbf{X}'\mathbf{X}$ being invertible. So the maximum possible value of R^2 is (only) achievable when $\mathbf{Z}$ is determined maximizing

$$\frac{1}{\mathbf{Z}'\mathbf{Z}} \sum_{i=1}^{n} \mathbf{Z}'\mathbf{X}_i(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{Z} = \frac{\mathbf{Z}'\mathbf{Q}\mathbf{Z}}{\mathbf{Z}'\mathbf{Z}}$$

where $\mathbf{Q} = \sum_{i=1}^{n} \mathbf{X}_i(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'$. The solution to this derived maximization problem is, of course, well known as $\mathbf{Q}$ is positive definite or semi-definite[4] allowing for an application of the spectral theorem. The eigenvector of $\mathbf{Q}$ corresponding to its largest eigenvalue solves the problem. In fact, the spectral theorem gives successive uncorrelated "best" $\mathbf{Z}$ variates essentially decomposing as much of the data spaces as the rank of $\mathbf{Q}$ allows. Once a given $\mathbf{Z}$ has been determined, the $\mathbf{A}_i$ may be determined by $\mathbf{A}_i = (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{Z}$ which will also be shown below.

A couple of brief computations using Lagrange multipliers establish the claims made above (alternatively Hilbert space methods may be used). Under the assumption $\mathbf{Z}'\mathbf{Z} = 1$ and on the constraint curve $\mathbf{A}_i'\mathbf{X}_i'\mathbf{X}_i\mathbf{A}_i = 1$ (which is a compact manifold guaranteeing maxima and minima and their occurence at critical points - with appropriate assumptions on $\mathbf{X}_i$), $r(\mathbf{Z}, \mathbf{X}_i\mathbf{A}_i) = \mathbf{Z}'\mathbf{X}_i\mathbf{A}_i$, the expression that will actually be optimized (as opposed to its square). The following computations were down with the transposes of the terms discussed above. Letting

$$\mathbf{\Psi} = \mathbf{Z}'\mathbf{X}\mathbf{A} - \frac{1}{2}\lambda(\mathbf{A}'\mathbf{X}'\mathbf{X}\mathbf{A} - 1)$$

(where $\mathbf{X}$ is some fixed $\mathbf{X}_i$) so that the critical point equation(s) is (lambda portion)

$$\frac{\partial \mathbf{\Psi}}{\partial \mathbf{A}} = \mathbf{Z}'\mathbf{X} - \lambda\mathbf{A}'\mathbf{X}'\mathbf{X} = 0$$

multiplying through by $\mathbf{A}$ on the right and applying the constraint shows that $\lambda = \mathbf{Z}'\mathbf{X}\mathbf{A}$ when the critical point equation is satisified. Consequently

$$\begin{aligned} 0 = \mathbf{Z}'\mathbf{X} - \lambda\mathbf{A}'\mathbf{X}'\mathbf{X} = \mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} - \lambda\mathbf{A}' &= \mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z} - \lambda\mathbf{A}'\mathbf{X}'\mathbf{Z} \\ &= \mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z} - (\mathbf{A}'\mathbf{X}'\mathbf{Z})^2 \end{aligned}$$

assuming that $(\mathbf{X}'\mathbf{X})$ is invertible and so we see that $(\mathbf{A}'\mathbf{X}'\mathbf{Z})^2 = \mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}$ at both the maxima and minima of $r(\mathbf{Z}, \mathbf{X}_i\mathbf{A}_i)$ so that the maximum value of $r(\mathbf{Z}, \mathbf{X}_i\mathbf{A}_i)^2$ is necessarily $\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}/\mathbf{Z}'\mathbf{Z}$ upon dropping the assumption $\mathbf{Z}'\mathbf{Z} = 1$. This last computation

[4]Carroll's paper [3] also allows for weights in the expression to be maximized but the analysis is the same. See Carroll for more.

also shows that $\mathbf{A}_i = \mathbf{Z}'\mathbf{X}_i(\mathbf{X}_i'\mathbf{X}_i)^{-1}$ is the appropriate choice for $\mathbf{A}_i$ (or really a scale). The equivalence of this technique and the classical technique may be verified by algebraic manipulations involving the eigenvector equation with $\mathbf{Q}$ and the transformation vector equations on the generalized side and the critical point equations (found in Anderson [2] for example) for the classical CCA. This is a key point but the actual verification itself is not as important and so is omitted. The only property of the classical approach that does not carry over exactly (for $n \geq 3$) is that of the mutual orthogonality or uncorrelatedness of the transformed vectors from each data set. In exchange, however, we obtain uncorrelatedness of the canonical variates themselves.

The above gives an "analytic approach" to the problem of generalizing the CCA, as Carroll points out (thinking of eigenvalue and eigenvector determination as analytic). As always, the specter of numerical instability with limited precision floating point numbers is looming in these computations (see the appendix for more information about numerical stability). It turns out, though, that this approach works out particularly well (numerically) when combined with PCA (principal component analysis) leading to numerically stable and efficient computational procedures.

3.6 Numerically Stable Computation

Carroll's CCA depends on the computation of the matrix $\mathbf{Q}$ as presented in his paper, a computation that isn't particularly numerically stable. In particular, small changes in the entries of the data matrices can cause large changes in the resulting canonical correlations and coefficients. Further explanation of this problem may be found in the appendix. We will now demonstrate that $\mathbf{Q}$ (with the weighting terms, w_i, given by Carroll) may be computed using the left singular vectors that arise from a singular value decomposition[5] (SVD) performed on the original data matrices. This importantly gives a stable computational method for the correlation scores and a generally stable method for the canonical variates. It also makes

[5]See the appendix.

certain features of this analysis clearer. We have not had time to find a reference for this material (which may be new or may be known to more seasoned researchers in the area).

In this section, we assume that each of the $\mathbf{X}_i$'s is full rank with mean vector zero and that $k \geq m_i$ for all i. By the "economical version" of the SVD theorem[6], then, $\mathbf{X}_i$ may be written as $\mathbf{X}_i = \mathbf{U}_i\mathbf{S}_i\mathbf{V}_i'$ where $\mathbf{U}_i$ is a partial isometry with dimensions $k \times m_i$ (that is, the columns of $\mathbf{U}_i$ form an orthonormal basis of a subspace of $\mathbb{R}^k$), $\mathbf{S}_i$ is a diagonal matrix with dimensions $m_i \times m_i$, and $\mathbf{V}_i$ is an orthogonal matrix with dimensions $m_i \times m_i$. Then,

$$\begin{aligned}
\mathbf{Q} &= \sum_{i=1}^{n} w_i \mathbf{X}_i (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i' \\
&= \sum_{i=1}^{n} w_i \mathbf{U}_i\mathbf{S}_i\mathbf{V}_i'(\mathbf{V}_i\mathbf{S}_i\mathbf{U}_i'\mathbf{U}_i\mathbf{S}_i\mathbf{V}_i')^{-1}\mathbf{V}_i\mathbf{S}_i\mathbf{U}_i' \\
&= \sum_{i=1}^{n} w_i \mathbf{U}_i\mathbf{S}_i\mathbf{V}_i'(\mathbf{V}_i')^{-1}\mathbf{S}_i^{-2}\mathbf{V}_i^{-1}\mathbf{V}_i\mathbf{S}_i\mathbf{U}_i' \\
&= \sum_{i=1}^{n} w_i \mathbf{U}_i\mathbf{S}_i\mathbf{S}_i^{-2}\mathbf{S}_i\mathbf{U}_i' \\
&= \sum_{i=1}^{n} w_i \mathbf{U}_i\mathbf{U}_i',
\end{aligned}$$

which avoids computations of inverses as well as involving only multiplications of well-scaled matrices. This expression has yet another advantage in that it may be written in a particularly convenient form using block matrices such that explicit matrix products may be avoided altogether. Forming the block matrix $\mathbf{U} = [\sqrt{w_1}\mathbf{U}_1\ \sqrt{w_2}\mathbf{U}_2 \ldots \sqrt{w_n}\mathbf{U}_n]$, the sum may be written $\mathbf{U}\mathbf{U}' = [\sqrt{w_1}\mathbf{U}_1\ \sqrt{w_2}\mathbf{U}_2 \ldots \sqrt{w_n}\mathbf{U}_n][\sqrt{w_1}\mathbf{U}_1\ \sqrt{w_2}\mathbf{U}_2 \ldots \sqrt{w_n}\mathbf{U}_n]' = \sum_{i=1}^{n} w_i\mathbf{U}_i\mathbf{U}_i' = Q$, and as we have seen, the eigenvectors of $\mathbf{U}\mathbf{U}'$ may be computed using a singular value decomposition. Specifically using the standard SVD, $\mathbf{U}$ is a $k \times \sum_{i=1}^{n} m_i$ matrix decomposable as $\mathbf{U} = \tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}'$ with $\tilde{\mathbf{U}}$ orthogonal of size $k \times k$, $\tilde{\mathbf{S}}$ of size $k \times \sum_{i=1}^{n} m_i$ with nonzero entries only on its diagonal, and $\tilde{\mathbf{V}}$ a $\sum_{i=1}^{n} m_i \times \sum_{i=1}^{n} m_i$ orthogonal matrix. Therefore, $\mathbf{U}\mathbf{U}' = \tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}'\tilde{\mathbf{V}}\tilde{\mathbf{S}}\tilde{\mathbf{U}}' = \tilde{\mathbf{U}}\tilde{\mathbf{S}}^2\tilde{\mathbf{U}}'$, showing that the columns of $\tilde{\mathbf{U}}$ are eigenvectors of $\mathbf{U}\mathbf{U}'$ and so of the matrix $\mathbf{Q}$.

[6]See Kincaid and Cheney [12], p. 295.

As for the transformation vectors, the computation as given by Carroll is $\mathbf{A}_i = (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{Z}$. Using the matrix factorizations above, the nature of this computation may be clarified and its computational qualities improved. Rewriting, we have

$$\begin{aligned}
\mathbf{A}_i &= (\mathbf{V}_i\mathbf{S}_i\mathbf{U}_i'\mathbf{U}_i\mathbf{S}_i\mathbf{V}_i')^{-1}\mathbf{V}_i\mathbf{S}_i\mathbf{U}_i'\mathbf{Z} \\
&= (\mathbf{V}_i\mathbf{S}_i^2\mathbf{V}_i')^{-1}\mathbf{V}_i\mathbf{S}_i\mathbf{U}_i'\mathbf{Z} \\
&= (\mathbf{V}_i')^{-1}\mathbf{S}_i^{-2}\mathbf{V}_i^{-1}\mathbf{V}_i\mathbf{S}_i\mathbf{U}_i'\mathbf{Z} \\
&= \mathbf{V}_i\mathbf{S}_i^{-1}\mathbf{U}_i'\mathbf{Z}
\end{aligned}$$

which is really an application of the Moore-Penrose pseudo-inverse of $\mathbf{X}_i$ to $\mathbf{Z}$. This shows that that $\mathbf{A}_i$ is a least-squares solution to $\mathbf{X}_i\mathbf{A}_i = \mathbf{Z}$ and gives a numerically improved computational method.

Chapter 4

Spectral Analysis

4.1 Introduction

In this chapter, we will transition from discussion of the multivariate ("uni-time" - that is, corresponding to drawing one number for each variable each time a sample is taken) setting, where we were concerned with finding relationships between different variables (e.g., representing different attributes of some population) under independent sampling from a fixed multivariate distribution, to the univariate ("multi-time") setting, where we will be concerned with sampling a single attribute some number of times in succession (over time) and the probabilistic relationships between the values that might be obtained at different time points. That is, instead of a multivariate random vector, we focus on a stochastic process consisting of (univariate) random variables. Although the distributions of interest are multivariate in both cases and may be handled identically from a mathematical standpoint, the additional structure of the time series our stochastic processes model (the natural, meaningful ordering by time and the assumed stationarity) may be exploited to some good effect. In particular, it is possible to develop a powerful, sensitive and useful statistic that provides a good description of the types of (univariate) stochastic processes we are interested in. This statistic can then be applied in the multivariate case (variable by variable) and combined with the CCA technique discussed previously simultaneously resolving the difficulties with applying CCA discussed previously. After providing additional motivation for the importance of sensitivity etc. in the scientific (data) setting, we begin the technical exposition by considering a basic statistical approach to time series analysis, followed by consideration of its limitations, and moving from there to discuss spectral analysis, how it improves upon

some of these limitations and the important relationship between the spectral statistics and the initial described "basic" statistics. The most significant portion of this chapter (both in length and in the value of the results) is concerned with estimation and the practical implementation of spectral analysis.

While we have appealed repeatedly to the flood of data in modern science in arguing for (improved) computational techniques, we have ignored the subtle point that in any single run of a given experiment, each variable may be sampled relatively few times and the size of the data set generated may stem more from the number of variables (for example, videos of biological processes often have this property). Further, in order for a model with assumed stationarity to be justifiable, it is often better to deal with data sets that were collected over a relatively short period of time. Therefore, we need techniques that are sensitive and robust when dealing with relatively short time series. For this reason, we will employ the multitaper spectral estimators that have these properties. There is yet another important reason to turn to spectral analysis, however, and that is the naturalness of working in the frequency domain with certain types of data. For example, in brain imaging studies (using, say, fMRI), the subjects may be exposed to a periodic stimulus such as a flashing light and so the expected response would likely be periodic. Even more importantly, though, the human body naturally generates a variety of periodic signals corresponding to periodic event such as heartbeats, breathing etc. It is often possible to relatively easily filter or at least separate these out in the frequency domain - something that is typically not easy in the time domain (see Mitra and Pesaran [13]). With this "practical side" motivation in hand, we turn first to the basic statistical development.

4.2 Autocovariance

The most basic measures from probability theory for quantifying the properties of and relationships within a collection of random variables consist of the numerical means, variances and covariances (or correlations) of the random variables and pairs of random variables. As

we are interested in how a stochastic process evolves in time, the focus will be on the covariances between the random variables for different time indices. In this situation, the index set is of course ordered and so it is natural to use this order when presenting the covariances giving rise to the notions of *autocovariance functions* and *autocovariance sequences*. It is these descriptions of the properties of a stochastic process we will initially focus on. These more straightforward statistical descriptions are crucially related to the spectral decompositions of interest.

For any continuous parameter stationary stochastic process $\{X(t)\}$, the *autocovariance function* is defined by

$$s(\tau) = \text{cov}\{X(t), X(t+\tau)\} = \text{cov}\{X(0), X(\tau)\}.$$

The properties of this descriptive tool as well as its relationship to the tools that will be developed in the course of this section are most illuminated by considering a specific type of stationary stochastic process called a *harmonic process* which underlies the area of spectral analysis and is the key to casting time series analysis in this light. Consequently, throughout the first part of this section we will consider the specific simple harmonic process $\{X(t)\}$ defined by $X(t) = a\cos(2\pi ft + \phi)$ where a is the constant amplitude, f is the constant frequency and ϕ is a random variable with a uniform distribution on $[-\pi, \pi]$. This process may be thought of as a model for the voltage across the conductors of a wall outlet (energized with $f$hz aV AC) with ϕ representing the random time at which voltage tracking commences. Calculating the autocovariance function for this process is quite instructive (note mean zero):

$$\begin{aligned} s(\tau) &= \text{cov}\{X(0), X(\tau)\} \\ &= E\{X(0)X(\tau)\} \\ &= \int_{-\pi}^{\pi} (a\cos(\phi))(a\cos(2\pi f\tau + \phi))\frac{d\phi}{2\pi} \\ &= \frac{a^2}{2}\cos(2\pi f\tau) \end{aligned}$$

and we see that calculating the autocovariance function amounts to discarding the random phase information while retaining (transformed) the important parameters of the model. If

we wished to compare different stochastic processes of this type, calculating the autocovariance functions of each would place all periodic components of each in phase making the autocovariance functions properly comparable (as opposed to the raw time series which are not robustly (statistically) comparable).

The above certainly suggests a candidate statistic that might be used to analyze and compare time series. However, there are a number of drawbacks to this approach. First and foremost among these is the difficulty of interpreting an autocovariance function or sequence. This is connected to the difficulty of visualizing a typical time series that might give rise to a given autocovariance function or sequence as well as to the statistical properties of autocovariance functions/sequences. For convenience, consider the (lag k) theoretical autocorrelation

$$\rho_k = E\{(X_t - \mu)(X_{t+k} - \mu)\}/\sigma^2$$

for the discrete parameter stochastic process X_t and the corresponding (lag k) sample autocorrelation for a given realization x_t

$$\hat{\rho}_k = \frac{\sum_{t=1}^{N-k}(x_{t+k} - \bar{x})(x_t - \bar{x})}{\sum_{t=1}^{N}(x_t - \bar{x})^2},$$

the first of which, as k varies, defines the theoretical autocorrelation sequence (acs) and the second of which, as k varies, defines the sample autocorrelation sequence (acs). Percival and Walden note that "Unfortunately, it takes a fair amount of experience to be able to look at a theoretical acs and visualize what kind of time series it corresponds to." (p8). So, even with the *theoretical* acs available, its usefulness as a direct description of the process by a typical user of time series analysis is limited. On the statistical side, this problem is related to the fact that in most cases "the estimators $\hat{\rho}_k$ and $\hat{\rho}_{k+1}$ are highly correlated and compounded by the difficulty in obtaining reliable estimates of ρ_k for k large relative to N (that is, k that represent a substantial proportion of the length N of a time series). In particular, Percival and Walden note that the standard deviation of $\hat{\rho}_k$ "...depends on k and the true acs in a complicated way - typically it increases as k increases..." and that this property along with

"the correlation between nearby estimators" cause significant interpretation issues. Further, these statistical properties also make statistical inference difficult. It is for these reasons as well as the general usefulness of having additional (more robust) methods that we turn to spectral analysis.

4.3 Spectral Analysis

One of the major themes present across all of the techniques surveyed in this thesis is that of (orthogonal/direct sum/spectral) decomposition. CCA (classical) decomposes vector spaces of random variables into sets of canonical variables, each of which captures a distinct portion of the behavior of the combined set of variables. The same may be said about the singular value decomposition detailed in the appendix. Mathematically, each of these is a type of spectral decomposition (for example, the singular values form the spectrum in the singular value decomposition case). The quintessential spectral decomposition from a physical standpoint may be said to be that of light. Light, among electromagnetic wave phenomena more generally, is naturally described by its spectrum. The spectrum of light from an unchanging source contains essentially all of the fixed information about the light in the form of the power level at the various constituent frequencies, from which it is possible to make statements about the color of the light, the total power output of the light source etc. It's worth pointing out, for the analogy that will develop, that this type of spectrum also does not contain generally irrelevant details such as phase information, which is, from the point of view of arbitrary sampling, basically random and usually not even coherent (across multiple points emitting radiation from non-laser sources). Interestingly, the mathematical version of this for deterministic functions, part of the area of Fourier analysis, was developed before the electromagnetic wave model for light propagation. The importance of spectral analysis in the field of time series analysis derives from the existence of a representation (with favorable properties) for any stationary stochastic process similar to the Fourier transform

for absolutely integrable deterministic functions allowing for a natural definition of a spectrum almost precisely analogous to the spectrum detailed in the light example above. This spectrum decomposes the variance of a stationary stochastic process in contrast with the decomposition of power provided by the electromagnetic spectrum example and so represents a kind of statistical spectrum. More generally, the spectral representation itself is the gateway to a variety of powerful techniques in addition to providing a more directly usable (by most researchers) statistic than the acs, in the form of the spectrum, for dealing with stationary stochastic processes.

In order to define a spectrum for stationary stochastic processes, it is necessary to develop a representation of such processes that allows for the type of spectral decomposition of interest. Beginning with a search for forms that can represent arbitrary *realizations* of stationary stochastic processes, the obvious options are to employ either a Fourier series or Fourier integral based representation (each of which has an existing comprehensive theory and notion of spectrum), but it is also clear that the assumption of stationarity precludes these approaches in general. In particular, there is no reason for realizations of a stationary stochastic process to be periodic, and the constant variance property also precludes realizations from tending to "die down" as $t \to \infty$. A candidate representation exists in what is sometimes called the Fourier-Stieltjes integral representation, developed rigorously by Wiener (1930) (in the same extensive paper, Wiener also introduced the term "generalized harmonic analysis" for a very comprehensive theory of which we are covering only a small part here). This representation has the form

$$g(x) = \int_{-\infty}^{\infty} e^{ifx} dP(f)$$

where f represents frequency and $P(f)$ is a complex-valued function which is the Fourier-Stieljes transform of $g(x)$. It is possible to represent periodic and absolutely integrable functions using this form, but also functions that have neither property. A good example of this

is given by Priestley taking

$$P(f) = \begin{cases} \frac{1}{2} & \text{if f} = \pm 1, \pm\sqrt{2} \\ 0 & \text{otherwise.} \end{cases}$$

In this case,

$$g(x) = \cos(x) + \cos(\sqrt{2}x)$$

which is neither periodic (due to the incommensurability of $\sqrt{2}$) nor absolutely integrable. This is certainly a representation that can be used to describe realizations of stationary stochastic processes in general. Amazingly enough, it turns out that it is possible not only to use this representation on a realization by realization basis, but it is in fact possible to develop a stochastic version of this representation for stationary stochastic processes themselves. This approach, also developed by Wiener in his seminal paper of 1930, is very similar (in spirit) to Fourier's original work using Fourier transforms to solve the heat equation by separating the temporal and spatial variables. A crucial result/property of the spectral theorem for stationary stochastic processes is that it decomposes the process into a portion that is a function of time and a separate portion that is random (that is, a function on the underlying sample space) with the two portions linked by frequency. This is exactly analogous to the Fourier approach to differential equations with frequency linking space and time variables. From the separated random part, a variance spectrum may be defined. We will flesh this approach out more fully after listing the desirable properties this variance spectrum possesses[1].

The spectrum of a stationary stochastic process, the existence of which is guaranteed by the spectral representation theorem, (will be seen to) possesses the following useful properties (primarily in contrast with the acs)

1. **Decomposition of the variance of the process (by frequency):** Makes it easier to visualize the types of time series that may be generated by the process and to

[1]The material discussed in this paragraph also receives a very nice treatment in the first chapter of Priestley's book.

get a sense of any tendencies for oscillation to occur at particular frequencies. Also structurally useful.

2. **Approximate uncorrelatedness of estimates of the spectral components**: Allows for the development of good statistical tests in contrast to the highly correlated increments for the sample acvs. This stems from the "diagonalization" of the process in frequency space provided by the spectral representation.

3. **Existence of a 'variance stabilizing' transformation**: The use of a 'decibel' ($10 \log_{10}$) scale equalizes the variance of estimates of the spectrum at different frequencies (no such variance stabilizing transformation is known for the acvs).

4. **Sensitivity and interpretability**: Interpretability stems primarily from the already mentioned properties. Percival and Walden give a fairly striking example of the greater sensitivity versus the acvs in a specific case on pages 147-149. In fact, the definition and study of the spectral estimator known as the periodogram was motivated in part by the question of discovering periodicities hidden by noise in noisy data (394, Priestley).

5. **Relationship between measurement limitations and frequency domain characterization**: The limitations of measurements recorded by instruments are usually most easily expressed and understood in the frequency domain. The frequency response or 'transfer' function is often known for a given instrument.

In particular, on a statistical level, the spectrum would typically be preferred just based on the statistically oriented properties in this list. Another critical aspect of this theory is the fact that the estimation procedures that will be fleshed out in the following sections serve the dual (connected) role of improving results on the deterministic or realization by realization level.

4.4 The Spectral Representation

The approach of Percival and Walden to the spectral representation theorem is fairly illuminating. They present the basis for the theorem without formally proving it and we will briefly summarize this approach while also attempting to clarify certain aspects of it. The special case of the harmonic processes provides the necessary motivation and so we will consider the real-valued discrete time harmonic process

$$X_t = \sum_{n=1}^{N} a_n \cos(2\pi f_n t + \phi_n), \qquad t = 0, \pm 1, \pm 2, \ldots,$$

with a_n and $0 < f_n < 1/2$ representing the constant real-valued amplitudes and frequencies and the ϕ_n terms representing independent random variables uniformly distributed on $[-\pi, \pi]$. Transitioning to the complex exponential representation of the right hand side reveals how the random variables may be separated from the time variable for this process. Specifically,

$$\begin{aligned} X_t = \sum_{n=1}^{N} a_n \cos(2\pi f_n t + \phi_n) &= \sum_{n=1}^{N} a_n \frac{e^{i\phi_n} e^{i2\pi f_n t} + e^{-i\phi_n} e^{-i2\pi f_n t}}{2} \\ &= \sum_{n=-N}^{N} B_n e^{i2\pi f_n t} \end{aligned}$$

with

$$B_n = a_n e^{i\phi_n}/2 \quad \text{and} \quad B_{-n} = a_n e^{-i\phi_n}/2, \qquad n = 1, \ldots, N$$

and, $B_0 \equiv 0$, $f_0 \equiv 0$ and $f_{-l} \equiv f_l$. The (random) coefficients $B_1, \ldots, B_N$ are clearly independent and thus are mutually uncorrelated (the expectations are finite here). Upon checking, we see that $\text{cov}\{B_n, B_{-n}\} = E\{B_n^* B_{-n}\} = (a_n^2/4)(E\{e^{-i2\phi_n}\}) = 0$ for $n = 1, \ldots, N$ and, therefore, that the random variables $B_{-N}, \ldots, B_N$ are all mutually uncorrelated. The means and variances of the (random) coefficients are

$$E\{B_n\} = 0 \quad \text{and} \quad \text{var}\{B_n\} = E\{B_n^* B_n\} = a_n^2/4$$

(defining $a_0 \equiv 0$ and $a_{-n} \equiv a_n$) and therefore

$$\begin{aligned} \text{var}\{X_t\} &= E\{X_t^* X_t\} \\ &= \sum_{m=-N}^{N} \sum_{n=-N}^{N} E\{B_m^* B_n\} e^{i2\pi(f_n - f_m)t} \\ &= \sum_{n=-N}^{N} a_n^2/4. \end{aligned}$$

This decomposes the variance of the stochastic process $\{X_t\}$ into the expected squared amplitudes of the oscillations at each frequency. So, a (variance) spectrum for this process with the properties mentioned previously may be defined by

$$S(f) \equiv \begin{cases} a_n^2/4 & \text{if } f = f_n,\ n = 0, \pm 1, \ldots, \pm N \\ 0, & \text{otherwise.} \end{cases}$$

It is important to note that while the B_n coefficients contain all of the information necessary to reconstruct the process X_t, the spectrum itself does not and, paralleling the discussion above, only captures the non-random information that accounts for the variability of this process (the fixed amplitudes). In general, however, we will in fact be finding the average contribution to variance at a given frequency (this is literally how much power is expected at a particular frequency in any given realization of the process - i.e., the variance spectrum is basically the average of the "power" spectra over all realizations[2]) when the amplitudes are allowed to vary as well. This is also non-random but it is important to distinguish the harmonic process case from the general stationary process case (this will manifest itself in the use of the expectation operator in expressions defining, for example, the "integrated" spectrum below).

The above is representative of the notion of, and process of deriving, spectral representations for stationary stochastic processes in general. By introducing the Fourier-Stieltjes integral discussed previously, we can remedy, in a natural manner, the discrete nature of the frequency increments essentially imposed by the sum based expression above. Using a

[2]See Percival and Walden section 4.2 for discussion of the limitations of this statement, which is, notwithstanding these, still useful for intuition.

stochastic form of this integral, we can transform any stationary stochastic process into a form from which a clear notion of spectrum emerges. To illustrate the specifics of this integral based representation, we begin by translating the above example into it. The expression

$$Z(f) \equiv \sum_{n=0}^{N} B_n, \qquad f_n < f \leq f_{n+1} \quad \text{for} \quad n = 0, \ldots, N,$$

defines a stochastic process with orthogonal increments. This means that

$$E\{(Z(f_1) - Z(f_2))(Z(\tilde{f}_1) - Z(\tilde{f}_2))\} = 0$$

whenever $f_1 < f_2 \leq \tilde{f}_1 < \tilde{f}_2$ (i.e., on non-overlapping intervals). Given such a process, it is then possible to define, for arbitrary continuous deterministic functions $g(f)$,

$$Y = \int_A g(f) dZ(f).$$

by taking the random variable Y to be the random variable satisfying

$$\lim_{n \to \infty} E\{(\sum_{j=1}^{n} g(f_j)\{Z(f_j) - Z(f_{j-1})\} - Y)^2\} = 0$$

where $\max(f_j - f_{j-1}) \to 0$ as $n \to \infty$. This is a stochastic version of the Riemann-Stieltjes integral which we see is defined in a mean-square sense over the underlying probability space[3]. It is reasonably straightforward to check that under this definition

$$\int_{-1/2}^{1/2} e^{i2\pi ft} dZ(f) = \sum_{n=-N}^{N} B_n e^{i2\pi f_n t} = X_t.$$

With this motivation, the spectral representation theorem may now be stated (notation will be explained following the theorem statement).

Spectral Representation 1 *Let $\{X_t\}$ be a zero-mean discrete parameter stationary stochastic process. Then there exists an orthogonal process $\{Z(f)\}$ with respect to which*

$$X_t = \int_{-1/2}^{1/2} e^{i2\pi ft} dZ(f)$$

in the mean-square sense for all integers t. The process $Z(f)$ has the following properties:

[3]For more information on the role of orthogonality in this definition and for the Lebesgue-Stieltjes approach, see Doob [8].

1. $E\{dZ(f)\} = 0$ *for all* $|f| \leq 1/2$,

2. *the expression* $dS^{(I)}(f) \equiv E\{|dZ(f)|^2\}$ *defines a bounded nondecreasing function* $S^{(I)}(f)$ *with the properties we desire a spectrum to possess*

3. *for any distinct frequencies* f *and* f' *contained in* $[-1/2, 1/2]$,

$$\text{cov}\{dZ(f), dZ(f')\} = 0$$

This theorem essentially says that any discrete parameter stationary process may be represented as an infinite sum of complex exponentials with random amplitudes $|dZ(f)|$ and random phases $\arg(dZ(f))$ and that in so doing a useful variance spectrum may be defined based on this representation. Further, the representation given by this theorem has the important property that its increments (the random variables $dZ(f)$ for $f \in [-1/2, 1/2]$) are uncorrelated. That is, the correlation (or covariance) matrix of the random variables extant in this representation is diagonal (the third property in the statement of the theorem). This is in contrast to the correlation (or covariance) matrix for the process itself which is usually not diagonal ($\text{cov}\{X_t, X_t + \tau\} \neq 0$ when $\tau > 0$ typically).

The notation used in the statement of this theorem is slightly obscure. The notation $dZ(f)$ when used either within the expectation operator or the covariance form essentially means $Z(f + df) - Z(f)$ for all $df > 0$ with df small (for the covariance property, the intervals must be non-overlapping). This notation is being used to indicate (important) features of Z in the context of Stieltjes integrals. The notation $dS^{(I)}(f) \equiv E\{|dZ(f)|^2\}$ means that there is a function $S^{(I)}(f)$, called the *integrated spectrum*, with the property that $S^{(I)}(f_2) - S^{(I)}(f_1) = E\{|Z(f_2) - Z(f_1)|^2\}$ for all $f_2 > f_1$ (Doob [8], 101). It is this function that is of interest.

The relationship between the spectral representation for a given process and its autocovariance sequence may be seen by a calculation that makes immediate use of one of the most useful properties of the spectral representation, the uncorrelated increments of the $Z(f)$

process. Since

$$\begin{aligned} X_t^* X_{t+\tau} &= \int_{-1/2}^{1/2} e^{-i2\pi f' t} dZ^*(f') \int_{-1/2}^{1/2} e^{i2\pi f(t+\tau)} dZ(f) \\ &= \int_{-1/2}^{1/2} \int_{-1/2}^{1/2} e^{-i2\pi f' t} e^{i2\pi f(t+\tau)} dZ^*(f') dZ(f), \end{aligned}$$

the acvs may be written

$$\begin{aligned} s_\tau &= E\{X_t^* X_{t+\tau}\} \\ &= \int_{-1/2}^{1/2} \int_{-1/2}^{1/2} e^{i2\pi(f-f')t} e^{i2\pi f\tau} E\{dZ^*(f') dZ(f)\}. \end{aligned}$$

The uncorrelated increments yield contributions in integrals only when $f = f'$ and consequently

$$s_\tau = \int_{-1/2}^{1/2} e^{i2\pi f\tau} E\{|dZ(f)|^2\} = \int_{-1/2}^{1/2} e^{i2\pi f\tau} dS^{(I)}(f).$$

Now, consider the case where $S^{(I)}(f)$ is differentiable with derivative $S(f)$. We may then write

$$s_\tau = \int_{-1/2}^{1/2} S(f) e^{i2\pi f\tau} df,$$

that is, s_τ is the inverse fourier transform of $S(f)$. Assuming that $S(f)$ is square integrable and appealing to Parseval's theorem, it is then the case that

$$S(f) = \sum_{\tau=-\infty}^{\infty} s_\tau e^{-i2\pi f\tau}$$

and so $S(f)$ and s_τ form a fourier transform pair. Although the autocovariance sequence and the spectrum can now readily be seen to contain the same amount of information, as has already been mentioned, the spectrum possesses a number of properties making it the (usually) preferred statistic. We will also see how to calculate improved estimates of the spectrum using techniques that are not directly applicable to calculating the acvs (although clearly the just detailed relationship could be used to apply them).

4.5 Estimation

In order to make use of the notions detailed above in practice, it will be necessary to estimate the spectrum of a stationary stochastic process from finite length discrete (digital) samples.

This raises two issues. The underlying processes of interest will generally be (mathematically) continuous time processes (such as laboratory experiment evolution over time) which, when the discreteness of sampling is brought into play, will become discrete parameter processes. So the first question is how accurately will the spectrum of the derived discrete parameter process reflect that of the continuous parameter process - we will see that aliasing is the major determinant of the answer to this question. However, this only takes us from dealing with infinite-length continuous parameter process to infinite-length discrete parameter processes. The second (issue) question is, therefore, a question of how accurately the spectrum of a discrete (infinite-length) parameter process may be determined from finite-length sample subsequences of that process. This question of course also arises in the deterministic fourier theory and the results are very much the same - resolution may be lost (that is, the fine features of spectra are "smeared" out) and power may be transferred between different spectral components. The primacy of spectral techniques will be seen to depend on powerful methods for controlling and reducing the effects of these undesirable traits.

The relationship between the acvs and the spectrum gives us a starting point from which to discuss the transition from continuous parameter processes to discrete. If $\{X(t)\}$ is a continuous parameter process then

$$X_t \equiv X(t_0 + t\Delta t), \qquad t = 0, \pm 1, \pm 2, \ldots$$

is a discretized version of it based on equally spaced sampling (the most common sampling scheme) for given sampling interval $\Delta t > 0$ and time offset t_0[4]. If $\{X(t)\}$ is stationary with sdf $S_{X(t)}(f)$ and acvf $\{s(\tau)\}$ then X_t will be stationary as well with some sdf $S_{X_t}(f)$ and acvs

$$s_\tau = \mathrm{cov}\{X_0, X_\tau\} = \mathrm{cov}\{X(t_0), X(t_0 + \tau\Delta t)\} = s(\tau\Delta t).$$

So in sampling $X(t)$ we effectively also sample $s(\tau)$ with the same sampling rate. Now, since $S_{X(t)}(f)$ and $s(\tau)$ form one fourier transform pair and $S_{X_t}(f)$ and $\{s_\tau\}$ form another subject

[4]We typically assume $\Delta t = 1$ in which case things match up directly with the statement of the spectral theorem, e.g., $f_{(N)} = 1/2$.

to the relationship above, it may be shown that

$$S_{X_t}(f) = \sum_{k=-\infty}^{\infty} S_{X(t)}(f + k/\Delta t), \qquad |f| < \frac{1}{2\Delta t} \equiv f_{(N)}$$

where $f_{(N)}$ is the Nyquist frequency due to the phenomenon known as aliasing. From this, it is clear that if a continuous process is sampled at such a rate that its spectrum is zero or nearly zero for $|f| \geq 1/(2\Delta t)$ then the spectrum of the discrete process will match the spectrum of the continuous process very closely for $|f| < 1/(2\Delta t)$ (note that the discrete process spectrum is periodic). On the other hand, if the spectrum of the continuous process is large at any frequency above the Nyquist frequency due to the sampling rate chosen, the estimate provided by computing the spectrum of the discrete parameter process can be poor. It is therefore important to either be sure that the sampling rate is high enough (arguments may be based on physical properties of conductors for example etc.) or to use a filter to remove frequencies that would be aliased due to the sampling rate.

Once aliasing is accounted for, we must determine the effect of finite sample sizes on estimates of spectra for discrete parameter processes. Suppose X_t is a discrete parameter process with zero mean and a purely continuous spectrum with sdf $S(f)$ forming a fourier transform pair with the acvs of the process s_τ. Continuing along the lines used for the transition from continuous to discrete, we will now estimate the acvs for this discrete parameter process given a sample and determine the effect of so doing on the spectral estimates we may base on this. For a variety of reasons (given by Percival and Walden), time series analysts often prefer to use a biased estimator of the acvs

$$\hat{s}_\tau = \frac{1}{N} \sum_{t=1}^{N-|\tau|} X_t X_{t+|\tau|}.$$

Applying the fourier relationship between the acvs and the spectrum we may define the following basic spectral estimator (using the simple approach amounting to considering s_τ

to be 0 for $|\tau| \geq N$ in the usual fourier pair):

$$\begin{aligned}
\Delta t \sum_{\tau=-(N-1)}^{N-1} \hat{s}_\tau^{(p)} e^{-i2\pi f \tau \Delta t} &= \frac{\Delta t}{N} \sum_{\tau=-(N-1)}^{N-1} \sum_{t=1}^{N-|\tau|} X_t X_{t+|\tau|} e^{-i2\pi f \tau \Delta t} \\
&= \frac{\Delta t}{N} \sum_{j=1}^{N} \sum_{k=1}^{N} X_j X_k e^{-i2\pi f (k-j) \Delta t} \\
&= \frac{\Delta t}{N} \left| \sum_{t=1}^{N} X_t e^{-i2\pi f t \Delta t} \right|^2 \equiv \hat{S}^{(p)}(f).
\end{aligned}$$

This estimator is known as the *periodogram*. One important property of this estimator (for this work and computational reasons more generally) is that at the fourier frequencies, its value is the squared modulus of the discrete fourier transform coefficient at such frequencies scaled by $1/(N\Delta t)$. We may now examine the consequences of estimating the spectrum using this estimator.

Percival and Walden (6.3) begin by summarizing the properties of $\hat{S}^{(p)}(f)$ in contrast with the properties of the spectrum $S(f)$ itself. The properties given in the spectral representation theorem for $dZ(f)$ coupled with the desire for an unbiased estimator lead to the claim that if $S^{(p)}(f)$ were a good estimator then we should have

1. $E\{\hat{S}^{(p)}(f)\} \approx S(f)$ for all f,
2. $\text{var}\{\hat{S}^{(p)}(f)\} \to 0$ as $N \to \infty$ and
3. $\text{cov}\{\hat{S}^{(p)}(f'), \hat{S}^{(p)}(f)\} \approx 0$ for $f' \neq f$.

It turns out, however, that the periodogram doesn't in general satisfy any of these properties. In particular, 2 doesn't hold at all for frequencies at which $S(f) > 0$ and 1 and 3 only hold with certain limitations. We will only consider the first item at present and will show how to deal with both it and the second item in later sections.

To see what happens when we compute the periodogram (on average) it is convenient to rewrite $E\{\hat{S}^{(p)}(f)\}$ as a convolution of $S(f)$ with some kernel (this strategy is motivated by

the clearer case of deterministic subsequences - Percival 3.7). The result is

$$E\{\hat{S}^{(p)}(f)\} = \int_{f_{(N)}}^{-f_{(N)}} \mathcal{F}(f - f')S(f')df'$$

where

$$\mathcal{F}(f) \equiv \frac{\Delta t \sin^2(N\pi f \Delta t}{N \sin^2(\pi f \Delta t)}$$

is *Fejér's kernel* (related to Dirichlet's kernel). Percival and Walden explicitly enumerate its properties but we will simply comment that it is possible to show that as $N \to \infty$ Fejér's kernel limits to a Dirac delta function and so with $S(f)$ assumed continuous it is the case that

$$\lim_{N\to\infty} E\{\hat{S}^{(p)}(f)\} = S(f),$$

that is, $\hat{S}^{(p)}(f)$ is asymptotically unbiased. However, asymptotic properties are often not particularly useful with real data and it is the "small sample" properties of the above that are of concern. Plotting Fejér's kernel for small N (see Percival and Walden [15] p. 200) we obtain an idea of what convolution by this kernel results in. The central lobe results in what is known as *loss of resolution*. Essentially, power at frequencies near the frequency we are attempting to estimate the power at is averaged together with the power at the actual frequency in question. This causes fine details of the spectrum to be lost. An even potentially more significant problem is caused by the sidelobes of this kernel. These sidelobes show that power from frequencies across the spectrum is averaged into estimates of power at a given frequency - this is known as leakage. Leakage becomes a significant issue when the frequency response in a given band is very low compared to frequency response in other bands (a crude metric of this is given by the notion of dynamic range, this is a significant issue when a process has high dynamic range and it is the case that many real physical processes do have high dynamic range). Leakage can contribute to significant bias in estimating the spectrum of a process and in general the periodogram is often too badly biased in many cases to be a usable statistical tool (see the Thomson quote, Percival and Walden [15] p. 199). The ability to demonstrate these properties using convolutions does suggest a way to repair the

problems noted: namely replace the kernel in the convolution. This turns out to be fairly straightforward due to the well-known *convolution theorem*.

4.5.1 Convolution Theorem

The convolution theorem immediately comes to mind if we wish to replace the kernels in the various integrals above with kernels with better properties. The various forms of this theorem amount to the fact that convolution in the frequency domain amounts to multiplication in the time domain and vice versa. In particular, this allows us to replace the kernel discussed above with another kernel by multiplying sample sequences pointwise by certain other sequences with desirable properties. Such sequences are called *tapers*.

The ideal taper would of course transform to a Dirac delta function in the frequency domain. Since we are dealing with finite length sequences, this is impossible. In particular, there are no non-trivial time-limited and band-limited functions (this may be seen in the continuous case by an appeal to analyticity). So it becomes necessary to optimize. That is, to try and find the sequences of a given fixed length (time) that are most concentrated in frequency with respect to some measure of concentration. This is known as the frequency concentration problem and it has a particularly useful solution with the appropriate setup.

4.5.2 Concentration Problem

The concentration problem for the discrete time, continuous frequency case was explored by Slepian (1978)[5] and the following approach and results are due to him. In the general setting, the concentration of a time-limited signal (with $\Delta t = 1$) in the frequency domain may be written as

$$\beta^2(W) \equiv \int_{-W}^{W} |G_p(f)|^2 df \Big/ \int_{-1/2}^{1/2} |G_p(f)|^2 df$$

[5]Also discussed in Percival and Walden [15], section 3.9.

where $W < 1/2$. This is simply the fraction of the total power of the signal in the band $|f| \leq W$. In the case of an index limited signal the fourier transform is defined by

$$G_p(f) = \sum_{t=-\infty}^{\infty} g_t e^{-i2\pi ft} = \sum_{t=0}^{N-1} g_t e^{-i2\pi ft}$$

and so we have in this case that

$$\begin{aligned}
\beta^2(W) &= \int_{-W}^{W} |G_p(f)|^2 df \Big/ \int_{-1/2}^{1/2} |G_p(f)|^2 df \\
&= \int_{-W}^{W} \sum_{s=0}^{N-1} \sum_{t=0}^{N-1} g_s^* g_t e^{i2\pi f(t-s)} df \Big/ \sum_{t=0}^{N-1} |g_t|^2 \\
&= \sum_{s=0}^{N-1} \sum_{t=0}^{N-1} g_s^* \frac{\sin(2\pi W(t-s))}{\pi(t-s)} g_t \Big/ \sum_{t=0}^{N-1} |g_t|^2
\end{aligned}$$

by making use of Parseval's equality to handle the denominator and the Euler relationship after integrating. In fact, the above expression does gloss over the terms where $s = t$ and for these terms the fraction in the numerator should be replaced by $2W$. In any case, this expression may be seen to be a quadratic form and may be written as a matrix. The problem may then be seen to be of the same type as seen earlier, namely the maximization of the form $\mathbf{g}^* \mathbf{A} \mathbf{g}$ subject to $g^* g = 1$ where the terms in the symmetric $\mathbf{A}$ matrix are the middle fractional terms of the above summation in the numerator. The complete solution to this problem is again given by the eigenvalues and eigenvectors of the matrix $\mathbf{A}$. The eigenvectors are known as the *discrete prolate spheroidal sequences* (dpss's) and the eigenvalues (as may be seen from the above expressions) quantify the concentration of each. One of the most useful outcomes of this approach from a statistical point of view is the fact that an orthogonal set of sequences is generated. The trade-off in using these sequences is in the width of the central lobe they generate in the replaced kernel (which as discussed before is related to loss of resolution). Aside from this caveat (which the multitaper approach addresses), these sequences form nearly ideal tapers. It is also an important fact that there is a stable and fast method for computing the tapers themselves based on a difference equation the dpss's satisfy.

4.6 Multitaper Approach

The multitaper spectral estimator is a direct spectral estimator, as is the periodogram, but with tapering applied to the data in an effort to improve bias properties. In particular, not only do multitaper methods potentially significantly reduce bias, decrease the variance of estimates, reduce the bounds of confidence intervals and allow for a straightforward quantification of the bias of estimates made into local bias (due to the user selected bandwidth W) versus broadband bias (due to power outside the $[-W, W]$ window around a given frequency), they achieve these results automatically - that is, without the need for user intervention - and hence are ideal candidates for use in software to help researchers to cope with the ever-increasing rate at which data is collected.

The simplest multitaper spectral estimator is formed in the manner suggested by its name by using several data tapers. In particular, for a realization $X_1, X_2, \ldots, X_N$ of a stationary process $\{X_t\}$ with zero mean the definition is

$$\hat{S}_K^{(mt)}(f) \equiv \frac{1}{K} \sum_{k=0}^{K-1} \hat{S}_k^{(mt)}(f) \quad \text{for} \quad \hat{S}_k^{(mt)}(f) \equiv \Delta t \left| \sum_{t=1}^{N} h_{t,k} X_t e^{-i2\pi f t \Delta t} \right|^2$$

where $\{h_{t,k}\}$ is the kth data taper used. That is, this simple multitaper estimator uses the average of the estimates given by K data tapers. The averaging strategy in general reduces the variance of the final estimate so that as long as bias is not introduced, the computed spectrum should match more closely the actual spectrum of the process. In particular, if the $\hat{S}_k^{(mt)}(f)$ estimates are pairwise uncorrelated then the variance of $\hat{S}_K^{(mt)}(f)$ should be approximately a multiple of $1/K$ of the variance of any of these individual direct estimates. In order not to introduce bias, it is necessary that the tapers also be chosen to provide good protection against leakage. These desired properties of the $\hat{S}_k^{(mt)}(f)$ being approximately uncorrelated and relatively leakage-free follow from the use of data tapers that are approximately uncorrelated with small sidelobes for processes whose spectral density functions have certain properties. This motivates the choice of the dpss's as good sequences to use as tapers. An intuitive justification for the uncorrelatedness of the $\hat{S}_k^{(mt)}(f)$ for spectra of appropriate

type follows from the easy to see property (consult the plots in Percival and Walden [15], p. 336-338) of the dpss tapers that each successive taper tends to accentuate and attenuate different regions of the data.

Chapter 5

Spectral Canonical Correlation Analysis and Results

5.1 Spectral CCA

It is possible to precisely characterize the problem with using CCA directly on the time series of stationary stochastic processes sampled independently by writing down an expression for the sample correlation statistic using the random variables in such a process. This same approach illustrates why the spectra can be compared using correlation[1]. With these facts in hand, we may move on to the method itself.

We are now ready to combine Carroll's generalization of canonical correlation analysis and Thomson's multitaper spectral estimators to obtain a technique we call "spectral CCA" or "spectral canonical correlation analysis" (or even just SpecCCA). This method will find correlations in the spectra of variables across data sets which amounts to determining which variables amongst the different data sets share similar features. It should be noted that the use of correlations to compare spectra in some sense amounts to throwing out white noise (the mean) and looking for tendencies for there to be more or less pronounced periodicities at given frequencies. At the same time, the averaging effect of the CCA technique amounts to averaging spectral estimates both within each data set (in the form of the transformation vectors projected into the spectra of a given data set) and across the sets (in the form of the canonical variates). This is obviously advantageous from the point of view that $S(f) = E\{|dZ(f)|^2\}$. After summarizing the technique, we will also discuss how this approach remedies the problem of applying CCA to data from stationary stochastic processes.

[1]These are omitted due to time constraints.

Given data sets $\mathbf{X}_1, \ldots, \mathbf{X}_n$ with time courses running down the columns and dimensions $k \times m_i$ where $k > m_i$, we begin by calculating basic multitaper spectral estimates at the fourier frequencies for each variable. This results in transformed data sets $\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_n$ where $\tilde{\mathbf{X}}_i$ has the form[2]

$$\tilde{\mathbf{X}}_i = \begin{pmatrix} \hat{S}^{(mt)}_{\mathbf{X}_i(:,1)}(f_1) & \hat{S}^{(mt)}_{\mathbf{X}_i(:,2)}(f_1) & \cdots \\ \hat{S}^{(mt)}_{\mathbf{X}_i(:,1)}(f_2) & \hat{S}^{(mt)}_{\mathbf{X}_i(:,2)}(f_2) & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}.$$

We then perform Carroll's canonical correlation analysis on these transformed data sets (using the numerically stable method discussed above). The result is a set of canonical variates in a matrix $\mathbf{Z}$ and the transformation vectors (containing the canonical coefficients) in the matrices $\mathbf{A}_1, \ldots, \mathbf{A}_n$. The canonical variates (or canonical spectra) represent spectra most similar to certain linear combinations of the spectra from each data set (they are the most similar across all the data sets). These variates also represent averaged spectral estimates from across all the data sets (see Carroll's comment in the case of two data sets for motivation for this interpretation). The transformation vectors in a given $\mathbf{A}_i$ indicate, on the one hand, which variables are most closely related to one another and to other variables across the data sets (in the sense of their spectra), and on the other hand, may be used to produce averaged spectra for a given data that are most similar to spectra contained in the other data sets. Finally, it should be noted that the above may easily be applied to data sets where $k \leq m_i$ by using a method of data compression such as the SVD or principal component analysis. It is important to note that data compression should be performed after tapering.

5.1.1 Results

We now give a brief example of some results obtained. The figures contain graphical representations of the transformation vectors found (which are images in this case) when the

[2]We use MATLAB notation for convenience - e.g., $\mathbf{X}(:,1)$, which means the first column of $\mathbf{X}$.

technique was applied to four videos taken of biological processes. The plots are of the corresponding "canonical spectra." The successive vectors and different spectra plotted within the same figure correspond to the four different data sets used. The figures are (unfortunately) on the pages after the explanatory text.

In figure 5.1, the blue line corresponds to vector 1, the red to vector 2, the black to vector 3 and the magenta to vector 4. Vectors 1-4 are transformation vectors associated with the "averaged" spectra plotted and the averaged spectra correspond to the first canonical variate. Ellipses have been added to try to pick out some similarities between the canonical spectra over part of the frequency range. The black and magenta seem to follow each other most closely (corresponding to vectors 3 and 4).

Note that the canonical spectra aren't exactly normal spectra in that they may be negative. The CCA process (and SVD used for data compression) is very effective at separating out power at different frequencies, exploiting sometimes small differences between the original spectra. The effect is that it can actually drive the "power" in the canonical spectrum down at a given frequency while driving the next frequency up. Often, of course, this process results in negative "power" at some frequencies. This is simply the result of the process selecting for power at some frequencies and against power at others within any given canonical variable. Power at frequencies selected against in any given variable is generally being put into other canonical spectra and is therefore being subtracted out of that given variable, sometimes causing the spectrum to be negative at those points.

In figure 5.2, the blue line corresponds to vector 5, the red to vector 6, the black to vector 7 and the magenta to vector 8. Vector 5 corresponds to vector 1 in the first set etc. Again, the last two data sets appear most similar in terms of their canonical spectra.

The above correspond to the first two canonical variates using multitaper spectral estimates from the four data sets with $NW = 4$.[3] The data was tapered producing the multitaper

[3]The time/bandwidth product, N is the number of time points and W is the bandwidth from the frequency concentration problem.

estimates, an SVD was performed and finally a CCA was performed. The remaining transformation vectors and spectra did not seem interesting. Due to significiant power at low frequencies and possible loss of resolution, it seemed important to also try the analysis with $NW = 2$. With some additional work, approximately five sets of transformation vectors and spectra contained features of interest. This work may be published in the future.

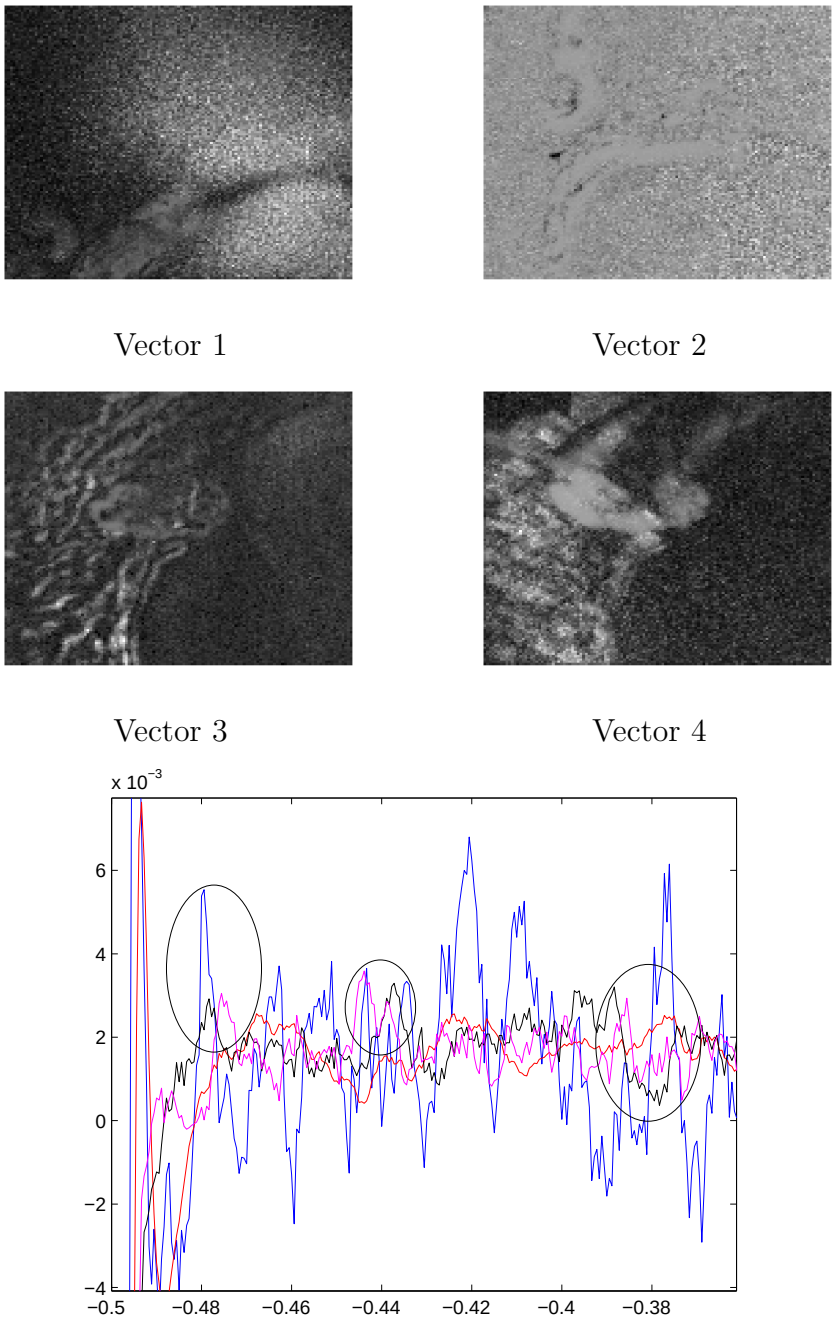
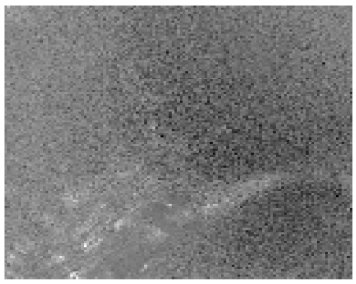
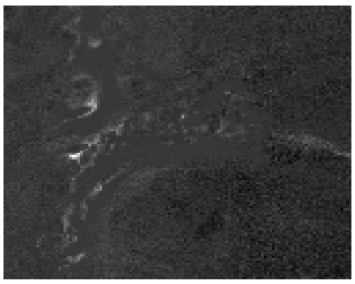


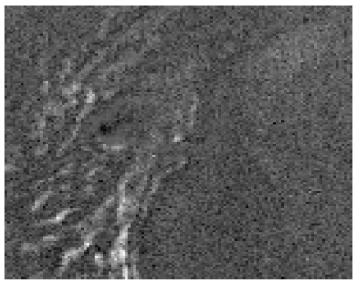
Figure 5.1: Canonical Spectra 1



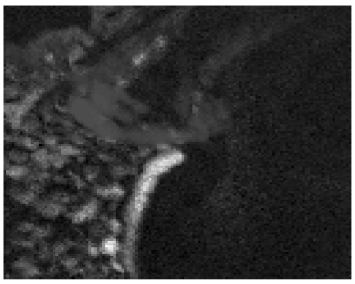
Vector 5



Vector 6



Vector 7



Vector 8

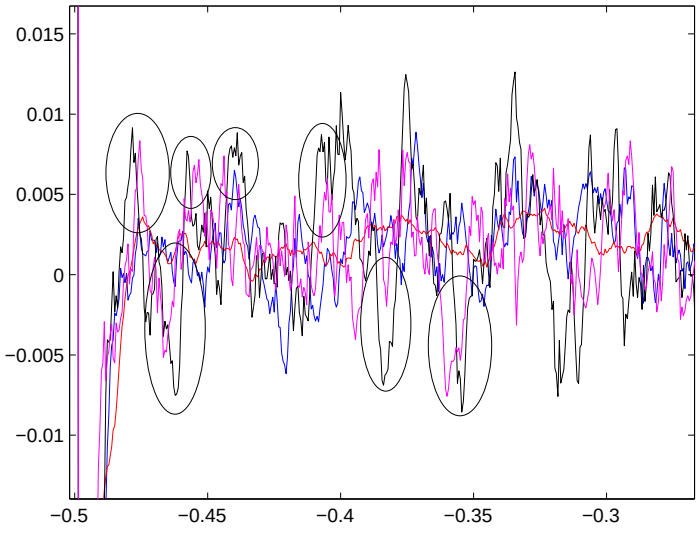


Figure 5.2: Canonical Spectra 2

Bibliography

[1] Anderson, E., Bai, Z. et. al. 1999. LAPACK Users' Guide. 3rd ed. SIAM, Philadelphia, PA.

[2] Anderson, T.W. 2003. An Introduction to Multivariate Statistical Analysis. 3rd ed. John Wiley & Sons, Inc., Hoboken, NJ.

[3] Carroll, Douglas J. 1968. Generalization of Canonical Correlation Analysis to Three or More Sets of Variables. *Proceedings, 76th Annual Convention, APA.*

[4] Casella, George; Berger, Roger L. 1990. Statistical Inference. Wadsworth Publishing Company, Belmont, CA.

[5] Cramér, Harald. 1946. Mathematical Methods of Statistics. Princeton University Press, Princeton, NJ.

[6] Demmel, James W. 1997. Applied Numerical Linear Algebra. SIAM, Philadelphia, PA.

[7] Demmel, J., and Kahan, W. 1990. Accurate Singular Values of Bidiagonal Matrices. SIAM *J. Sci. Stat. Comput.*, 11:873-912.

[8] Doob, J.L. 1953. Stochastic Processes. John Wiley & Sons, Inc., New York, NY.

[9] Goldman, Dana P., and Smith, James P. 2002. Can Patient Self-Management Help Explain the SES Health Gradient? *Proceedings of the National Academy of Sciences.* 99(16):10929-34.

[10] Golub, Gene H., and Van Loan, Charles F. 1996. Matrix Computations. 3rd ed. The Johns Hopkins University Press, Baltimore, MD.

[11] Kettenring, J.R. 1971. Canonical Analysis of Several Sets of Variables. *Biometrika*. 58(3):433-451.

[12] Kincaid, David, and Cheney, Ward 2002. Numerical Analysis: Mathematics of Scientific Computing. 3rd ed. Brooks/Cole, Pacific Grove, CA.

[13] Mitra, P.P., and Pesaran, B. Analysis of Dynamic Brain Imaging Data. *Biophysical Journal*. 76:691-708.

[14] Panik, Michael J. 2005. Advanced Statistics from an Elementary Point of View. Elsevier Academic Press, San Diego, CA.

[15] Percival, Donald B., and Walden, Andrew T. 1993. Spectral Analysis For Physical Applications: Multitaper and Conventional Univariate Techniques. Cambridge University Press, Cambridge, UK.

[16] Priestley, M.B. 1981. Spectral Analysis and Time Series. Academic Press Inc., New York, NY.

Appendix A

Singular Value Decomposition

A.1 Introduction

The *singular value decomposition* provides us with a factorization of a rectangular matrix into a product of orthogonal or unitary matrices and a diagonal matrix[1]. It is essentially a generalization of the spectral decomposition for square symmetric matrices. It is extremely useful theoretically as well as practically and is, due to the computational properties of certain methods for computing it, essential in applying certain statistical techniques to data (such as principal component analysis and canonical correlation analysis). The next few paragraphs define and motivate this decomposition and the following sections detail its usefulness computationally.

Given an $m \times n$ (data) matrix A, the singular value decomposition (SVD) is a factorization of A as $A = USV^*$ with U an $m \times m$ unitary matrix, V an $n \times n$ unitary matrix and S an $m \times n$ matrix with nonzero entries only on its diagonal. Given such a factorization, we readily see that $A^*A = VS^*SV^*$ and $AA^* = USS^*U^*$, that is, the columns of U and V eigenvectors of the respective sample covariance matrices. So, in fact, the SVD gives us not only the eigenvectors of the two sample covariance matrices, but in so doing, also gives us the projections of these eigenvectors in the data (ie $AV = US$ and $AU^* = SV^*$). This factorization is also important as we may compute variations of it such as the "thin SVD" and so avoid completely orthonormally decomposing the null-space of the matrix A.

[1]This appendix contains a rough outline of some important numerical material and may not adhere very well to the notation set up in the body. Readers may consult Demmel [6] or Golub and Van Loan [10] for a more careful exposition and more information.

A number of simple proofs of the existence of the singular value decomposition for an arbitrary matrix are known. To motivate the existence of this decomposition based on the better known spectral theorem, consider the computation $(AA^*)AV = A(A^*AV) = AVD$, which demonstrates that the columns of AV are eigenvectors of the matrix AA^* so we ought to be able to write $AV = US$ for some unitary matrix U and some diagonal matrix S. To analyze the scalars S must contain, computing the norm of a given column of AV, say $A(V_i)$ where V_i denotes the i**th** column of V, we see that $\|AV_i\|^2 =< AV_i, AV_i >=< V_i, A^*AV_i >=< V_i, D_iV_i >= D_i < V_i, V_i >$ which implies that $\|AV_i\| = \sqrt{D_i}\|V_i\|$, that is, the singular values are the square roots of the eigenvalues of the matrices AA^* etc.

When working with data and computers, two additional problems arise that are not encountered in the "purely" mathematical realms. The first of these is measurement error, a problem that was previously mentioned in the statistics section. Error of this type may be dealt with using statistical methods but is also amenable to treatment through the selection of methods with good numerical properties. In particular, The second problem occurs due to the use of digital computers which are limited to finite-precision computation.

The first is measurement error which could potentially effect our ability to even accurately determine parameters of the (actually) sampled population. The second is the effect of finite-precision computation and storage of numbers within computers on the solutions obtained using computers. Due to these two factors, it is very important to use computational algorithms that exhibit a form of Lipschitz continuity (with a small multiplier). That is, it would be quite undesirable to use algorithms for which the results can swing wildly with even small changes in the inputs as, given that error is practically guaranteed to be present in our data, the validity of any results thus obtained would be highly suspect. Further, it is important that the computations themselves when carried out on a finite-precision computer do not introduce significant error. An algorithm for which small changes in inputs can only (in relevant cases) result in small changes in the outputs is called *numerically stable* or simply *stable*. This notion may also be used to justify the claim that an algorithm itself car-

ried out on finite-precision hardware does not introduce significant error by introducing the complementary notion of *backward stability*. Alternatively, *forward stability* may be shown directly. The specifics will be offered after a motivating example. It is important to note that the simple computation used to motivate the existence of the SVD lacks stability and is not robust if implemented on a computer.

A.2 Example

As mentioned previously, the computation given to motivate the existence of the SVD is unsuitable for the purpose of actually computing the SVD. Specifically, for our purposes, it is unwise to actually form the sample covariance matrix. To see this, consider the following computations which are carried out in a floating point number system using base 10 with 6 digits worth of precision (the exact meaning of this is explained in the next section). The function fl maps a matrix with entries in $\mathbb{R}$ to its floating point representation in this system.

Let

$$A = \begin{pmatrix} 1 & 1 \\ 10^{-3} & 0 \\ 0 & 10^{-3} \end{pmatrix}$$

which, for our purposes, could be a data matrix with each column corresponding to the time course of one variable so that computing the sample covariance matrix would involve the formation of A^TA. The result of actually computing this matrix in the given floating point system is

$$fl(A^TA) = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

which has eigenvalues 1 and 0 (implying singular values of the original matrix of 1 and 0) regardless of how high a precision we use to compute them. However, it is clear by forming A^TA without rounding that the smaller of the two eigenvalues is 10^{-6} and so the original matrix has the singular value 10^{-3} as its smaller singular value. So, simply forming the matrix A^TA results in an immediate loss of precision of absolute size 10^{-3} which is $100\times$

greater than the unit of roundoff in the given floating point system and represents a relative error of size 1 indicating essentially no significant digits in the solution obtained. Hence, any algorithm based on forming A^TA or AA^T cannot (reasonably) be said to be backward stable.

The above may be contrasted with the behavior of the algorithms that will be specified later in this section. With the same floating point system, the bidiagonalization of A yields

$$fl(B) = \begin{pmatrix} -1 & -1 \\ 0 & 0.00141421 \\ 0 & 0 \end{pmatrix}$$

which has the smaller singular value .000999997 computed using the high accuracy SVD method of Demmel and Kahan. In this case, this is basically full relative precision. In general, for this case, we are guaranteed an answer in the interval $[.00099, .00101]$ which is equivalent to a relative error of about 10^{-2}, essentially indicating at least 2 significant digits in the solution, significantly better than the result when A^TA was formed.

The matrix used in this example is also used in a similar way by Golub and Van Loan in their discussion of the least-squares problem. Their example illustrates the numerical difficulties arising when the normal equations are used to solve the least-squares problem. This is relevant to material that will be covered in the CCA section. This is also discussed by Demmel who notes that the normal equations method can lose twice as many digits as methods based on the SVD and that the normal equations method isn't even necessarily stable (p. 118). There is a very nice discussion of the use of the SVD in the case of the rank-deficient least-squares problem given by Demmel in section 3.5.1 indicating that dropping singular values below some tolerance can change a very ill-conditioned problem to a well-conditioned problem in an appropriate manner. (*) This is important in the computation of the transformation vectors in the CCA section (see also Golub [10] p. 243, 250 - about Hansen regularization - and 571).

A.3 Floating Point Numbers and Error Analysis

In order to describe more precisely and generally the accuracy of computational methods carried out with finite-precision, a few additional concepts are needed. In this section, a few basics will be given, followed by a model for floating point number systems and, finally, some reference material for matrix norms.

A.3.1 Error

Let $x \in \mathbb{R}^n$ for some $n \geq 1$ (when n is one, x is a scalar). Typically, an approximation to such a vector will be represented by placing a hat over the name for the vector, as in $\hat{x}$. To quantify the relationship between the two, either the notion of *absolute error* or that of *relative error* may be used. The *absolute error* (for a given vector p-norm) is defined by the expression

$$\|\hat{x} - x\|_p$$

and the *relative error* (for a given vector p-norm) by

$$\frac{\|\hat{x} - x\|_p}{\|x\|_p}.$$

One important comment is that results involving relative error are typically preferred, especially when the norm involved is the (scalar) absolute value. Relative error in this case essentially indicates how many of the nonzero digits in a nonzero computed result are correct (regardless of size). This was alluded to before and a more exact statement, following Golub and Van Loan [10] p. 54, is: if

$$\frac{|\hat{x} - x|}{|x|} \approx 10^{-p}$$

then $\hat{x}$ has approximately p correct significant digits. This is in contrast to absolute error bounds where the number of correct digits in a computed result then depends entirely on how large the correct value is in a particular case (specifically, if the correct value is small, there may be no useful guarantee on the error at all).

A.3.2 Floating Point Numbers

Computers are, of course, unable to represent arbitrary real numbers. Instead, floating point number systems are usually employed as an approximation to the real number system. A floating point number system F is essentially characterized by four integers (Golub and Van Loan [10] p. 61), the *base* β, the *precision* t and the *exponent range* $[L, U]$. Nonzero numbers in F are of the form

$$\pm .d_1 d_2 \ldots d_t \times \beta^e \qquad 0 \leq d_i < \beta,\ d_i \neq 0,\ L \leq e \leq U$$

and so (following a comment of Demmel [6] p. 9), floating point representation is analogous to scientific notation based representation. To distinguish between an exact (real) computation and a floating point result, the function fl is typically introduced. For our purposes, this function will have the signature $fl : \mathbb{R} \to F$ and, given $x \in \mathbb{R}$, will be defined by: $fl(x)$ is the closest number to x representable in F, rounding in a specified manner (depending on the floating point system) in the case of a tie.

With this background, the key features the relevant floating point systems possess may be illuminated. The symbol $\odot$ will be used to denote any of the arithmetic operations $+, -, \times, \div$. If a and b are floating point numbers, $a \odot b$ will stand for the exact result of such a computation (implicitly an element of $\mathbb{R}$), $fl(a \odot b)$ will be the floating point result and $(a \odot b) - fl(a \odot b)$ is often termed the *roundoff error*. The *maximum relative representation error* may be approximated based on the following computation

$$\begin{aligned} \frac{|fl(\beta^0 + \frac{1}{2}\beta^{-t+1}) - (\beta^0 + \frac{1}{2}\beta^{-t+1})|}{|\beta^0 + \frac{1}{2}\beta^{-t+1}|} &= \frac{\frac{1}{2}\beta^{-t+1}}{1 + \frac{1}{2}\beta^{-t+1}} \\ &< \frac{1}{2}\beta^{1-t} \end{aligned}$$

and so the constant ϵ is typically defined by $\epsilon = \frac{1}{2}\beta^{1-t}$ and termed the *machine epsilon* or *machine precision*. It then follows that $fl(a \odot b) = (a \odot b)(1 + \delta)$ where $|\delta| \leq \epsilon$. Therefore,

$$\frac{|fl(a \odot b) - (a \odot b)|}{|a \odot b|} \leq \epsilon \qquad a \odot b \neq 0,$$

that is, the relative error in any given single arithmetic operation is small. According to Demmel, "this is the most common model for roundoff error analysis."

Details have been left out, including the handling of overflows and underflows, the phenomenon of catastrophic cancellation and other possibilities for rounding, in favor of presenting the most relevant features of the model underlying the IEEE floating point standard (754 - implemented in virtually all modern computers including those with Intel, AMD, IBM Power, SUN SPARC CPUS etc.). The analyses on which the following stability results depend have, seemingly in all cases, been carried out with respect to the above model and are valid on any computer implementing a system conforming to this model (in particular, the IEEE standard). Demmel [6] provides some additional illuminating comments regarding the IEEE standard on p. 13 of *Applied Numerical Linear Algebra*.

A.3.3 Matrix Norms

All that will be offered here is the definition of the class of norms from which the norms used later are drawn. Consult Golub and Van Loan for more detail. The matrix p-norm of a matrix A is defined by

$$\|A\|_p = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}$$

which it should be noted utilizes the corresponding vector p-norm.

A.4 Algorithms and Stability

MATLAB uses the LAPACK function DGESVD to compute the SVD of double-precision floating-point matrices (the default floating-point matrix type in MATLAB). This driver function makes use of a number of other LAPACK functions to compute singular value decompositions by bidiagonalizing input matrices (when necessary) and then applying the "implicit zero-shift QR algorithm" of Demmel and Kahan [7]. The bounds on the resulting errors are as follows (as detailed in the LAPACK users guide)

Let $A = U\Sigma V^T$ where $U\Sigma V^T$ is the exact singular value decomposition of A. The computed SVD, $\hat{U}\hat{\Sigma}\hat{V}^T$, will be nearly the exact SVD of a perturbed matrix $A + E$ (an example of backward stability) with $\|E\|_2/\|A\|_2 \leq p(m,n)\epsilon$, where $p(m,n)$ is a modestly growing function of m and n, so that

$$|\hat{\sigma}_i - \sigma_i| \leq p(m,n)\epsilon\sigma_1.$$

The qualification of "nearly the exact SVD" stems from the use of finite precision representations of the orthogonal matrices. As noted in the LAPACK users guide, this bound means that "large singular values (those near σ_1) are computed with high relative accuracy and small ones may not be."

The singular vectors computed by LAPACK also satisify a couple of error bounds. First, they are nearly orthogonal to machine precision whether or not they are close to the true singular vectors. That is,

$$|\hat{u}_i^T \hat{u}_j| = O(\epsilon)$$

when $i \neq j$. Second, the closeness of a computed singular vector corresponding to the singular value σ_i depends on the separation between σ_i and the rest of the singular values of the matrix. Specifically, the following approximate bound holds

$$\theta(\hat{u}_i, u_i) \lesssim \frac{p(m,n)\epsilon\|A\|_2}{g_i}$$

where $g_i = \min_{j \neq i} |\sigma_i - \sigma_j|$ is the absolute gap between σ_i and the nearest distinct singular value.

Finally, it is worth pointing out that the error bounds are much improved when dealing with bidiagonal matrices which is also detailed in the LAPACK users guide.

As noted by Golub and Van Loan [10] (p. 65), quoting Wilkinson, a priori error bounds themselves are often less precise than they could be and it may be argued that it is the

potential instabilities revealed in the analysis that are most important. For this reason and for the sake of completeness, a brief overview of some of the details leading to the above bounds follows.

Given a matrix A for which the SVD is desired, the Demmel and Kahan algorithm consists of the following phases

1. Compute orthogonal matrices P_1 and Q_1 such that $B = P_1 A Q_1$ is a bidiagonal matrix (a type of banded matrix with nonzero entries only on its diagonal and immediate superdiagonal).

2. Perform implicit (zero-shift) QR iteration to remove the superdiagonal entries of B while preserving its singular values. This is equivalent to the formation of orthogonal matrices P_2 and Q_2 such that $P_2 B Q_2 = \Sigma$ is diagonal with the singular values of B on its diagonal. The actual computation is carried out iteratively without explicit QR-factorizations (in forming P_2 and Q_2) in the same manner as the traditional Golub-Kahan algorithm, but with all shifts set to 0 and organized in such a way as to compute all intermediate matrices as well as the final result to nearly full machine precision.

The paper of Demmel and Kahan [7] discusses the fact that the singular values of a bidiagonal matrix are determined very precisely by the entries of such a matrix and the above algorithm is constructed in such a way as to determine these to nearly the precision with which they are determined by the data. The stability of algorithms such as the above comes generally from the use of orthogonal matrices as detailed in Demmel sections 3.4.3 and 3.4.4. The specific precision achieved by the Demmel and Kahan algorithm depends on an accurate Givens rotation construction function detailed in section 3 of Demmel and Kahan and the manner in which error propagates within the QR algorithm as given by Lemma 6 of section 8 yielding bounds on the error in each step which are given in Lemma 7 of section 8. Finally, this is combined with what Demmel and Kahan term the "central result" of section 2, a theorem that details the change in the eigenvalues of a related type of matrix depending on

perturbations, to conclude with Theorem 6 bounding the relative error to nearly machine precision. This error bound applies only to bidiagonal matrices in general and the looser error bound given above is due primarily to the (possible) introduction of error during the initial phase of the SVD computation involving a dense matrix, the bidiagonalization phase.