Principal Components Analysis

Overview

Principal components analysis (or PCA in informal circles) is a method of reducing a very large number of data points down to a manageable size. One ERP often contains between 100 and 300 or more data points (averaged voltage samples). Most studies have more than one subject, so the resultant data set can be very large; e.g., with only 30 subjects, each having only one ERP of 200 data points, the data set contains 6000 numbers--a veritable multivariate nightmare. It is useful to think of such a **D**ata set as a matrix:

$$\mathbf{D}_{Nxn} =$$

Subject #1	$[t_0, t_1, t_2, \ldots, t_{n-1}$	Where N = Number subjects
Subject #2	$t_0, t_1, t_2, \ldots, t_{n-1}$	n = Number sample points per average
Subject #3	$t_0, t_1, t_2, \ldots, t_{n-1}$	t = voltage at time point 0, 1, ...
...	...	
...	...	
Subject #N	$t_0, t_1, t_2, \ldots, t_{n-1}]$	

Within each of the N rows, the data points represent the sample values that comprise a subject's ERP. Within each of the n columns, all data points correspond to the same point in time for different subjects. The goal of PCA is to reduce the number of columns in this matrix, from 200 to approximately 3 or 5, in such a way that most of the meaningful information in the original ERP waveform is preserved. In the case of this example, this amounts to reduction of the data set to $1/40^{th}$ of its original size while not losing meaningful information. The reduced matrix has scores for each subject on several hypothetical variables called components; the matrix is called a component **S**core matrix and looks like:

$$\mathbf{S}_{Nxm} =$$

Subject #1	$[s_1, s_2, s_3, \ldots, s_m$	Where N = Number subjects
Subject #2	$s_1, s_2, s_3, \ldots, s_m$	m = Number of components
Subject #3	$s_1, s_2, s_3, \ldots, s_m$	s = score on component 1, 2, ...
...	...	
...	...	
Subject #N	$s_1, s_2, s_3, \ldots, s_m]$	

A way to reduce this same data set without PCA would be to assign each subject a score on essential bumps in the waveform. For example, instead of 200 voltage values per subject, there could be five amplitude values for each subject: P1, N1, P2, N2, P3. Although such a data reduction would tell us little about the latencies for a given subject, much of the meaningful amplitude information for each subject would be summarized by these five scores. PCA summarizes information in a similar fashion, but does so in a way that captures the maximal amount of information from the original data set with the fewest scores possible.

No, Really? How Can This Be Possible?

This can be accomplished because many of the data points within an ERP waveform are correlated with one another. The ERP data set is not unlike a questionnaire with many questions tapping the same construct; much of the information is redundant. For example, the person who endorses the item "I hate loud noises" would probably be very likely to endorse similar items such as "I dislike the sound of an air-hammer", "My lab instructor speaks too loudly", “Thumper low-riders should be banned”, and "The sound of an airplane passing overhead is aversive." Responses to all these items are highly intercorrelated and may be manifestations of a higher order construct, perhaps sensitivity to noise.

Now consider the ERP data set. Much of the information contained within each ERP is redundant. If a subject shows a large P300 peaking at a latency of 500 milliseconds, the sample values on either side of this 500 millisecond peak will be similarly large. If the subject has a small P300 amplitude, the adjacent time points will have small values. All the sample values surrounding 500 milliseconds are therefore highly intercorrelated. Rather than examining each of the individual data points in the region of P300, we might describe the ERP in terms of a higher-order construct, namely P300 amplitude.

Now such a description may seem rather obvious, and you may ask "why do we need some fancy statistical hocus pocus to reveal the obvious?" Good question.

1. First, some of the important information may not be obvious. There may exist several processes that overlap in time to produce the observed waveform. For example, the observed positive peak at 500 milliseconds may be the result of two or more

positive voltage fields summing at the scalp very closely in time (call them P475 and P550 for example). To the extent that these temporally overlapping processes differ between persons, PCA can uncover these processes by examining the pattern of intercorrelations between various time points and extracting orthogonal (uncorrelated & independent) components.

2. Second, PCA produces a reduced set of data that maximally captures the information available in the original large data matrix. (You might hit upon such an efficient reduction by eyeballing the data if you were given several years of trial and error; you'll finish graduate school faster if you try PCA). PCA extracts components in decreasing order of amount of variance accounted for in the original data set (#1 is max, then #2 ...); usually 3 to 5 factors account for most of the variance in an ERP data set, often over 90% of the original variance. To account for all of the original variance, one would need n components--the same number as variables. For our example of 200 time points, assume that five components account for 90% of the variance in the original data set; the remaining 195 components collectively account for only 10%. These last 195 components are therefore ignored in the interest of parsimony. (Think of wealth distribution in the United States and you've got an apt metaphor.)

But What do These Component Scores Mean?

Another Matrix!. Before this question can be answered, I must tell you "the rest of the story." In addition to the component score matrix $\mathbf{S}_{Nxm}$, there exists a component **L**oading matrix which tells you how much each time point "loads" on a particular component:

$\mathbf{L}_{mxn}$ =

Component #1 $[l_0, l_1, l_2, \ldots, l_{n-1}$
Component #2 $l_0, l_1, l_2, \ldots, l_{n-1}$
Component #3 $l_0, l_1, l_2, \ldots, l_{n-1}$
... ...
Component #m $l_0, l_1, l_2, \ldots, l_{n-1}]$

Where m = Number of components
n = Number sample points per average
l = component loading for time point 0, 1, ...

So, the data reduction is not as large as I may have led you to believe because there exists this extra matrix $\mathbf{L}_{mxn}$. In the case of 30 subjects and 200 data points per subject, 6000 data points clutter the original data matrix; a PCA yielding 5 components produces 30 * 5 component scores + 5 * 200 component loadings = 1030 data points. The savings are still remarkable. Moreover, it is usually only the component score matrix $\mathbf{S}_{Nxm}$ which is used in subsequent analyses.

The Meaning of Component Scores. One way to interpret the meaning of the components is to plot the loadings of the various time points for each of the components (from the loading matrix $\mathbf{L}_{mxn}$). For example, if most time points have near-zero loadings for component #1, but the time points in the region of 500 milliseconds have high positive loadings on component #1, one might conclude that this is the P500 component (a late P300). Therefore subjects who have a high score on component #1 probably have large P500's; subjects with small scores on component #1 would have small P500's. You could also imagine a situation where a control group has a P300 and a "pathological" group has a P500; PCA might extract a component representing each of these. The pathological group would have higher component scores on the P500 component, and the control group would have higher scores on the P300 component. The following plot shows the loadings of time points on various components in a study by Donchin & colleagues. In this plot, component #2 appears to be a P300-type component (at latency 520 milliseconds).

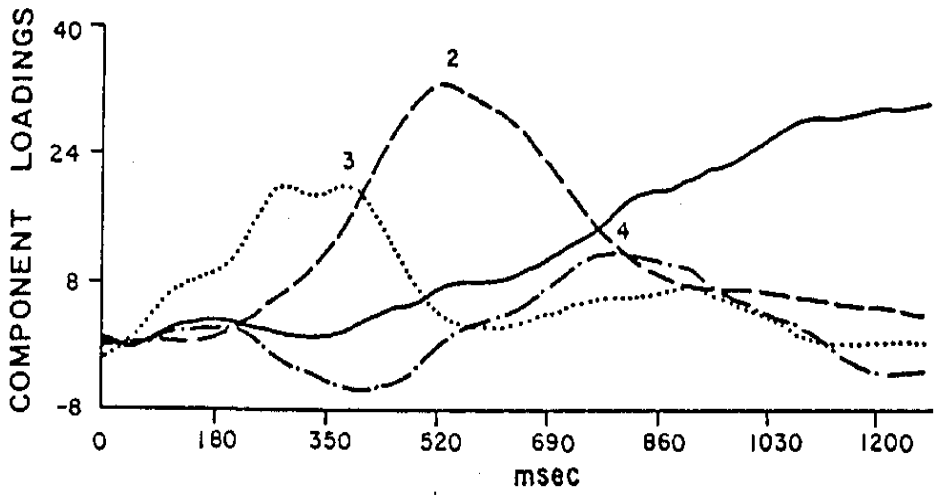

Figure 10-4. Plot of four sets of component loadings derived from a principal-components analysis (PCA) of an ERP data set. Each of the component loading vectors is composed of 128 points corresponding to 128 time points (100-Hz digitizing rate) in the waveforms.

But How Do I Know How Well PCA Did? Can I Reconstruct the Original Data?

To approximate the original **D**ata matrix, one simply post-multiplies the component **S**core matrix by the **L**oading matrix:

$$\mathbf{D}_{Nxn} \sim= \mathbf{S}_{Nxm} * \mathbf{L}_{mxn}$$

This reconstructed **D**ata matrix will differ slightly from the original **D**ata matrix because not all n components are used. To the extent that the m components account for most of the variance in the original data set, the reconstructed data matrix will closely approximate the original data matrix. This reconstruction is basically a multiple regression prediction equation, predicting a score based upon a weighted sum of other variables. These variables are not observed variables, but the PCA extracted component scores (**S** matrix), and the weights are the loadings (**L** matrix).

Two basic approaches: R & Q

Correlation vs Covariance as the Starting Point. PCA begins with the computation of either a correlation or covariance matrix. The correlation matrix is simply the covariance matrix that has been standardized by setting all variances equal to one. For ERP's, the covariance matrix is often preferred because correlations obscure some of the actual variations. In an ERP experiment, everything is fairly highly correlated with everything else. For example, an ERP with a small P300 may correlate about .80 with an ERP with a large P300, simply because the two waveforms are quite similar in morphology. Yet for the purposes of most analyses, these differences between these two waveforms are important. It is only the covariance matrix which adequately reflects this difference. By contrast, if you wish to minimize the effects of deviant waveforms, use the correlation matrix.

Correlation matrixes are used in conventional factor analyses with variables differing in units of measure (e.g., inches of rain, severity of depression, number of mental health admissions). This avoids the problem of "comparing apples and oranges."[1] Covariance matrices are appropriate when all variables use the same units of measure (e.g., microvolts).

R-PCA. The matrices **D**, **S**, & **L** illustrated above reflect the R-PCA approach. This approach seems to be more commonly used. R-PCA involves computation of time-point x time point matrix from which the PCA follows. The R-PCA correlation or covariance matrix summarizes the relationships among the time-points across the voltage-time function; it contains the correlation of each time point with every other time point. The resultant correlation matrix is then of dimensions n x n and contains n^2 entries. For an ERP with 200 sample points, this amounts to 40,000 entries.

To review, using the R-PCA method, each subject has a score on several hypothetical components--parts of an ERP waveform; each time point loads on these hypothetical components. The meaning of the components is determined by plotting the loadings of each time point as a function of time.

Q-PCA. Q-PCA involves computation of waveform x waveform correlation or covariance matrix, which is essentially a matrix of unlagged cross correlations between various subject's ERP's. This correlation or covariance matrix summarizes the relationship among waveforms across subjects. The resultant correlation matrix is then of dimension N x N and contains N^2 entries. Instead of 40,000 entries, this matrix would only have 900 entries if there were 30 subjects in the original matrix.

For those of you who are curious, the Q-PCA method decomposes the data matrix $\mathbf{D}_{nxN}$ (now many rows, few columns) into a component score matrix $\mathbf{S}_{nxm}$ and a component loading matrix $\mathbf{L}_{mxN}$. This means that each time point has component scores, and each subject has loadings on the various components. In other words, each of the components is now like a hypothetical ERP waveform, and each subject "loads" on some or all of these hypothetical waveforms. Sketch out the matrices yourself and perhaps you'll see what's happened.

Stability of Data Reduction. A general guideline in performing any form of factor analysis is that there should be at least 10 observations for each variable in order to extract reliable factors. (In the case of the R-PCA method, 10 subjects [N] for each time point [n]). Because only reliable variance may be held in common between variables, only reliable variance may extracted as a higher-order factor. (There is no higher order common factor of unreliable variance!) Stable (reliable) correlations are produced with a larger number of observations; unstable correlations based upon too few observations will therefore lead to less reliable factors which may not replicate using a fresh sample. For the R-PCA method, with 200 variables and 30 observations, there exist .15 observations per variable. For the Q-PCA method, with 30 variables (subjects) and 200 observations (time points), there exist 6.67

[1] But see Sandford , S.A. (1995). Apples and Oranges-A Comparison. *Annals of Improbable Research (AIR), 1(3).* Available at http://www.improbable.com/airchives/paperair/volume1/v1i3/air-1-3-apples.html

observations per variable. The Q-PCA method thus is more in line with this general guideline of 10 observations per variable, and may therefore be more likely to extract reliable factors.

In partial defense of the R-PCA method is that ERP time points are different and perhaps more reliable than conventional variables. Another difference is that ERP time points are highly intercorrelated, with correlations in the range seldom seen in psychological research; such large correlations are unlikely to be solely chance phenomena. Some time points (e.g., P300) will prove more reliable because there exists more variance at these time points. Additionally, for a component like P300, there is more signal relative to the background noise. It seems likely that the more reliable time points will load highly on first or second principal components, and less reliable time points will not load as highly on those components.

Applications of R's & Q's; My Own Speculations. If individual differences are of interest, use the Q-PCA technique, examining intercorrelations of waveforms. You may find two, three, or even more patterns of ERP waveform. By examining subjects with similar waveforms, you may be able to gain a better understanding of the variables which influence the observed variations in waveforms. Such variations are often glossed over or obscured when investigators present grand-mean waveforms.

If every waveform is expected to be very similar, and if you wish to examine similarities across individuals, use the R-PCA method examining time point x time point correlations. You may uncover unseen processes putatively influencing the observed waveform.

If in doubt about how to proceed, try both and compare. Using R-PCA, you will obtain component loadings for each time point, and component scores for each subject. Using Q-PCA, individuals will have component loadings and the time points will have component scores.

A Few Assumptions of PCA, Briefly

1. PCA is a linear model; assumes the components sum together without interaction to produce the actual waveform
2. Sources of variance are orthogonal; if two sources are highly correlated, may result in a composite PCA component reflecting both
3. Component invariability in terms of latency jitter across subjects
   a. PCA does not distinguish between variations in amplitude vs variations in latency
   b. Especially a problem in comparing control vs pathological groups; pathological groups will typically be more variable
   c. Allen & Collins unpublished simulation study:
      (1) Two groups: Control & Pathological
      (2) Identical waveforms for each group differed only in latency
      (3) The two groups differed significantly on three of four principal component scores
      (4) In other words, if one indiscriminately interprets these as amplitude or morphology differences, one would be WRONG!!!

Summary

Always examine the results of PCA and compare to the waveforms in your original data set. PCA is a mathematical decomposition of variance, and it will extract factors accounting for many sources of variance, perhaps lumping several sources together in one component if these sources covary in the original data set (e.g., longer latency and diminished amplitude of late components may covary in a pathological group). It is incumbent upon you the investigator to determine what the sources of variance are. Examine variations in latency, amplitudes of components, and morphology of waveforms.