Understanding Principal Components and the MBH98 Results

It is common to use index numbers to reduce large data sets to one or two series, in order to make it easier to understand the numbers. For instance, a stock market index (like the Dow Jones) summarizes the change in value of the all the stocks in the New York market by averaging the price of some leading firms and ignoring all the others. This yields one index series, call it y, which "stands in" for a large group of k series $x_1,...,x_k$. So the Dow Jones Industrial Index stands in for the thousands of individual shares that are traded each day, to summarize whether the market grew in value or not.

One way to calculate y is to make it a weighted sum of the other k series:

$$y = a_1 x_1 + a_2 x_2 + \cdots + a_k x_k ,$$

where $a_1,...,a_k$ are the weights. For instance, a simple average uses $\frac{1}{k}$ for each of the weights.

In the case of the hockey stick study, indexes were constructed to stand in for larger groups of tree ring series. The weights were chosen using "Principal Component Analysis", a common algebraic tool. In the period under controversy, 70 tree ring series were replaced with a small number of indexes called "principal components".

Principal component analysis (PCA) is a method for choosing the weights so that y will "explain" as much of the variance in the group of variables $x_1,...,x_k$ as possible. It is relatively easy to explain the behaviour of a series that never changes. But if a series varies a lot, you can only summarize it with another series that tracks its movements to some extent. PCA works by examining the variance of each of the k series, and selecting higher weights for those series that vary a lot, so that they influence the weighted sum y relatively more. (The weights $a_1,...,a_k$ are restricted so that the sum of their squared values equals 1; this is necessary for computational reasons.) The graph of y will most strongly reflect the shape of whichever of the k series $x_1,...,x_k$ have the highest variance, and hence the largest weights.

Once the weights are chosen to maximize the explained variance, y is called the "first Principle Component" of $x_1,...,x_k$, or PC1. The standard interpretation is that the graph of y shows the dominant

pattern of variance in the *x*'s. The analysis also produces an estimate of how much of the variance in the *x*'s is explained by the PC1. If the explained variance associated with PC1 is very high, it implies that there is one dominant signal in the *k* underlying series. Hence the graph of *y* is an important summary of the information in the underlying data.

Since *y* doesn't usually explain everything going on in the underlying data, there is some left-over, unexplained variance associated with each of the *k* columns. PCA yields a set of *k* residual vectors

$$z_1,\ldots,z_k$$

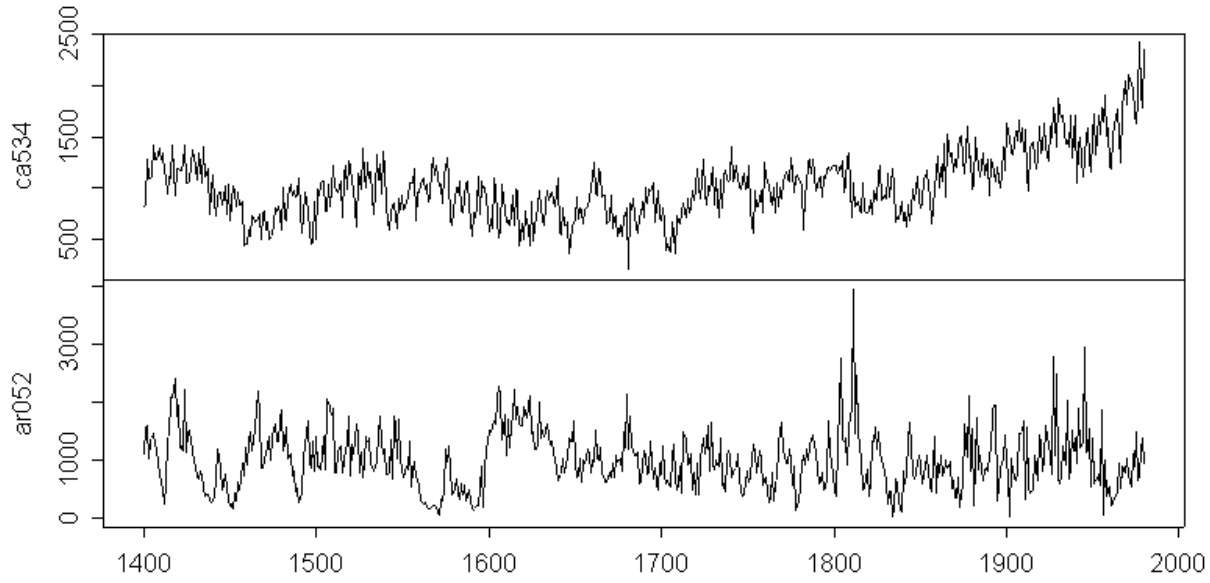
one for each of the *x*'s, which shows that part of the behaviour of the corresponding *x* series that is not explained by *y* or a scalar multiple of *y*. The PCA process can be repeated on the *z*'s, yielding another set of weights and another principal component. In this case it is the PC1 of the *z*'s, but more importantly it is the "second Principal Component" of the *x*'s, or PC2. PC2 is the best summary of the variability left over after the PC1 has explained the dominant variability. The analysis also yields an estimate of how much variance in the *x*'s is explained by PC2.

By repeating the process one can compute PC3, PC4, PC5 and so on, at each step obtaining a series that explains progressively less and less of the variance in the *x*'s. Eventually you would get a PC consisting of a column of near-zeroes, which means there is no variability left to explain. There is no formal rule for deciding how many PCs are needed to adequately explain all the variability in the *k* original *x*'s. If, say, the first PCs explain over 90 percent of the variance in the *x*'s, that may be sufficient for the purpose at hand. In PC analysis, regardless of how may PCs are calculated, successive PCs (PC2, PC3 etc) capture progressively less important patterns in the data.

PCA can have difficulty detecting important patterns in the data if the *x*'s are in very different units. If this is a problem the data may be re-scaled prior to doing PCA. The transformation called "standardization" involves, for each series, subtracting its mean and dividing by the square root of its variance. This re-scales all the series so that the standardized versions end up with the same mean (0) and variance (1). This makes PCA less likely to pick weights based on differing nominal units rather than the inherent variability in the underlying data. In the case of MBH98 the data are tree ring indexes that are in common dimensionless units, so apart from shifting to a zero mean it is not necessary to standardize them for the purpose of running PC analysis.

But in MBH98 an unconventional transformation was applied to the *x*'s. Each series in the period under controversy runs from 1400 to 1980. Rather than compute the mean over this interval they computed the mean over 1901 to 1980 (and likewise for the variance). For most of the 70 series in the North American tree ring group ("NOAMER") this does not matter too much, since they have no trend and the variance doesn't change much over the interval. So a PC analysis applying the usual standardization would yield pretty much the same results as the MBH98 method.

But a few tree ring series in the MBH98 data base series undergo a dramatic change in behaviour at the start of the 20th century. Back in the 1980s, two tree ring researchers (Donald Graybill and Sherwood Idso) had sampled a network of bristlecone and foxtail pine trees at high elevation sites in the western USA. They found the trees had experienced a growth spurt in the 20th century, but nearby weather station records showed no corresponding temperature increase over the interval. There are various theories about why these trees grew the way they did, but the important thing to note is that the growth spurt is *not* a temperature signal. However, Mann et al. included the Graybill-Idso tree ring records in their data base. As an example, the following graph shows a "normal" tree ring proxy (bottom panel, Mayberry Slough AR), and one of the Graybill-Idso proxies (top panel, Sheep Mountain CA).



You can see that the top panel tells a very different story from the bottom one.

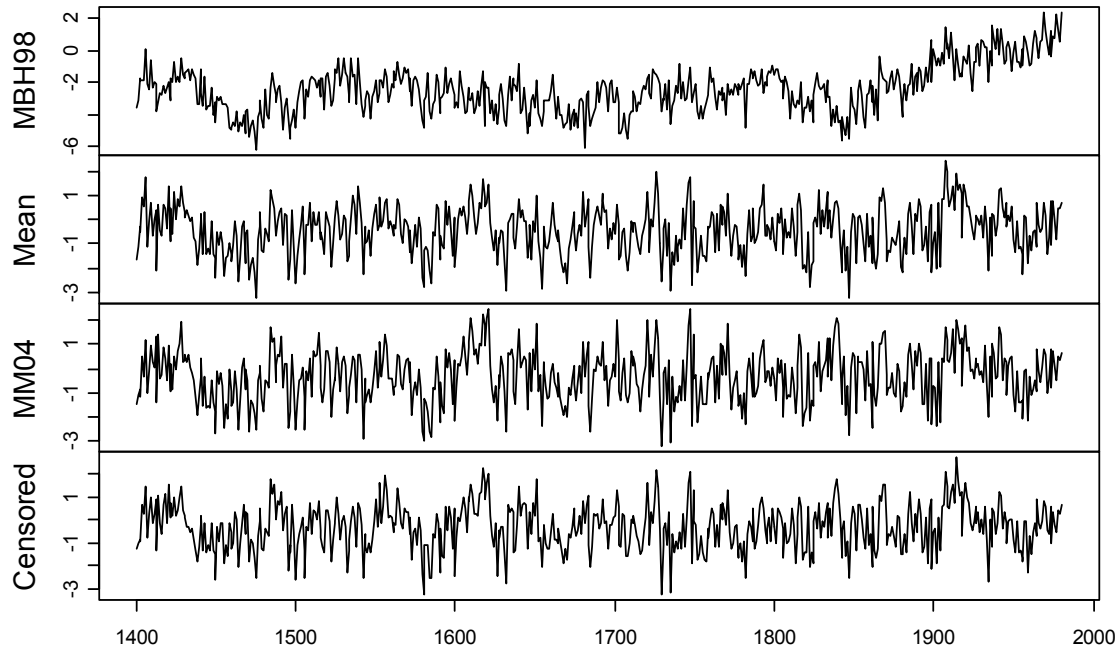
The MBH98 transformation had the effect of radically inflating the variance of series like the one in the top panel, relative to ones in the bottom panel. The reason is that, in the top panel, the mean of the ending

segment is different from the mean of the whole series. So rather than center the series on a zero mean, their method de-centers it so that most of the series has a non-zero mean. The variance grows with the square of the deviations away from a nonzero mean. Then PCA gives these high-variance, decentered series the heaviest weights. In the above graph, Mann's PC method gives 390 times more weight to the top series than the bottom series. As a result, a few hockey stick shaped-series dominate the PC1 and from there dominate the final shape of the hockey stick graph.

A reader (Professor Richard Muller of Berkeley) sent us the following numerical example to help explain.

> *Suppose we have a flat distribution of temperature for the years 1000 to 1980 between limits +-1. Then the variance, the mean value of x^2, is 1/3. Let's assume that the variance during the period 1902-1980 is also 1/3. Then the normalized variance becomes 1. That's the way it should be.*
>
> *Now let's consider a set of data with a hockey-stick shape. Let it have a flat distribution between -1 and +1 for all years prior to 1902, followed by a steady rise from 1902 to 1980 from +1 to +2, a slight hockey stick end. The mean value for this period is M0 = +1.5, and the variance for this period about that mean is V0 = 1/12.*
>
> *Now we return to the bulk of the data, covering the period 1000 to 1902. Following the procedure of Mann et al., first we subtract M0. The data is now a flat distribution between -0.5 and -2.5. The (incorrect) variance V (about the false average) is calculated as the mean value x^2 between -0,5 and -2.5 giving a value V = 31/12. We divide this variance by V0 to get a normalized variance of (31/12)/(1/12) = 31, So this hockey stick data, random data with a rising tail at the end, is made 31 times as important as the other data! When PCA is used, it will pick a hockey-stick shape as the function that best explains this huge variance.*

The following graph shows the NOAMER tree ring data summarized 4 different ways. The second panel is just the mean of the (70) series. The top panel ("MBH98") is the PC1 computed using the MBH98 method, while the third panel ("MM04") is the PC1 using a standard method (on the covariance matrix). Applying a standard PC methodology to the data would lead to the conclusion that the "dominant" pattern in the data suggests there was no change in the climate over the 600 years covered by these series. Applying the MBH98 method implies a dominant warming pattern starting around 1900.



If PCA is used to compute PC2, PC3, etc for the NOAMER network, standard analysis shows that the Graybill-Idso proxies don't get much weight until PC4, which accounts for less than 8% of the explained variance of the NOAMER group. This makes sense, as they are not a dominant signal in the data and they only represent a minor local pattern in the data. Under the MBH98 method they get very heavily weighted in the PC1, and they are said to account for nearly 40% of the variance in the NOAMER group.

If they are removed from the NOAMER group altogether, then we are left with about 50 series that look like the Mayberry Slough example (i.e. no change in the mean over the 6 centuries). For this subset it doesn't matter if you use the MBH98 method or a standard method, the PC1 looks the same. The bottom panel ("CENSORED") shows the PC1 computed using the MBH98 method but with the Graybill-Idso series (and 1 other similar one) removed. As you can see it is virtually identical to the conventional PC1 in the 3rd panel.

The reason the bottom panel is labeled "Censored" is that these data are on Professor Mann's FTP site in a folder called

ftp://holocene.evsc.virginia.edu/pub/MBH98/TREE/ITRDB/NOAMER/BACKTO_1400-CENSORED.