Summarising stratigraphic data using principal curves

Gavin L. Simpson

May 30th, 2013

Abstract

Keywords

Introduction

Stratigraphic data are generally high dimensional, often exceedingly so in the case of speciose proxies such as diatoms. It is difficult, if not impossible, to identify the major changes in such data directly, and some form of data reduction is required to highlight these changes. Ordination methods, most notably principal components analysis (PCA) and correspondence analysis (CA) and its detrended version (DCA), are often used by palaeolimnologists to identify the main pattern or patterns in a stratigraphic sequence. The locations of samples on the first one or two ordination axes are usually presented alongside the main taxa or variables in the proxy data and are essentially used as a summary of change in the set of proxies analysed.

Continual compositional change along ecological gradients

Figure 1 shows a small artificial data set of 3 species observed at 19 sites along a gradient, taken from Table 9.7 of P. Legendre and Legendre (2012, 482). From the upper panel in Figure 1 it is clear that species composition changes progressively along the gradient represented by the sampling locations. The lower panel of Figure 1, a PCA applied to the data, exhibits a strong curvilinear pattern in the ordination space of components 1 and 2 (note the curvature extends into the thrid dimension). Such curvilinear patterns are known as a *horseshoe* (when the end points of the curve bend back in on the curve itself) or an *arch* (when the end points of the curve do not bend back on themselves) and have long been

noted in the ecological literature (e.g. Goodall 1954; Noy-Meir and Austin 1970; Swan 1970). In this case the end points bend very strongly inwards due to the PCA considering the absence of species 2 a similarity between these samples.

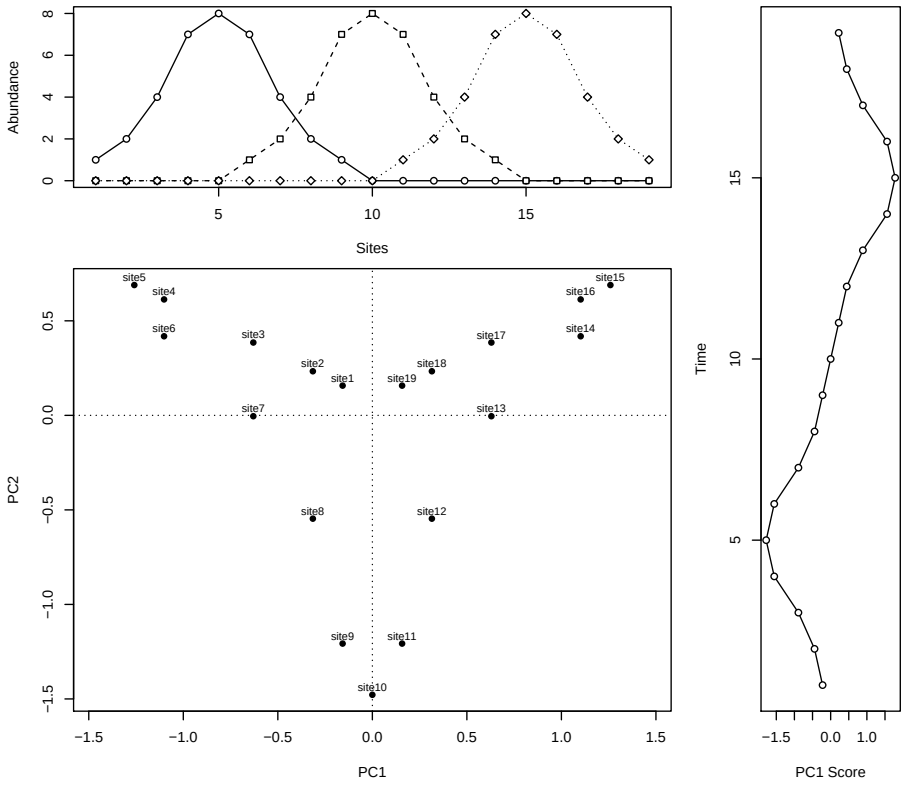

Figure 1:

For the sake of argument, assume the gradient shown in Figure 1 is temporal and the samples form a stratigraphic sequence. If one were to take the first principal component axis scores from such an ordination as a summary of change in the data the pattern shown in panel C of Figure 1 would be observed, suggesting oscillatory behaviour. Such behaviour is totally contrary to the true response along the gradient that we know to be present because these are articificial data. In this case, at least two PCs would be required to summarise the main features in this data set, where in practice there is only a single gradient, which, given a more suitable technique, could be recovered as a single component. It is in this regard that I introduce principal curves as one way to extract such a gradient from palaeoecological data.

Such horseshoe or arched configurations often occur in real palaeocological data. Consider the classic, well-studied late-glacial pollen sequence from Abernethy Forest, Scotland (Birks and Mathewes 1978), the main pollen taxa from which

are shown in Figure 2.

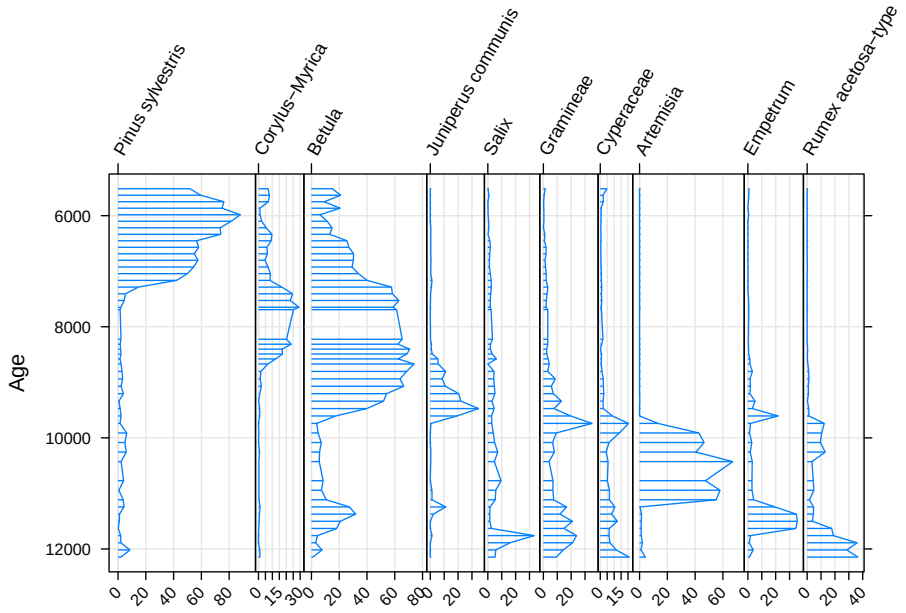

Figure 2:

When PCA is applied to Abernethy data a pronounced arched configuration is observed with two strong directions of change that align with the first and second principal components (Figure 3a). The first principal component largely acocunts for change after 9000

Principal Curves

A principal curve (PCr) is a smooth, one-dimensional curve that passes through the middle of a set of points (observations) m dimensions, where m refers to the number of variables (species for example). A PCr can be thought of as a generalisation of the first principal component line but, instead of being linear, is a smoothly varying curve. PCrs are also related to various forms of smooth regression used as scatterplot smoothers for example. To demonstrate the associations between PCrs, linear regression, scatterplot smoothing and PCA, consider a simple 2-d sample of data with observations on two variables x and y, with 50 observations of each. x takes values in the range $0, ..., 1$, and y is generated from the cubic equation

$$y = -0.9x + 2x^2 + -1.4x^3 + \varepsilon$$

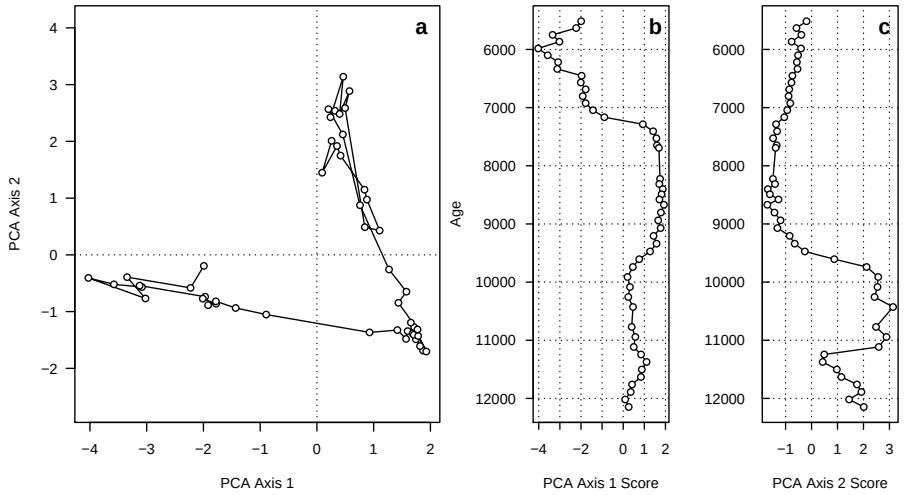

Figure 3:

where ε is Gaussian noise with $\mu = 0$ and $\sigma = 0.05$.

In the least squares linear regression relating x and y with y in the response role the line of best fit through these data is the one that minimises the sum of squared errors in y: all the error is assumed to be in y with x known wihtout error. Panel a in Figure 4 illustrates the linear regression fit to these data, with the narrow vertical bars illustrating the distances between the observations and the fitted line whose squares are minimised. In PCA, neither variable plays the role of response or predictor and errors are now minimised in both x and y. This situation is shown in panel b of Figure 4. Note that the distances between the observations and the fitted line (principal component) are measured orthogonal to the line in constrast to the regression situation and mean that the squared error in both x and y is evaluated and minimised.

The models discussed so far are both straight lines through the data. Depending on the relationship between x and y a more complex non-linear relationship may be required. Smoothing splines are a semi-parametric generalisation of the linear least squares model, where a smooth curve is fitted to the data that minimises the sum of squared distances in y subject to some constraint on the complexity of the fitted curve. The relationship between x and y is derived from the data themselves rather than specified *a priori* by the analyst, but, as with linear regression, the curve is fitted by consideration of the error in y only. The smoothing spline fit to the example data, using ~5.49 degrees of freedom (as selected via generalised cross-validation; GCV) is shown in panel c of Figure 4. Principal curves combine the features of orthogonal errors from PCA with the non-linear, data analytical fit from smoothing splines. Panel d shows a principal curve fitted to the example data.

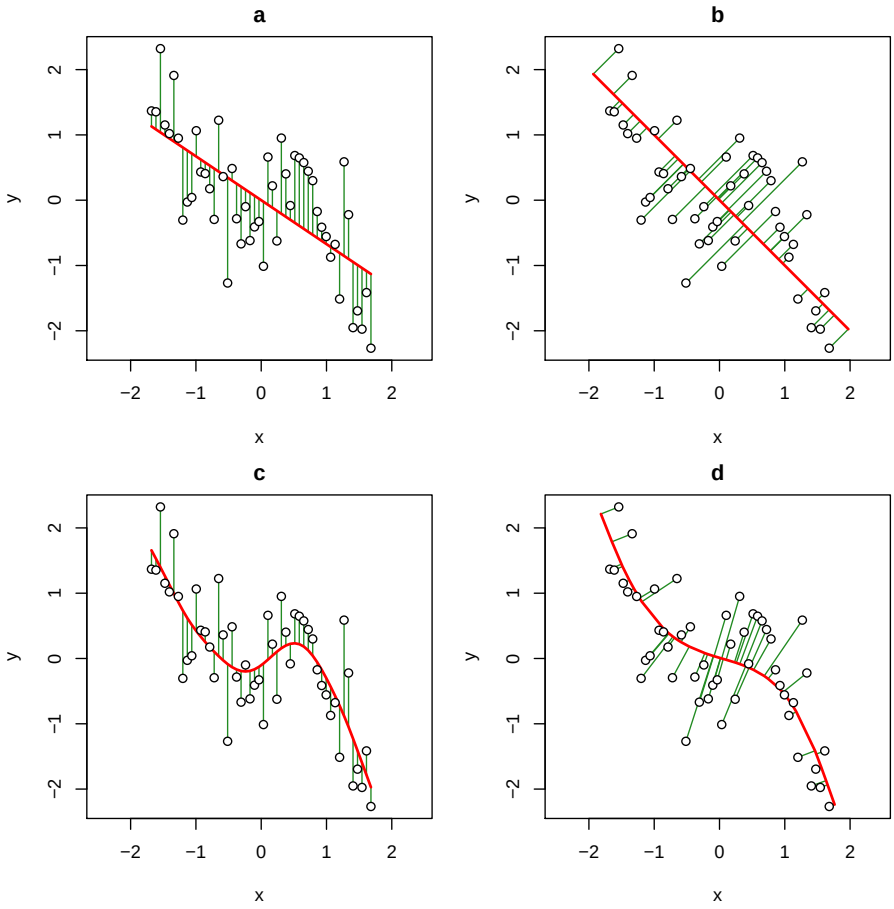

Figure 4: Relationship between (a) linear regression, (b) principal components analysis, (c) smoothing splines, and (d) principal curves each applied a simple 2-dimensional data set.

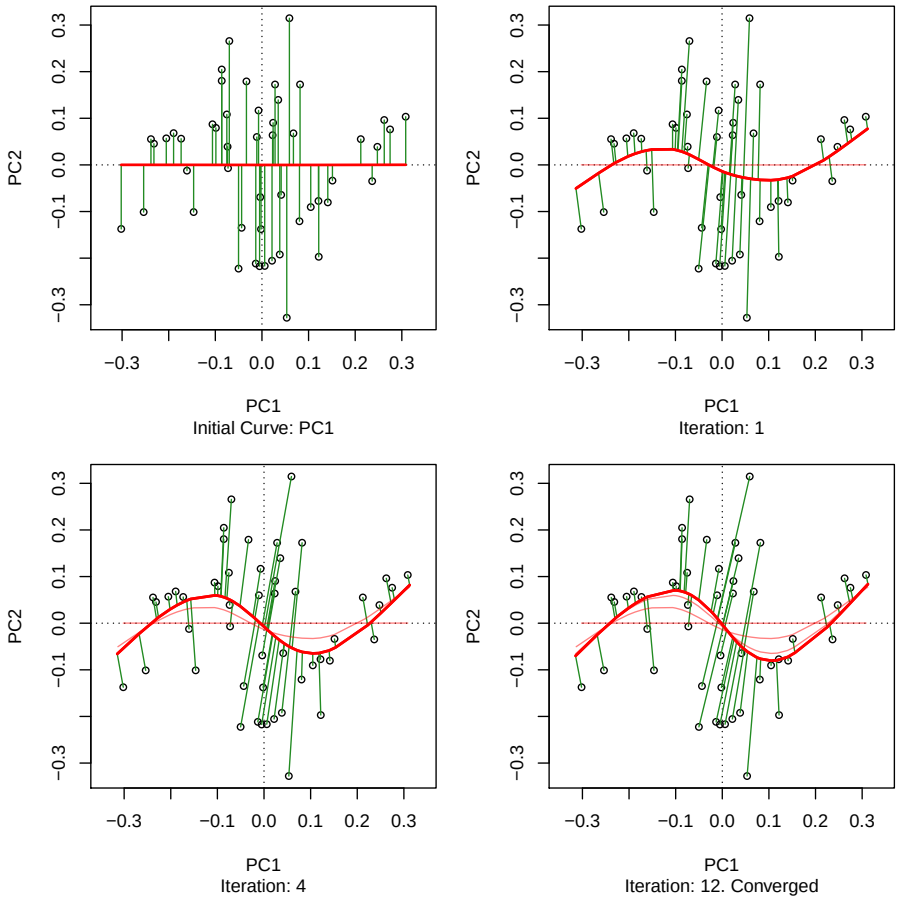
PC2
PC1
Initial Curve: PC1
PC2
PC1
Iteration: 1
PC2
PC1
Iteration: 4
PC2
PC1
Iteration: 12. Converged

Figure 5:

Principal curves are fitted using a two-stage iterative algorithm, which starts from any smooth curve through the data. Usually a PCA or CA axis is used as the starting curve. The first stage of the algorithm is the projection step; the observations are projected onto the current curve. Each observation projects onto a the location on the curve to which it is closest (shortest Euclidean distance). The distance along the curve from an arbitrarily chosen end is recorded for each projected observation. These distances are the arc lengths along the curve and can be thought of as the axis score for the principal curve.

A local-averaging step forms the second stage of the algorithm. The current curve is bent towards the data through the fitting of smooth functions to the variables in the data. A smooth function is fitted to each variable in turn using the values of the variable (for example the species abundance) as the response and the distance along the curve as the predictor. The fitted values of this smooth function for the ith variable represent the coordinates of the new curve in the ith dimension. The set of fitted values from the smooth functions for all m variables gives the location of the updated curve.

At this stage a self-consistency check is made through the prejection step; the observations are projected on to the closest point on the updated curve and the distance along the curve determined. If the updated curve is sufficiently close to the previous curve the updated curve is said to be self-consistent and convergence is declared and the algorithm terminates. If the updated curve is not self-consistent the algorithm continues to iterate the two stages until the curve become self-consistent or the maximum nuber of allowed iterations is reached.

Figure 5 illustrates the iterative principal curve algorithm for the cubic polynomial data in Figure 4. In the upper left panel the initial curve is shown, here the first principal components was used. At the first iteration of the algorithm the curve has been bent markedly towards the data (Figure 5, upper right panel) such that the squared orthogonal distance between the observations and the curve is reduced. By the fourth iteration (Figure 5, lower left panel) the curve has largely stabilised and the improvement of fit in subsequent steps is minor. The algorithm converged after 12 iterations to the curve shown in the lower right panel of Figure 5, upper right panel.

Whilst the principal curve algorithm is decidely simple it allows for a great deal of customisation by the user. The intial curve needs to be supplied and in the initial description of principal curves, the first or second principal component line was a natural choice. For ecological data the first or second correspondence analysis axis may provide a better starting point. Alternatively, an principal coordinate analysis axis may be used, allowing for any dissimilarity coefficient to be chosen by the user. The choice of starting curve is particularly important; choosing an initial curve that is far from the optimal princpal curve may lead to increased time to convergence or, more importantly, an over-fitted principcal curve. Several starting configurations should be used to achieve an optimal fit, and the resulting curves investigated to identify over fitting. We return to this shortly.

In the description of the principal curve algorithm I referred simply to a smooth function being fitted to each variable in turn. This was intentional as this aspect of the principal curve method can be best thought of as a plug-in element; there are are large number of ways in which a smooth function relating proxy values to distance along the curve, such as LOESS, kernel smoothing, a generalised additive models (GAM) or the many forms of smoothing or regression splines. In the examples presented here, cubic smoothing splines are used. GAMs may be particularly useful in fitting data whose error distribution is non-normal.

Having selected a method for fitting the smooth functions, the next choice involves identifying an appropriate number of degrees of freedom or complexity for each function. In the case of GAM fits, penalised regression splines can be used where the complexity is determined from the data themselves either via GCV or, more robustly, via REML smoothness selection during fitting. In the case of LOESS or kernel smoothing, the window span or bandwidth can be varied, and with cubic smoothing splines, the degree of freedom for each function can be specified.

Hastie and Stuetzle (1989) suggest starting with a large span of 0.6 (60% of the data in the smoothing window) or a low number of degrees of freedom and then iterate the curve to convergence. The resulting curve is then used as the intial curve for a second principal curve fit, this time employing smoothers with a span or 0.5 and again iterating until convergence. This curve is then used for a final time, with a span of 0.4. Addtionally, this last curve can be used in one subsequent *iteration* of the algorithm, this time with the span for each smooth function selected via cross-validation (Hastie and Stuetzle 1989),.This iterative process is designed to stop the principal curve from moving too close to the individual data points during early iterations of the algorithm. The initial fit using a large span allows a principal curve to conform to the gross features of the data without over-fitting. The subsequent runs with gradually smaller spans essentially allow for this general curve to be slowly improved in terms of fit to data.

De'ath (1999) remarks that gradually increasing the complexity of the fitted curve is not needed; De'ath (1999) suggests an alternative straegy to fitting principal curves

1. Initialise the curve,
2. Fit smooth functions to each variable using cross-validation to select the optimal complexity (span, bandwidth or degrees of freedom),
3. Take as the complexity value to use to fit the principal curve the median of the complexity values over the set of m variables,
4. Iterate the curve to convergence,

5. Optionally, perform a final iteration of the algorithm using cross-validation to select the complexity for each smooth function fitted to distance along the converged principal curve.

For the summary of sediment core data, I have observed that the final cross-validation step tends to lead to overly-complex fitted curves. In addition, I have found De'ath's suggestion (1999) to use the median of the m complexity values to be somewhat wasteful of degrees of freedom; there will be many variables that do not warrant the fitting of such complex curves (variables that respond in a linear or monotonic fashion), whilst there will be some variables that require more complex fits than the media complexity.

I favour a more data-driven approach wherein at step 2 above, instead of using the median complexity, the cross-validated complexity for each variable is used to fit the curve. As such, species that vary markedly along the initial curve are fitted using more complex smooth functions than those species that do not vary as much. One obvious issue with this approach is that the complexity for each smooth function is determined with respect to the initial curve. A hybrid approach may be more robust to specification of the initial curve; first a principal curve is fitted with low complexity smooth functions and the resulting curve used as the initial curve to a second run of the algorithm, this time using the cross-validated complexity for each variable.

Once a principal curve has been fitted it is essential that the resulting curve is investigated as over-fitting is easy to achieve. The algorithm can usually be monitored during fitting, to see how the curve adapts to the data. This coupled with the trace output for the residual variance at each iteration can help to assess the complexity of the fitted curve. Subsequently, the response curves for the variables along the fitted curve should also be checked. Overly-fitted principal curves can usually be diagnosed through complex, multimodal response curves for some or all variables. Such response curves can be observed even if the principal curve itself does not appear to be complex when plotted.

Methods

Results

Discussion

References

Birks, Hilary H., and Rolf W. Mathewes. 1978. "Studies in the vegetational history of Scotland. V. Late Devensian and early Flandrian

pollen and macrofossil stratigraphy at Abernethy Forest, Inverness-shire." *New Phytologist* 80: 455–484. doi:10.1111/j.1469-8137.1978.tb01579.x. http://onlinelibrary.wiley.com/doi/10.1111/j.1469-8137.1978.tb01579.x/abstract.

De'ath, Glenn. 1999. "Principal Curves: a new technique for indirect and direct gradient analysis." *Ecology* 80: 2237–2253. http://www.esajournals.org/doi/abs/10.1890/0012-9658(1999)080%5B2237:PCANTF%5D2.0.CO%3B2.

Goodall, D. W. 1954. "Objective methods for the classification of vegetation. III. An essay in the use of factor analysis." *Australian Journal of Botany* 2 (jan): 304–324. http://www.publish.csiro.au/paper/BT9540304.

Hastie, Trevor, and Werner Stuetzle. 1989. "Principal Curves." *Journal of the American Statistical Association* 84 (jun): 502–516. doi:10.2307/2289936. http://www.jstor.org/stable/2289936.

Legendre, Pierre, and Louis Legendre. 2012. *Numerical Ecology*. Elsevier.

Noy-Meir, I., and M. P. Austin. 1970. "Principal Component Ordination and Simulated Vegetational Data." *Ecology* 51 (may): 551–552. doi:10.2307/1935398. http://www.jstor.org/stable/1935398.

Swan, J. M. A. 1970. "An Examination of Some Ordination Problems By Use of Simulated Vegetational Data." *Ecology* 51 (jan): 89–102. doi:10.2307/1933602. http://www.jstor.org/stable/1933602.