Stratigraphic Data

Gavin L. Simpson

U Adelaide 2017 • Feb 13–17 2017

Introduction

Introduction

Stratigraphic data are usually a multivariate time series

Common things we want to do to these data are

1. summarise changes in the data
2. estimate rates of change
3. zone or cluster the data
4. fit models to the time series

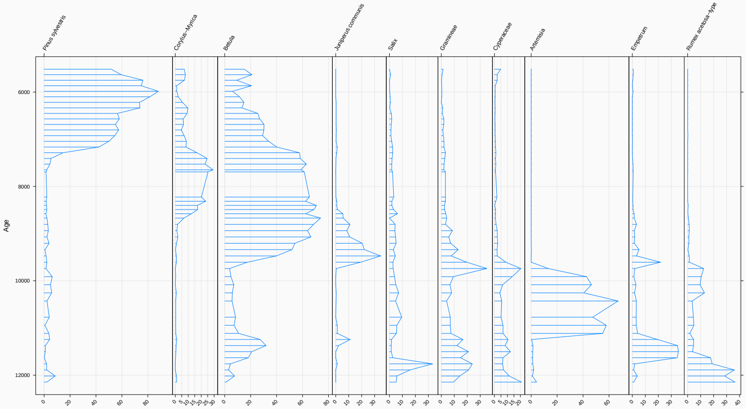
Summarising stratigraphic data

Summarising change in stratigraphic data

- Ordination commonly used to describe patterns of change in multivariate sediment core data
- PCA, CA, or even DCA axis 1 and 2 scores commonly used
- These methods capture largest patterns of variation in underlying data under assumption of particular model
- Can be upset by data sets with a dominant gradient
- Can apply all techniques learned earlier in workshop to stratigraphic data
- Can we do any better than these methods?

Principal Curves

A single long or dominant gradient in an (palaeo)ecological data set poses particular problems for PCA and CA — horseshoe or arch

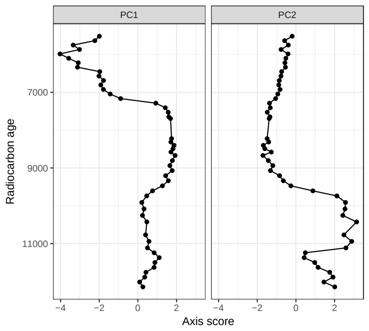


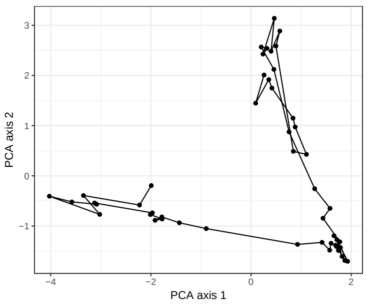
Abernethy Forest pollen data (Birks & Mathewes, 1978)

Principal Curves

A single long or dominant gradient in an (palaeo)ecological data set poses particular problems for PCA and CA — horseshoe or arch

Trend is broken over two or more axes:



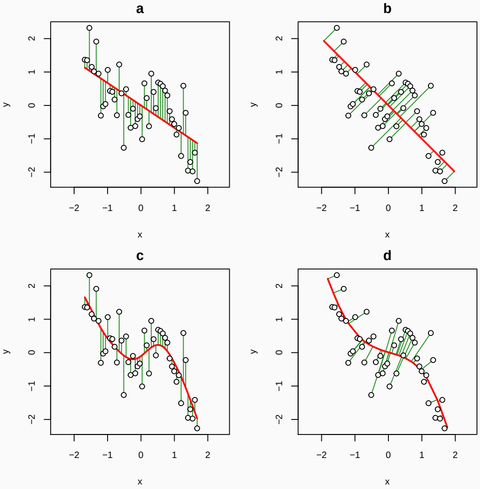


Can we generalise the PCA solution to be a smooth, non-linear surface?

Principal Curves — Comparison of estimation techniques

- In OLS regression, y is the response and x is assumed without error
- Errors are minimised in y only — sums of squared errors
- PCA can be seen as a regression of y on x where neither y nor x plays the role of response or predictor
- In PCA, errors in both x and y are minimised — sums of squared orthogonal errors

Principal Curves — Comparison of estimation techniques



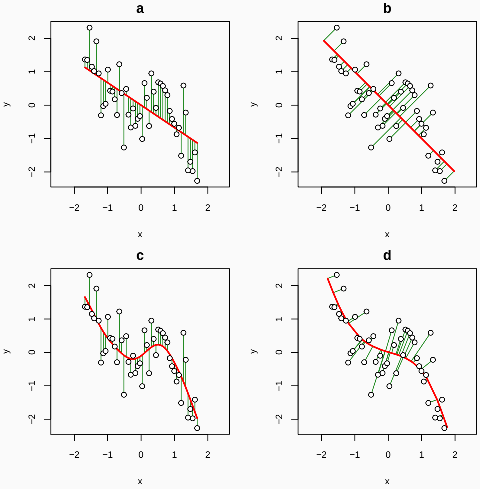
Principal Curves — Comparison of estimation techniques

We can generalise the OLS model to a regression of y on x using a smooth function of x, $f(x)$, as the predictor $f(x)$ can be estimated using a multitude of techniques

- Loess smoother
- Local-linear smooths
- Smoothing splines
- Regression splines
- ...

$f(x)$ is usually estimated from the data, with smoothness determined by minimising a penalised sums of squares criterion under CV (or GCV): **Errors are still assessed in y only**

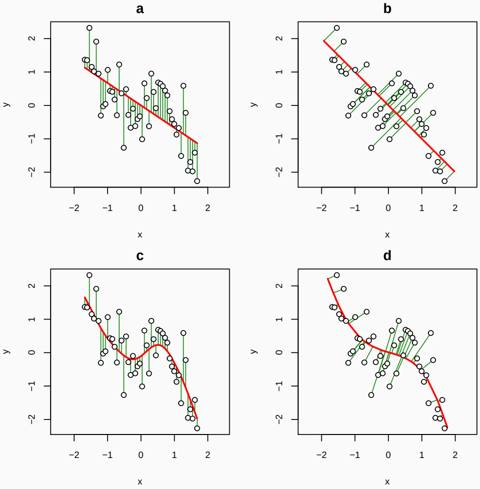
Principal Curves — Comparison of estimation techniques



Principal Curves — Comparison of estimation techniques

- Ideally we would generalise PCA to find non-linear manifolds in the same way as we went from OLS to semi-parametric regression using smoothers
- This is exactly what is done in the method of principal curves
- Our aim is to estimate as the principal curve, a 1-d manifold that passes through the data in high-dimensions that minimises the sum of squared orthogonal errors
- We bend the principal component (for example) towards the data to achieve a better fit to the data
- How far and how flexibly we can bend the curve towards the data is determined from the data to minimise a penalized criterion during fitting

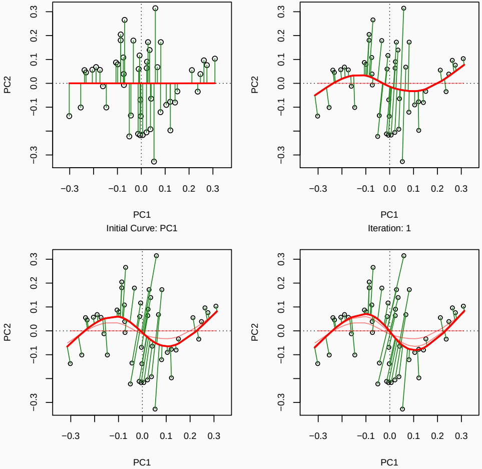
Principal Curves — Comparison of estimation techniques



Principal Curves — Fitting algorithm

- Start with any smooth curve — the first or second PCA or CA axis
- Begin the Projection Step
    - All objects are projected onto a point on the smooth curve that they are closest too
    - The distances of the points along the curve that each object projects onto are determined
- Begin the Local Averaging Step
    - Bend the current smooth curve towards the data so that the sum of squared orthogonal distances is reduced
    - Taking each species (variable) in turn as the response, fit a smoother to predict the response using distance along the current curve as the predictor variable
    - Repeat for all species (variables) and collect the fitted values of the individual smoothers into a matrix that described the new location of the curve in high dimensions
- If the new curve is sufficiently similar to the current curve, declare convergence
- If algorithm has not converged, iterate the projection and local averaging steps until convergence

Principal Curves — Fitting algorithm



Initial Curve: PC1

Iteration: 1

Iteration: 4

Iteration: 12. Converged

Principal Curves — How Smooth?

- The smoother fitted to produce the principal curve is a plug-in element of the algorithm
- Can use any smoother; here used cubic regression splines
- Important to not over fit the data by allowing too-complex a curve
- Several options
  - Fit PCs with a large span (few df), note error, then reduce span (increase df), note error, etc. Use screeplot to determine optimal span
  - Fit smoothers to each species using starting curve, allowing (G)CV to choose optimal smoothness for each species. Fit the PC using the median of the smoothness values over all species
  - Allow the optimal degree of smoothness to be determined for each species individually during each local averaging step
- Advantage of latter is that those species that vary along the curve more strongly can use more degrees of freedom than those species that only vary lineally

Principal Curves — Abernethy Forest

```
> aber.pc <- prcurve(abernethy2, trace = FALSE, vary = TRUE, penalty = 1.4)
> aber.pc
```

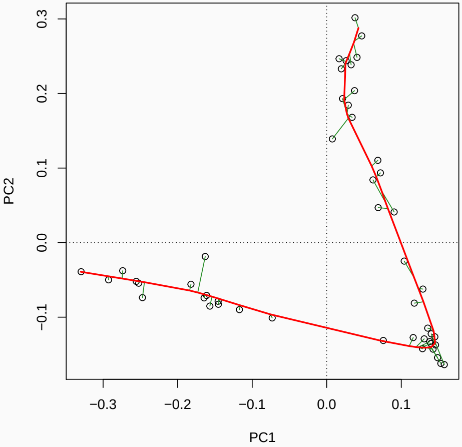
```

    Principal Curve Fitting

Call: prcurve(X = abernethy2, vary = TRUE, trace = FALSE, penalty
= 1.4)

Algorithm converged after 6 iterations

           SumSq Proportion
Total     103234      1.000
Explained  98864      0.958
Residual    4370      0.042

Fitted curve uses 218.3391 degrees of freedom.
```

```
> varExpl(aber.pc)
```

```
      PrC
0.9576693
```

Principal Curves — Abernethy Forest

Visualise the fitted curve in PCA space



Principal Curves — Comparison with PCA and CA

- The PC describes the long, sequential gradient in vegetation in a single variable
- The PC explains 96% of the variance in the absolute pollen data
- PCA axis 1 explains 47% and CA axis 1 31% of the variance in the data
- We need at least 2 PCA axes to fully capture the single gradient (80.2%)
- Distance along the curve between adjacent time points is a measure of compositional change
- Can be expressed as a rate of compositional change — highlights the periods of rapid compositional change in the Abernethy sequence

Rate of change analysis

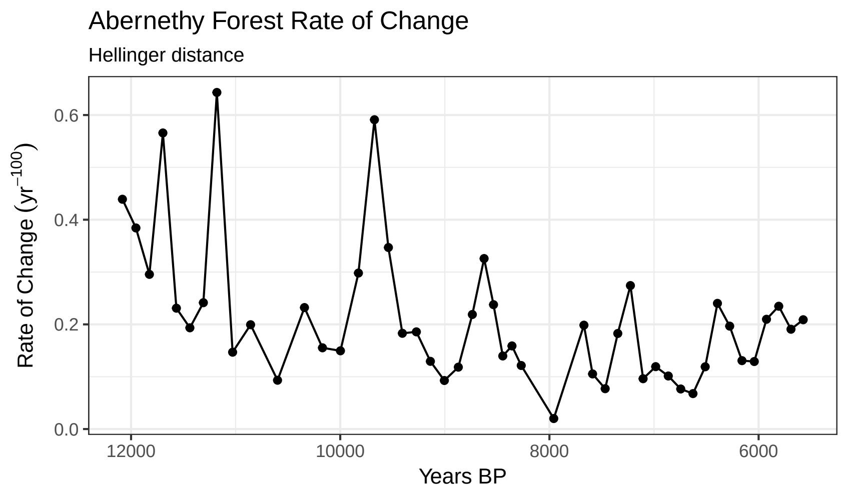
Rate of change analysis

- Stratigraphic sequences record changes over time
- How quickly do these changes take place?
- Rate of change analysis aims to answer this
- Two general approaches:
    - change in ordination units
    - change measured in dissimilarity
- Could also use
    - derivatives of splines from principal curve
    - derivatives of GAM(s) fitted to variables of interest

Rate of change analysis

- Jacobsen & Grimm (1988) method involves
  - smooth the data
  - interpolate to constant time intervals
  - ordinate smoothed, interpolate data (e.g.~using DCA)
  - calculate change in ordination/axis score units as measure of RoC
- Dissimilarity-based approach can be performed two ways
  - Smooth the data & interpolate, *then* compute dissimilarity between interpolated levels
  - Compute dissimilarity between adjacent samples directly, then standardise dissimilarity by time interval between samples.

Rate of change analyis



Chronological clustering (zonation)

Chronological clustering

- Chronological (or constrained) clustering commonly used to partition a sediment sequence into 2 or more zones
- Useful for, *inter alia*
  - delineating periods of similar species composition
  - identifying discontinuities or periods of rapid change
  - to facilitate description of a stratigraphic sequence
- As with standard cluster analysis, plethora of methods available
  - Optimal partitioning
  - Binary (divisive) splitting
  - Agglomerative partitioning
- Can be used with any dissimilarity (in theory), but common ones are
  - Cluster sums of squares (within-group Euclidean distance)
  - Cluster-wise information statistic

Chronological clustering

- Optimal partitioning
  - Identifies optimal locations for splits to form *k* zones
  - Non-hierarchical, 3 zone solution **not** found by splitting one of the zones from the two zone solution
  - Split placed to minimise within-cluster sum of squares or information conten
- Binary (divisive) splitting
  - Similar to optimal method but *is* hierarchical
  - Split sequence into two zones, then split one of the 2 resulting zones, repeat
  - Zone that is split is the one that would reduce within-group SS or IC the most
- Agglomerative partitioning
  - Start with all samples in separate zones and fuse the most similar adjacent samples
  - Repeat, each time fusing most similar samples or zones

Chronological clustering

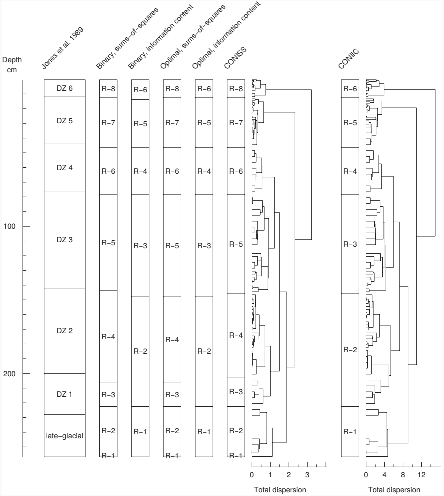


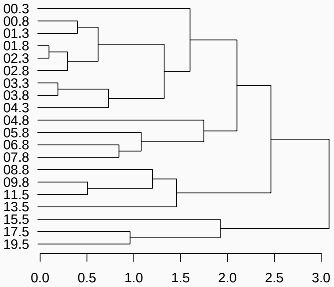
Figure 1:

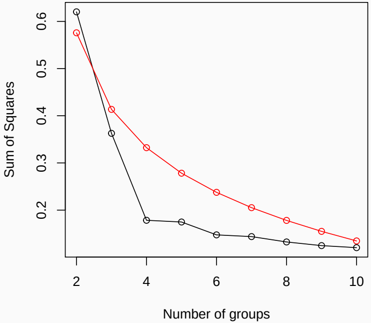
CONISS — Number of Zones?

CONISS applied to Round Loch of Glenhead short core

Number of zones determined by

- screeplot
- broken stick (Bennett, 1996, *New Phytologist* **132**, 155–170)



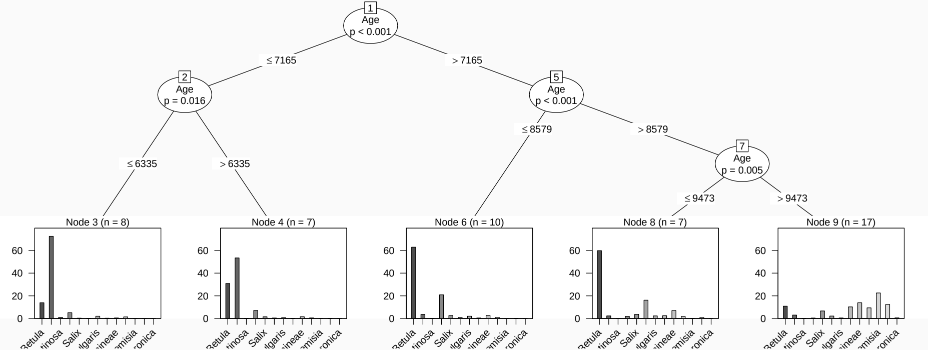


Binary splitting via MRT

Binary splitting or recursive partitioning applied to the Abernethy Forest pollen record

Fitted via a multivariate regression tree with **Age** as only predictor & number of zones determined by

- Cross validation and cost-complexity pruning (see Simpson & Birks, 2011, Chapter 8, DPER Vol. 5)
- Conditional inference (R package **partykit** and `ctree()`)



Re-use