Data visualization (and beyond) with local principal curves and manifolds

Jochen Einbeck

Department of Mathematical Sciences, Durham University

jochen.einbeck@durham.ac.uk

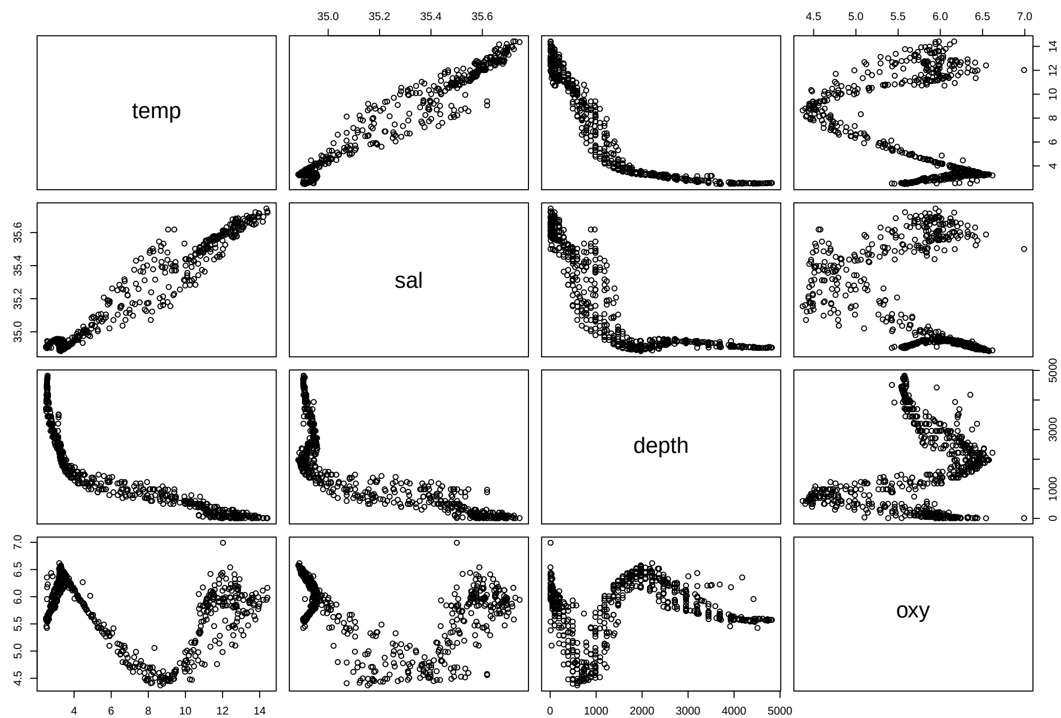
joint work with Ludger Evers (University of Glasgow),

London, 15 December 2013

Durham University

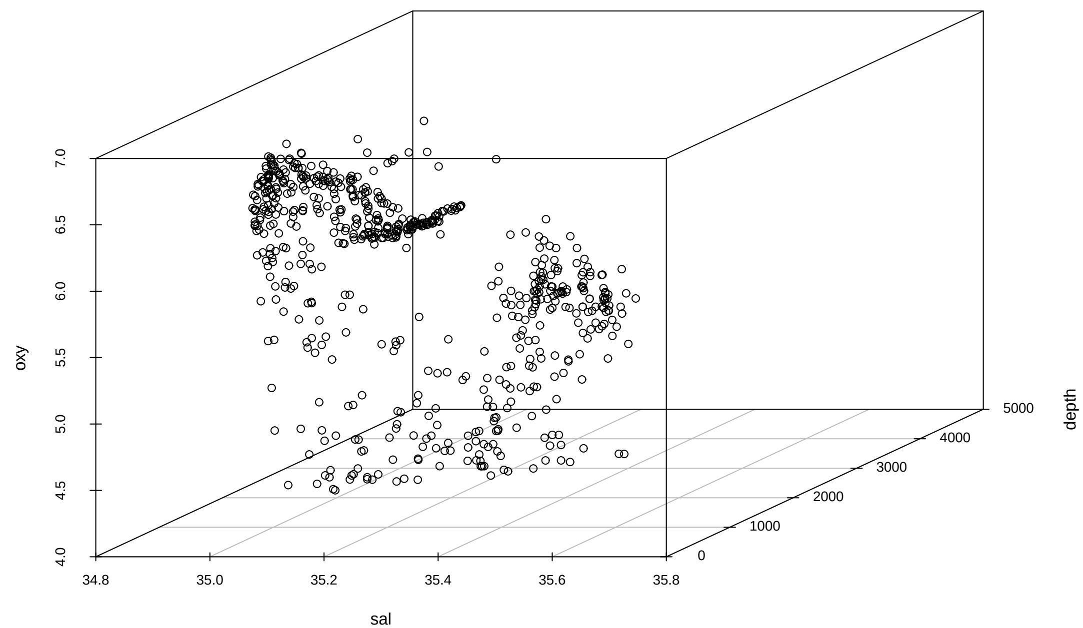
Motivation

- Consider oceanographic data recorded by the German vessel "Gauss" in May 2000 southwest of Ireland.
- $N = 643$ Measurements on water temperature (response), salinity, water depth, oxygen content.



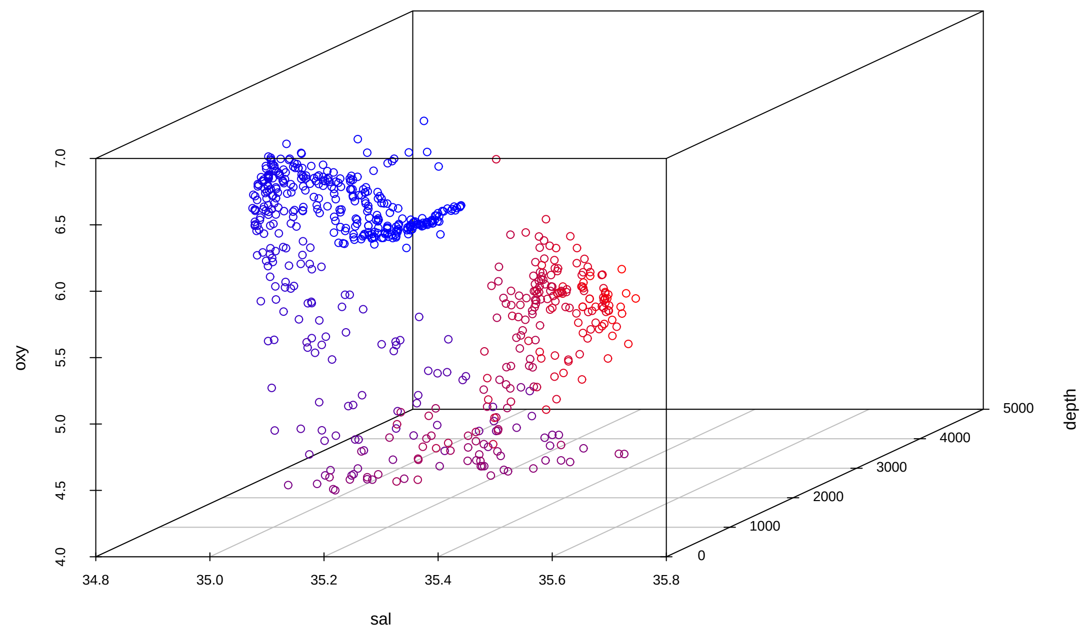
Motivation (cont.)

- This is a 3-variate regression problem, with the predictor space given by salinity, water depth, and oxygen:



Motivation (cont.)

- This is a 3-variate regression problem, with the predictor space given by salinity, water depth, and oxygen:



- We shade higher water temperatures red.
- Can we make use of the one-(?) dimensional inner structure?
- This is a task for principal curves.

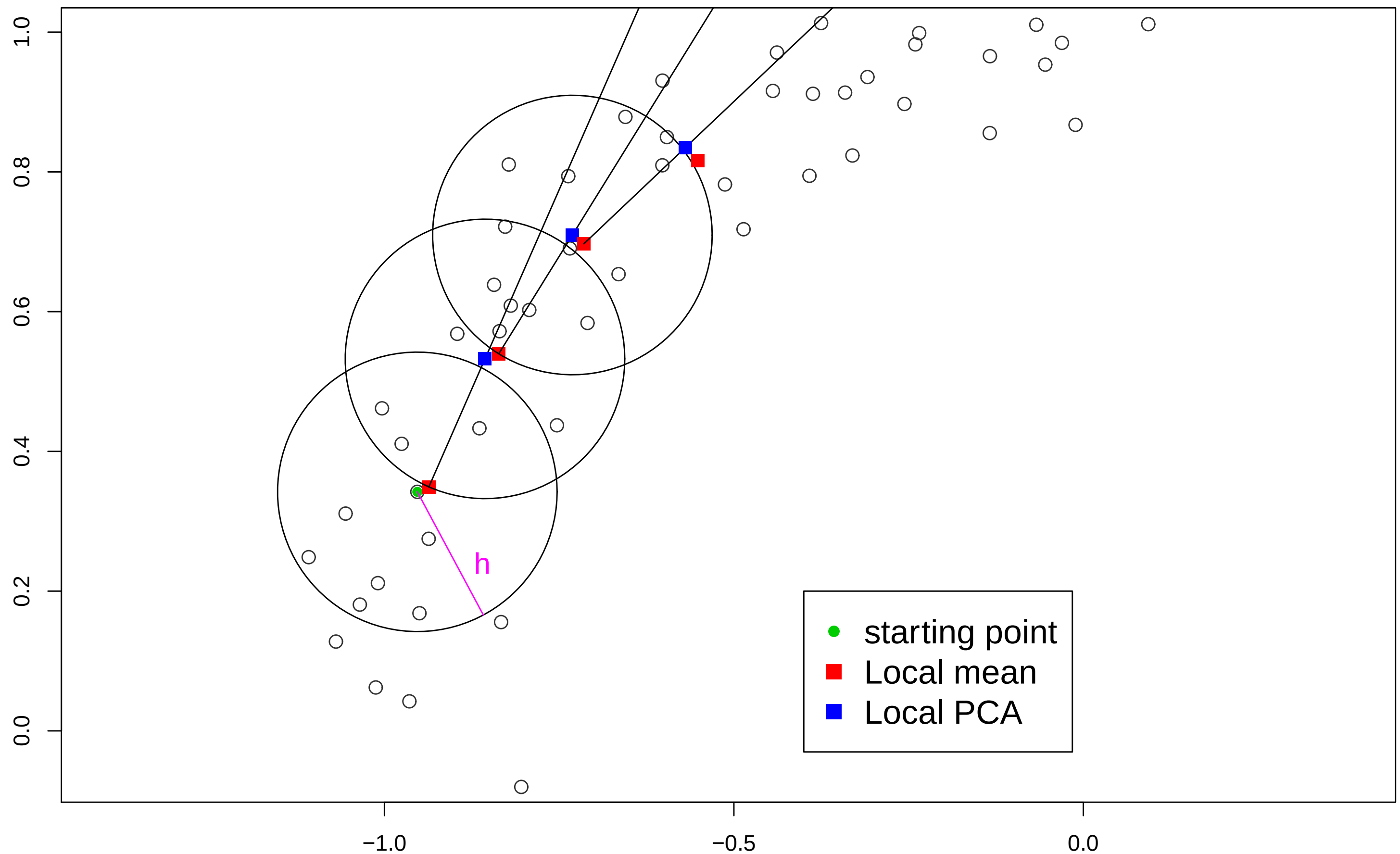
Principal curves

Principal curves are 'smooth curves through the middle of a data cloud'. Different principal curve algorithms vary in how the 'middle' of the data cloud is defined/found:

- Traditional: Global ('top-down') techniques.
  - Hastie & Stuetzle 1989: HS principal curves (R packages **pcurve** and **princurve**)
  - Tibshirani 1992: Probabilistic principal curves (no public implementation)
  - Kégl et al. 2002: Polygonal line algorithm (available as Java applet)
- Alternative: Local ('bottom up') methods.
  - Delicado 2001: Principal curves of oriented points (C++ programme)
  - Einbeck et al. 2005: Local principal curves (R package **LPCM**)
  - Ozertem & Erdogmus 2011: 'Locally defined principal curves' (no public implementation ?)

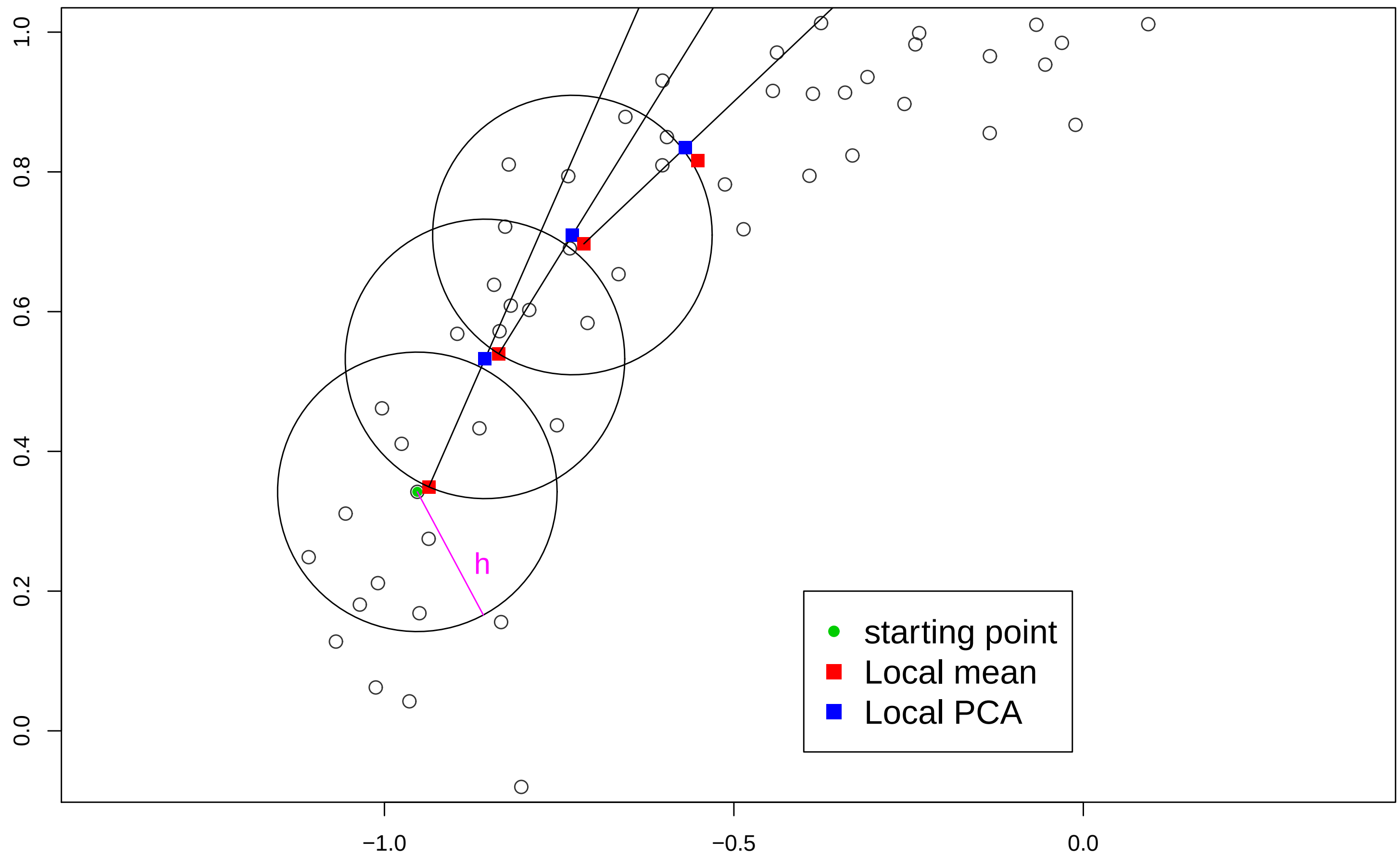
Local principal curves (LPC)

- Calculate alternately a local mean and a first local principal component, each within a certain bandwidth h.



Local principal curves (LPC)

- Calculate alternately a local mean and a first local principal component, each within a certain bandwidth h.

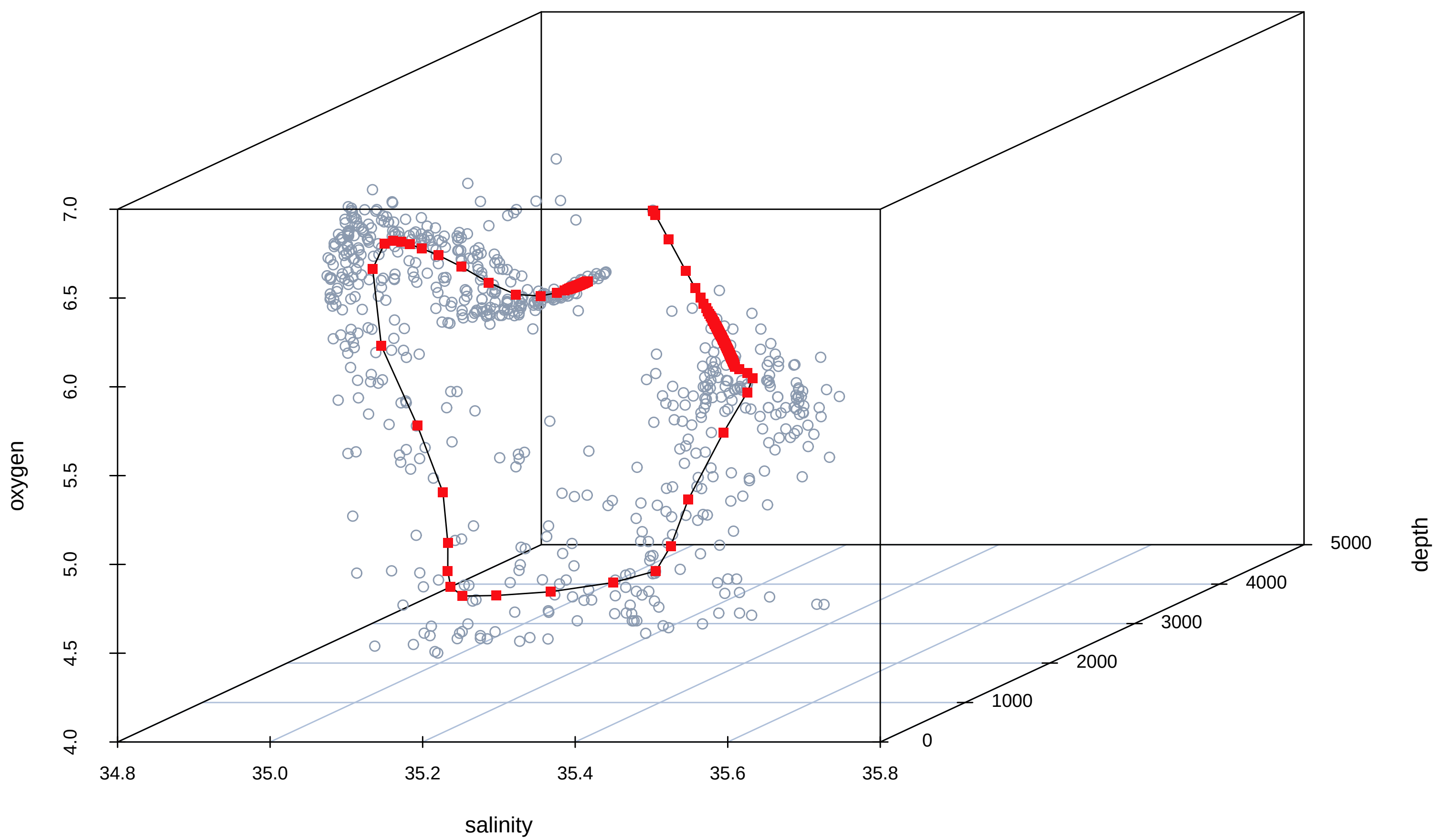


- The LPC is the series of local means.

Fitting the LPC

- LPC through oceanographic data set, with local centers of mass:

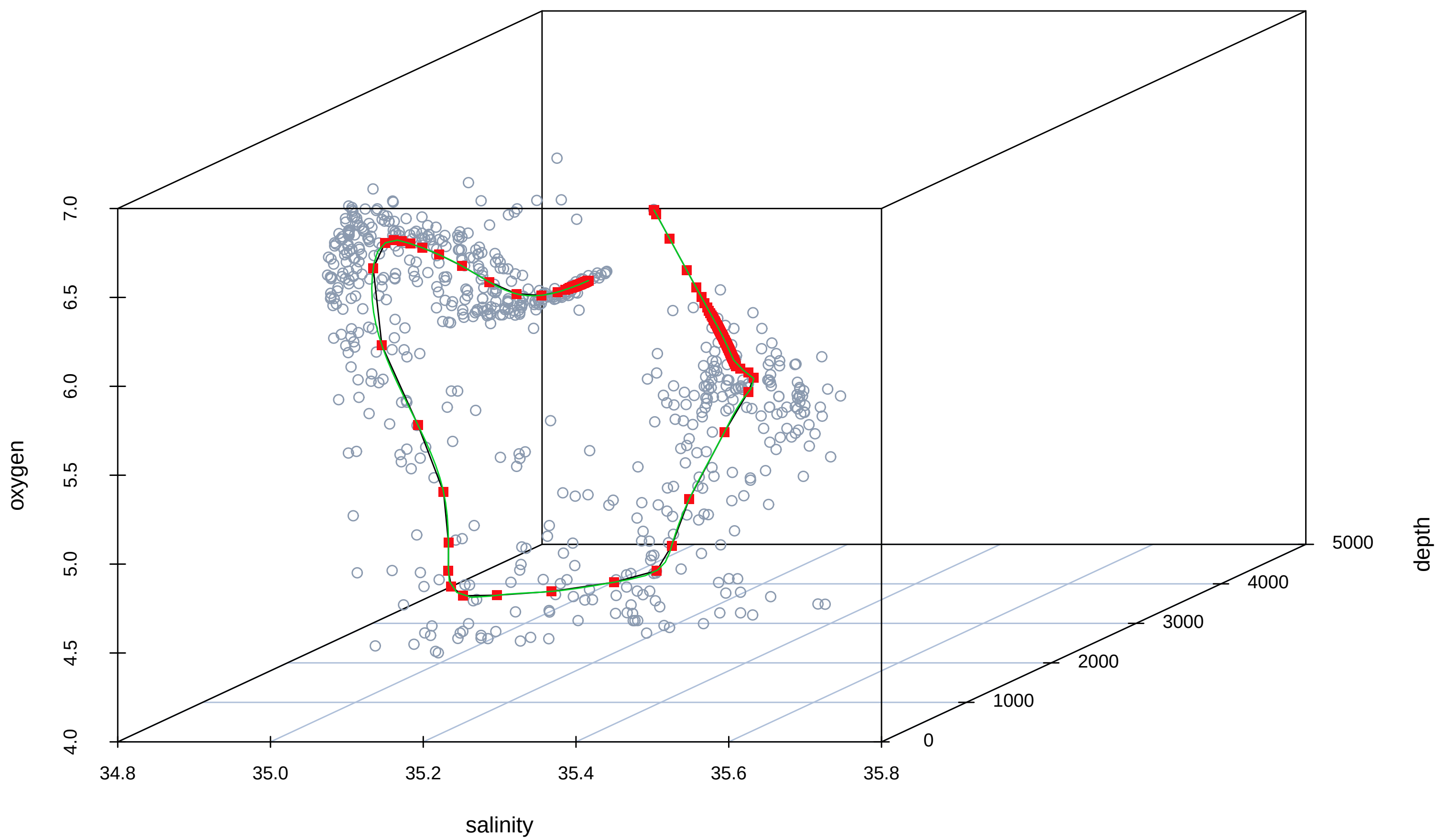
```
> require(LPCM)
> ocean.lpc <- lpc(ocean, h=0.12)
> plot(ocean.lpc, type=c("curve", "mass"))
```



Parametrizing the LPC

- We parametrize the LPC through the arc length of a cubic spline through the local centers of mass (Einbeck, Evers & Hinchliff, 2010).

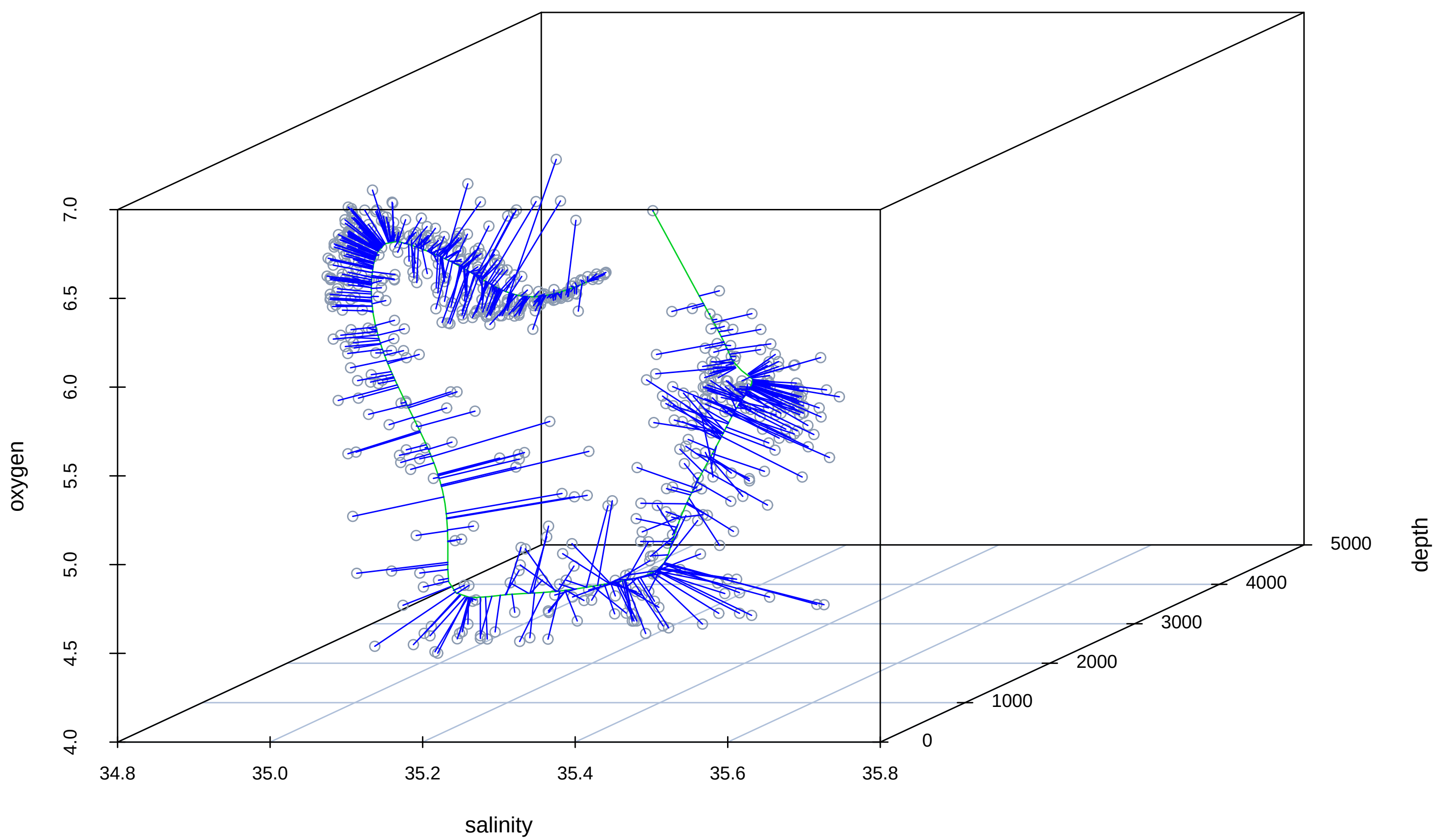
```
> ocean.spline <- lpc.spline(ocean.lpc)
> plot(ocean.spline, type=c("curve", "mass", "spline"))
```



Projecting onto the LPC

- We project each data point $\boldsymbol{x}_i \in \mathbb{R}^d$ onto the nearest point on the curve, yielding a one-dimensional projection index $p_i \in \mathbb{R}$

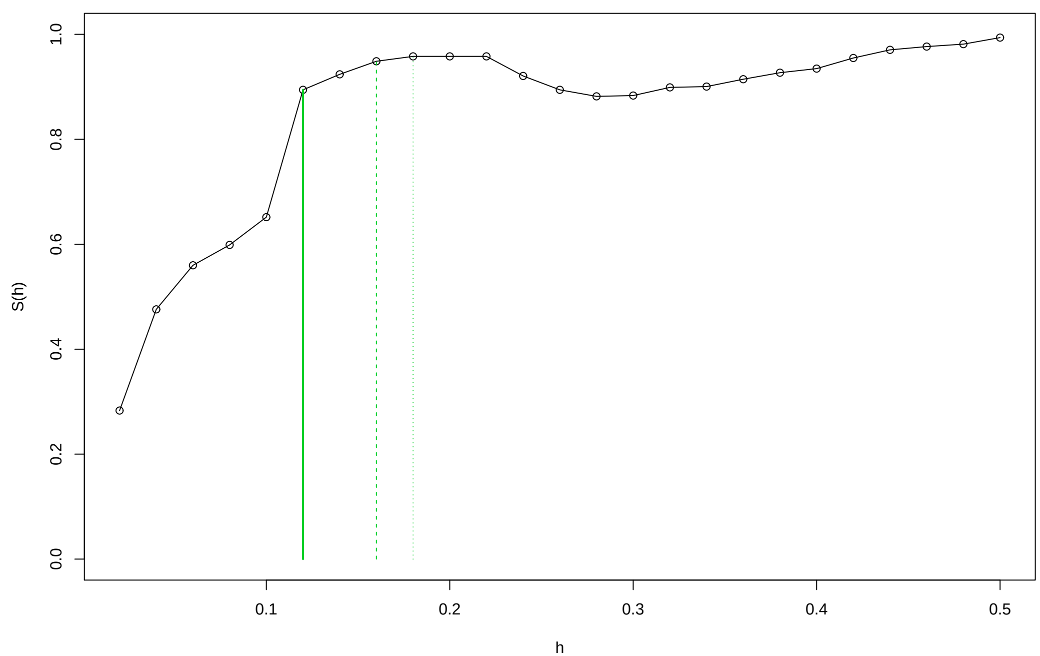
```
> plot(ocean.spline, type=c("spline", "project"))
```



Bandwidth selection

- Self-coverage: Proportion of data points within tubes around the curve of the same radius as the bandwidth used to fitted the curve (Einbeck, 2011).

```
> ocean.self<- lpc.self.coverage(ocean)
```

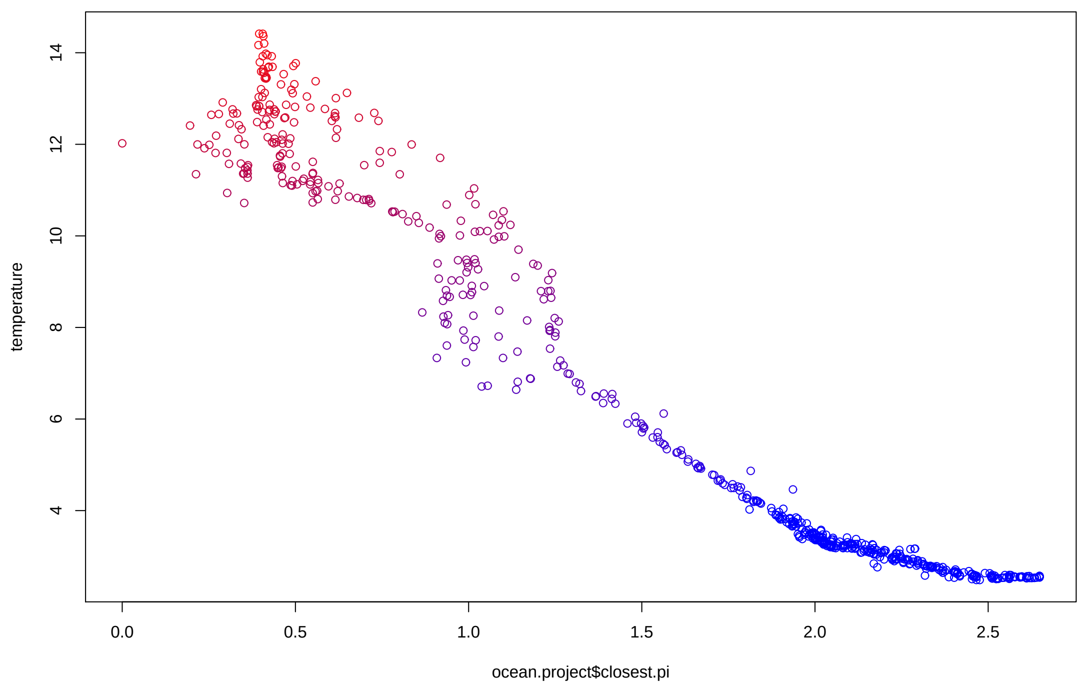


- gives $h = 0.12$

Regression based on the LPC

- Back to initial problem: With $y =$ temperature as response, it remains a univariate nonparametric regression problem $y_i = g(p_i) + \varepsilon_i$.

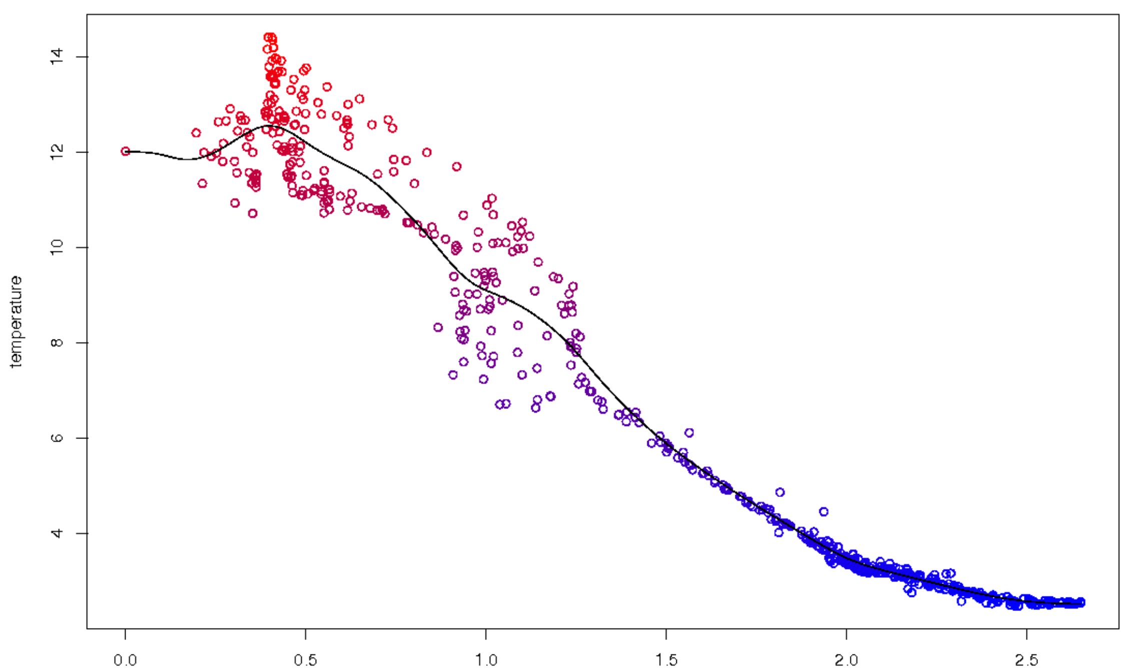
```
> pi <- lpc.spline(ocean.lpc, project=TRUE)
> plot(pi, temperature, ...)
```



Regression based on the LPC (cont.)

- This can be fitted by any nonparametric smoother; for instance, a local linear smoother.
- Could be considered as a single-index model with nonparametrically constructed index.

```
> require(KernSmooth)
> fit<- locpoly(pi[order(pi), temperature[order(pi)],...))
> lines(fit)
```



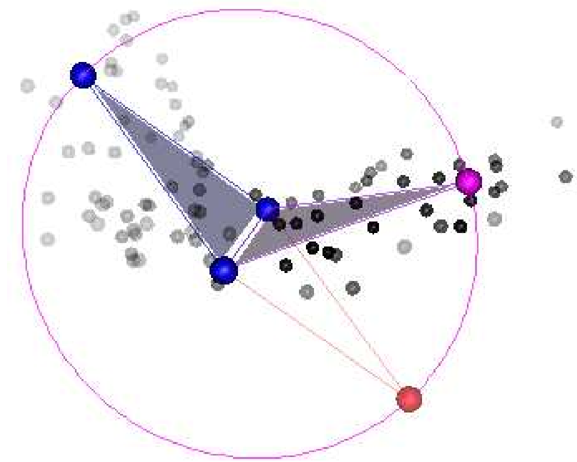
Principal surfaces

- Idea for local principal surfaces:
  - Build a mesh of "locally best fitting triangles".
  - Local PCA is (only) used to define the initial triangle.

Starting from the initial triangle, iteratively . . .

(1) glue further triangles at each of its sides.

(2) adjust free vertexes via a constrained mean shift. Dismiss a new triangle if the new vertex
  - falls below a density threshold
  - is too close to an existing one.



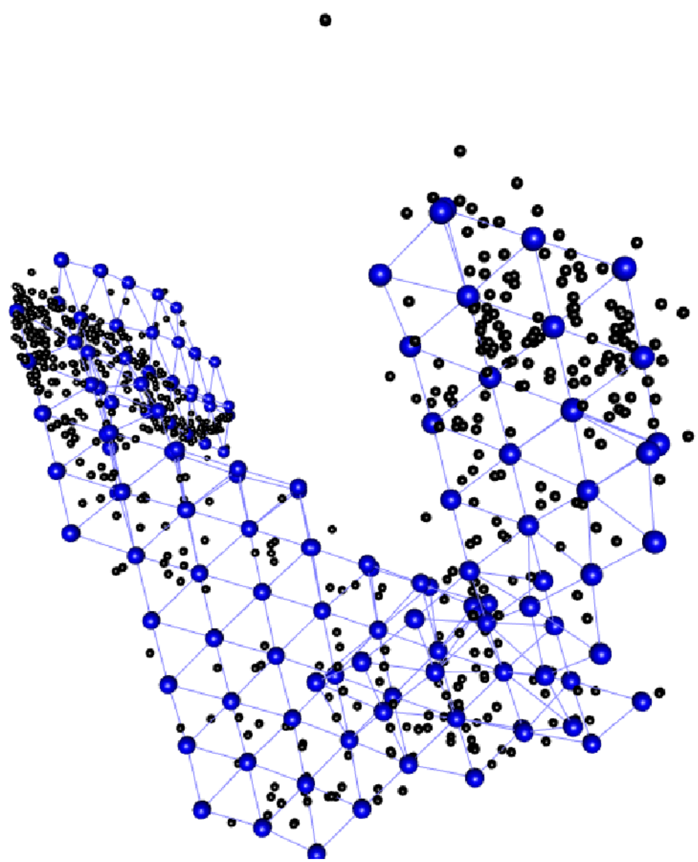
. . . until all triangles have been considered.

(Einbeck, Evers & Powell, 2010)

Principal surfaces (cont.)

- Local principal surface fitted to oceanographic data:

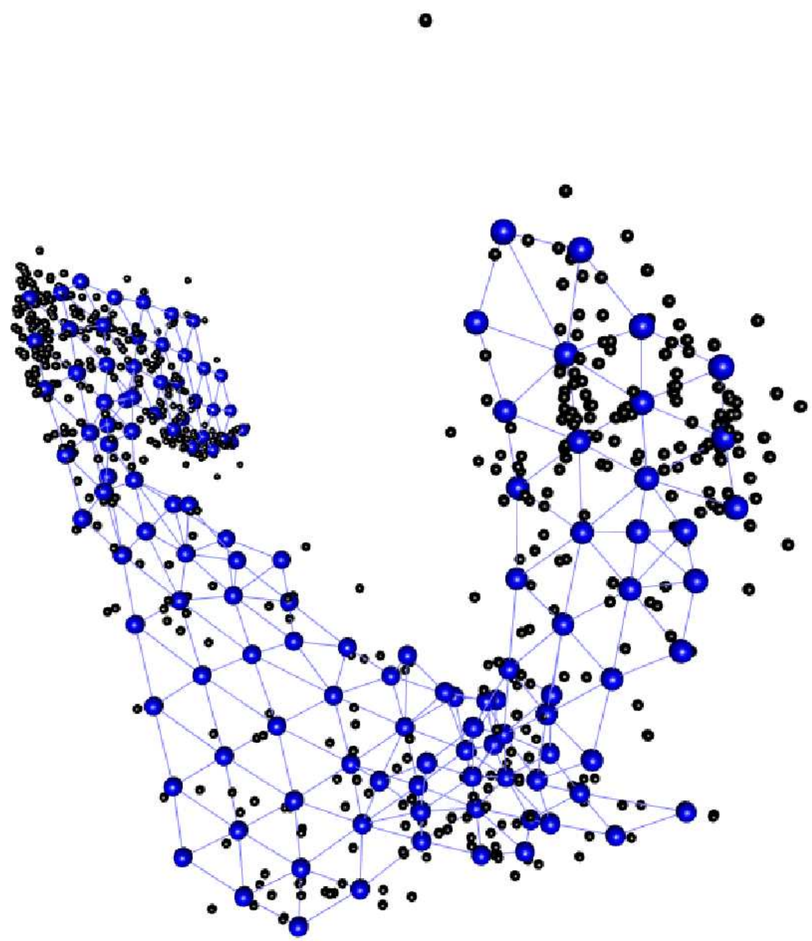
```
> library(lpmforge)   # by L. Evers, under construction
> ocean.lpm <- lpm(ocean, h=120)
> plot3d(ocean.lpm)
```



Principal surfaces (cont.)

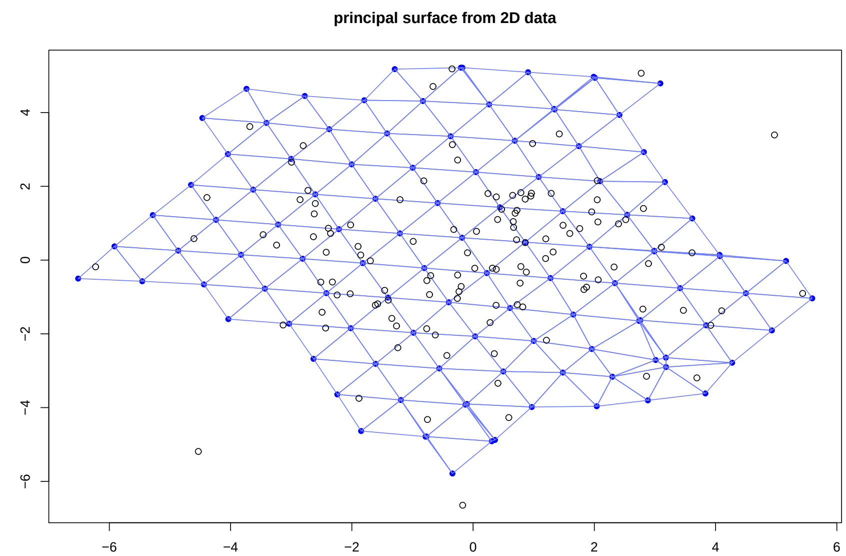
- Postprocessing via elastic net (Gorban and Zonovyev, 2005)

```
> ocean2.lpm<- postprocess.lpm(ocean.lpm)
> plot3d(ocean2.lpm)
```



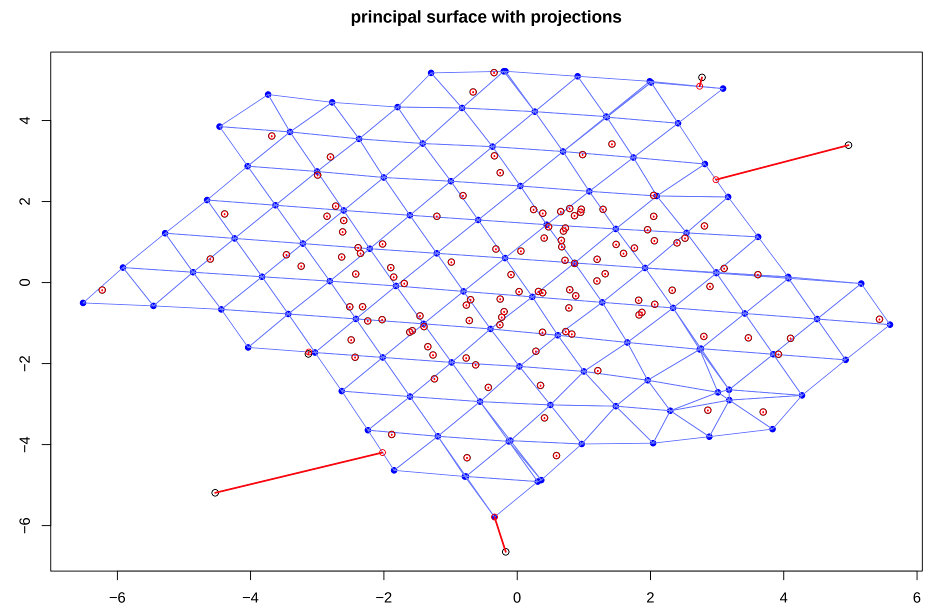
Regression on principal surface

- Toy example: A principal surface for bivariate data.



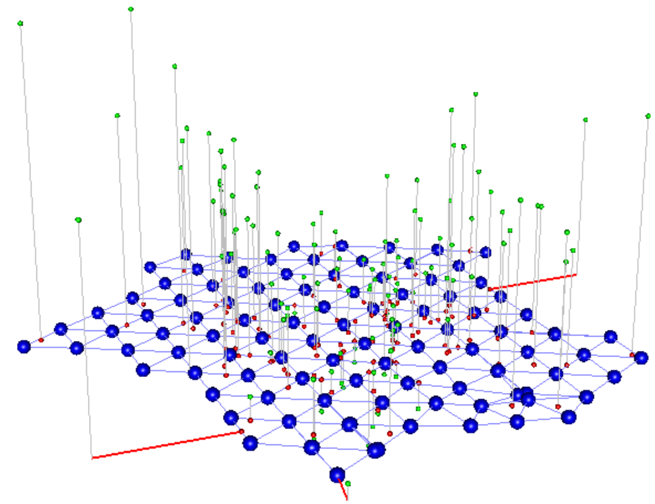
Regression on principal surface

- Toy example: A principal surface for bivariate data.
- Initially, each data point $\mathbf{x}_i$ is projected onto the closest triangle (or simplex), say t_i.



Regression on principal surface

- Toy example: A principal surface for bivariate data.
- Initially, each data point $\mathbf{x}_i$ is projected onto the closest triangle (or simplex), say t_i.



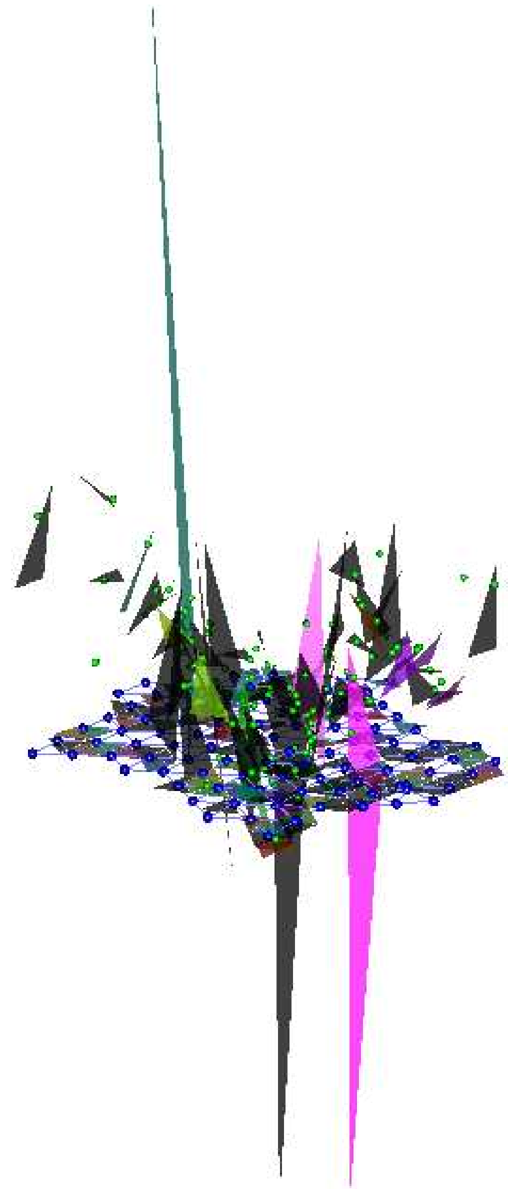
- Next, consider a response y_i.
- We can fit separate regression models for each triangle j

$$y_i = \mathbf{c}^{(j)}(\mathbf{x}_i)'\boldsymbol{\beta}_{(j)} + \epsilon_i \qquad \text{for all } i \text{ with closest triangle } t_i = j,$$

where $\mathbf{c}^{(j)}(\mathbf{x}_i)$ are the coordinates of the projected point using the sides of the $j-$th triangle as basis functions.

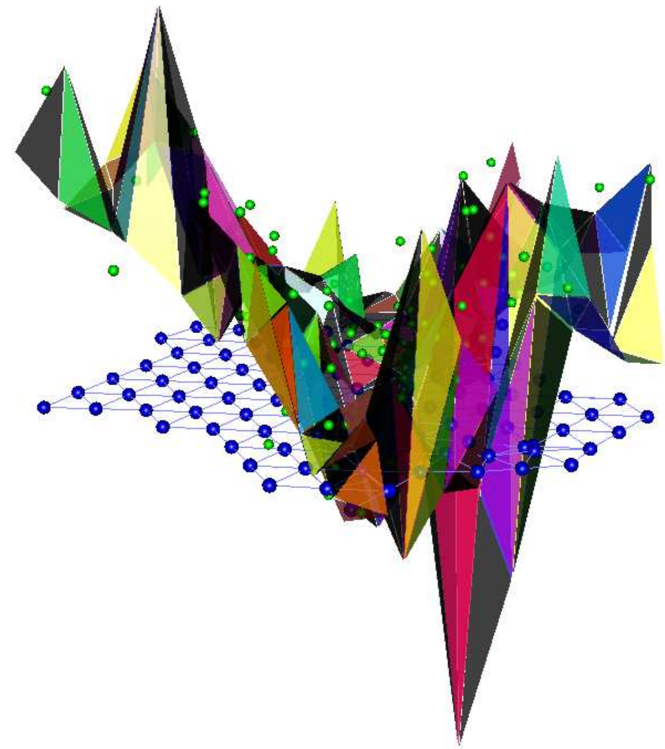
Penalized regression

- Fitting totally unrelated regressions within each triangle is clearly unsatisfactory.



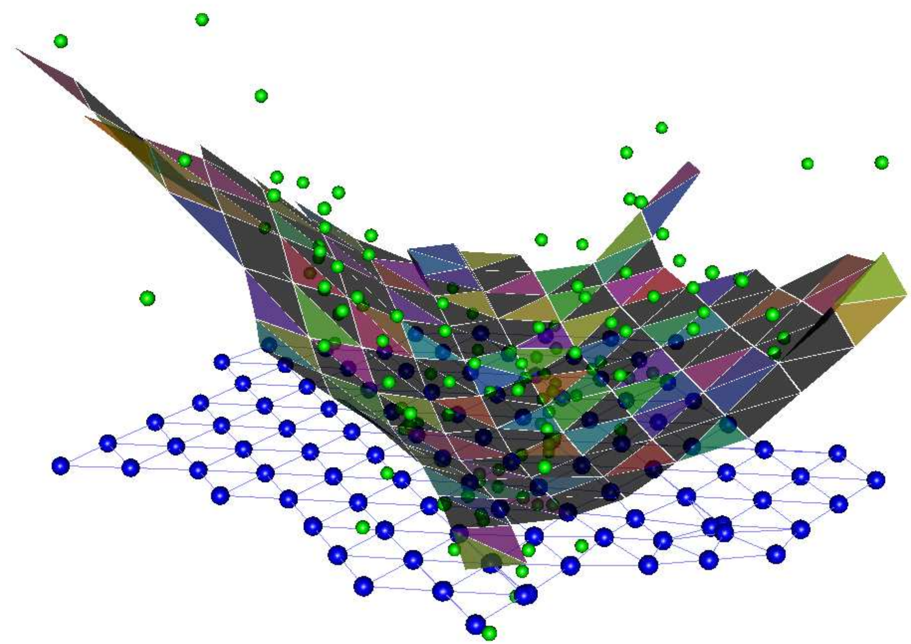
Penalized regression

- Fitting totally unrelated regressions within each triangle is clearly unsatisfactory.
- Therefore, we apply an continuity penalty which which penalizes differences between predictions of neighboring triangles at shared vertices.



Penalized regression

- Fitting totally unrelated regressions within each triangle is clearly unsatisfactory.
- Therefore, we apply an continuity penalty which which penalizes differences between predictions of neighboring triangles at shared vertices.
- Additionally, we apply a smoothness penalty which penalizes difference in regressions at adjacent triangles.



Penalized regression (cont'd)

- Define
  - the parameter vector $\boldsymbol{\beta}' = \left(\boldsymbol{\beta}'_{(1)}, \boldsymbol{\beta}'_{(2)}, \ldots\right)$,
  - the design matrix $\boldsymbol{Z}$ (which is a box product of $(\mathbf{c}^{(t_i)}(\mathbf{x}_i))_{1\le i\le n}$ and an adjacency matrix);
  - appropriate penalty matrices $\boldsymbol{D}$ and $\boldsymbol{E}$.
- Then the entire minimization problem can be written as

$$\|\mathbf{Z}\boldsymbol{\beta} - \mathbf{y}\|^2 + \lambda\|\mathbf{D}\boldsymbol{\beta}\|^2 + \mu\|\mathbf{E}\boldsymbol{\beta}\|^2. \tag{1}$$

- The solution is given by

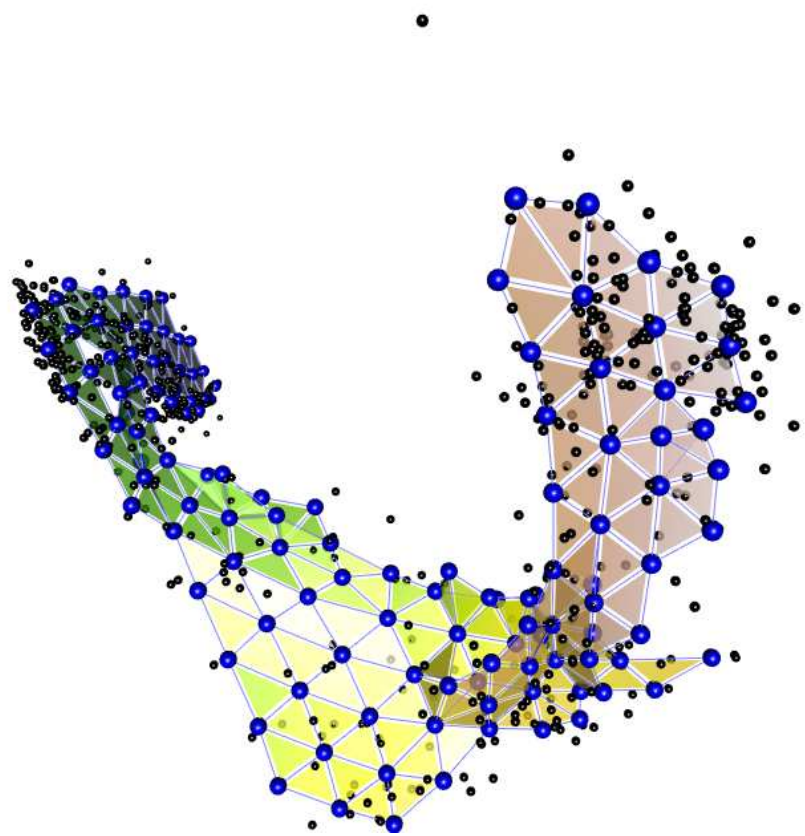
$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z} + \lambda\mathbf{D}'\mathbf{D} + \mu\mathbf{E}'\mathbf{E}\ )^{-1}\mathbf{Z}'\mathbf{y}.$$

(Einbeck, Isaac, Evers & Parente, 2012)

Back to oceanographic data

- Penalized regression of water temperature on principal surface

```
> ocean.reg <- regression.lpm(ocean2.lpm, temperature,
      penalty.continuity=1, penalty.smoothness=1)
> plot(ocean.reg)
```



Conclusion

- Principal curves and surfaces can be used as a building block for further statistical procedures (such as, nonparametric regression).
- Techniques are only suitable for data with very high inter–variable correlations.
- R package **LPCM** (on CRAN)
  - Principal curve fitting (incl. parametrization and projection)
  - Bandwith selection
  - Measuring goodness–of–fit
  - Mean shift (clustering) tools
- R package **lpmforge** (in development, L. Evers)
  - Fitting principal surfaces and manifolfs of higher dimension
  - Includes functionalities for post-processing (elastic net), projection, and regression.
  - No automated smoothing parameter selection yet.
  - Finding the 'right' dimension of the manifold is another issue...

References

Delicado (2001): Another Look at Principal Curves and Surfaces, *Journal of Multivariate Analysis* **77**, 84–116.

Gorban & Zinovyev (2005): Elastic principal graphs and manifolds and their practical application. *Computing* **75**, 359–399.

Hastie & Stuetzle (1989): Principal Curves. *JASA* **84**, 502–516.

Kégl, Krzyzak, Linder,& Zeger (2000): Learning and Design of Principal Curves. *IEEE Transactions Patt. Anal. Mach. Intell.* **24**, 59–74.

Ozertem & Erdogmus (2011): Locally defined principal curves, *Journal of Machine Learning Research* **12**, 1249–1286.

Tibshirani (1992): Principal Curves Revisited. *Statistics and Computing* **2**, 183–190.

References (cont.)

Einbeck, Tutz & Evers (2005): Local principal curves. *Statistics and Computing* **15**, 301–313.

Einbeck, Evers & Hinchliff (2010): Data compression and regression based on local principal curves. In Fink et al. (Eds): Advances in Data Analysis, Data Handling, and Business Intelligence, Heidelberg, pp. 701–712, Springer.

Einbeck, Evers & Powell (2010): Data compression and regression through local principal curves and surfaces. *International Journal of Neural Systems* **20**, 177–192.

Einbeck (2011). Bandwidth selection for mean-shift based unsupervised learning techniques - a unified approach via self-coverage, *Journal of Pattern Recognition Research* **6**, 175–192.

Einbeck, Isaac, Evers & Parente (2012): Penalized regression on principal manifolds with application to combustion modelling. *Proc' of the 27th International Workshop on Statistical Modelling*, University of Economics, Prague.