Localized regression on principal manifolds

Jochen Einbeck

Department of Mathematical Sciences, Durham University

jochen.einbeck@durham.ac.uk

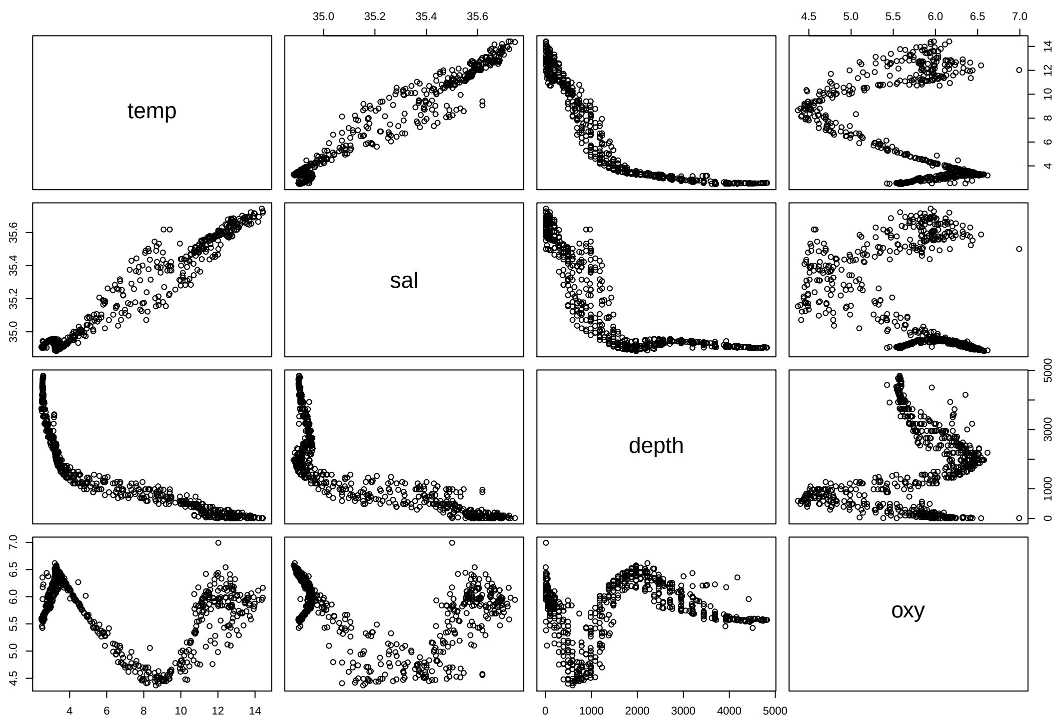
joint work with Ludger Evers (University of Glasgow),

Glasgow, 8th of July 2010

Durham University

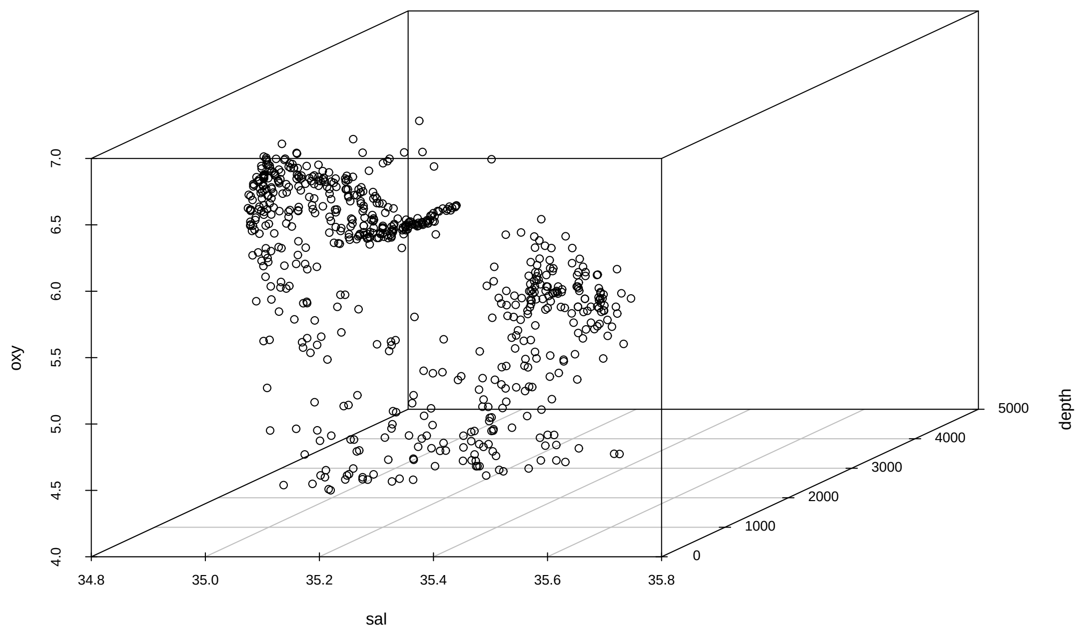
Motivation

- 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.



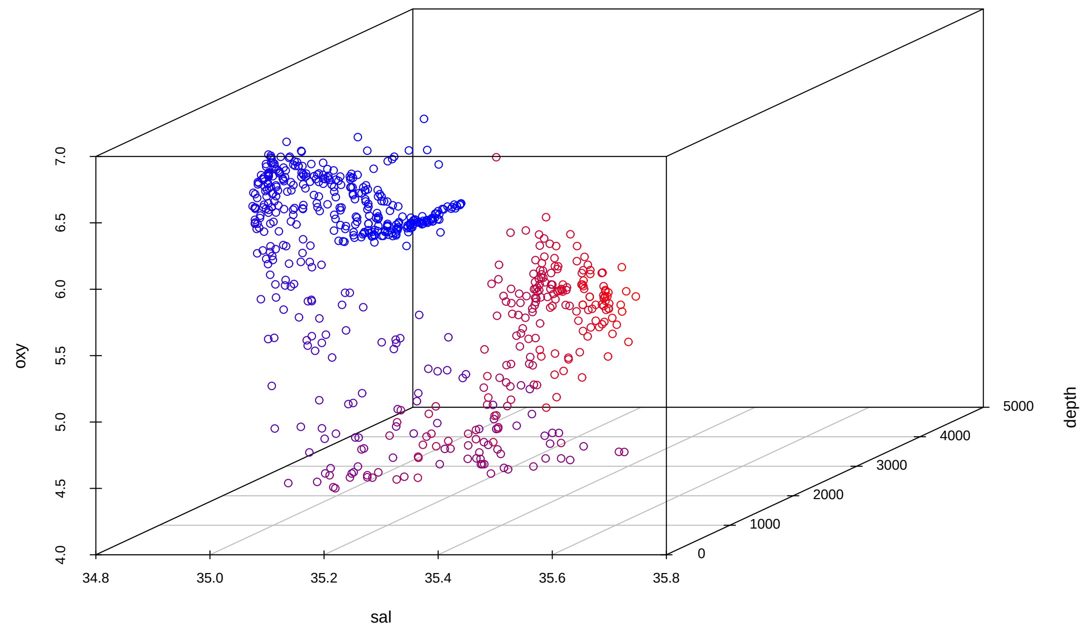
Motivation (cont.)

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



Motivation (cont.)

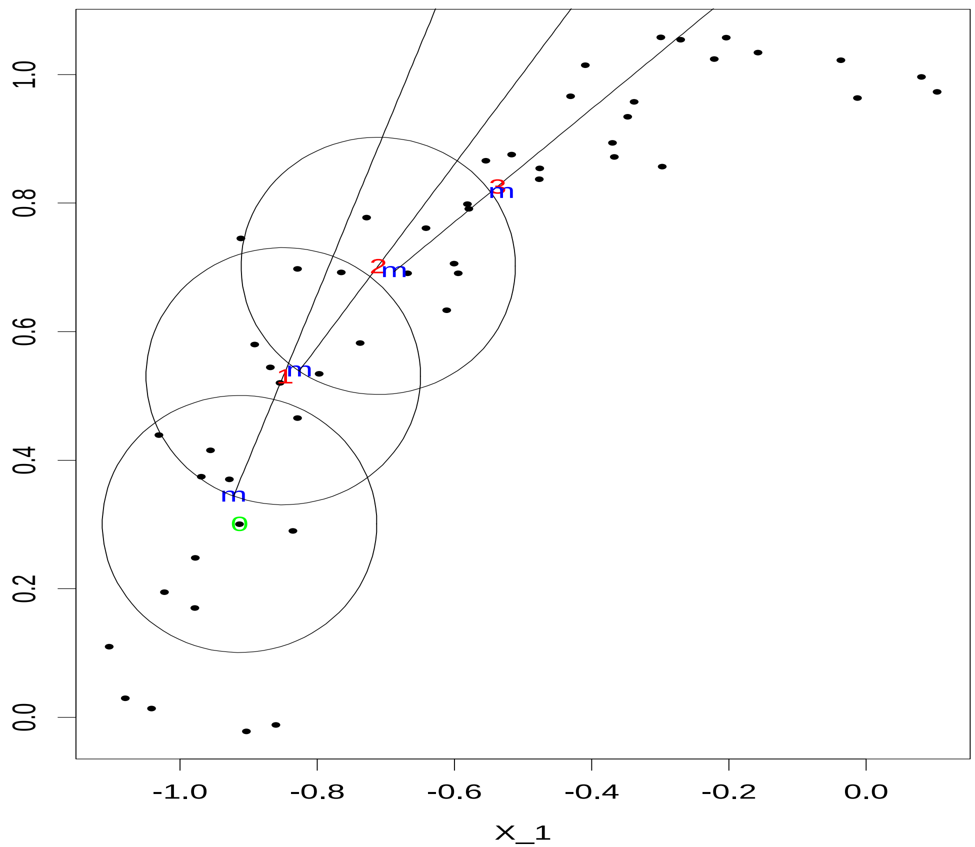
- This is a 3-variate regression problem, with the predictor space given by salinity, water depth, and oxygen:



- We shade higher water temperatures red.
- Can we make use of the one-(?) dimensional inner structure?
- This is a task for principal curves (Hastie & Stuetzle, 1989).

Local principal curves (LPCs)

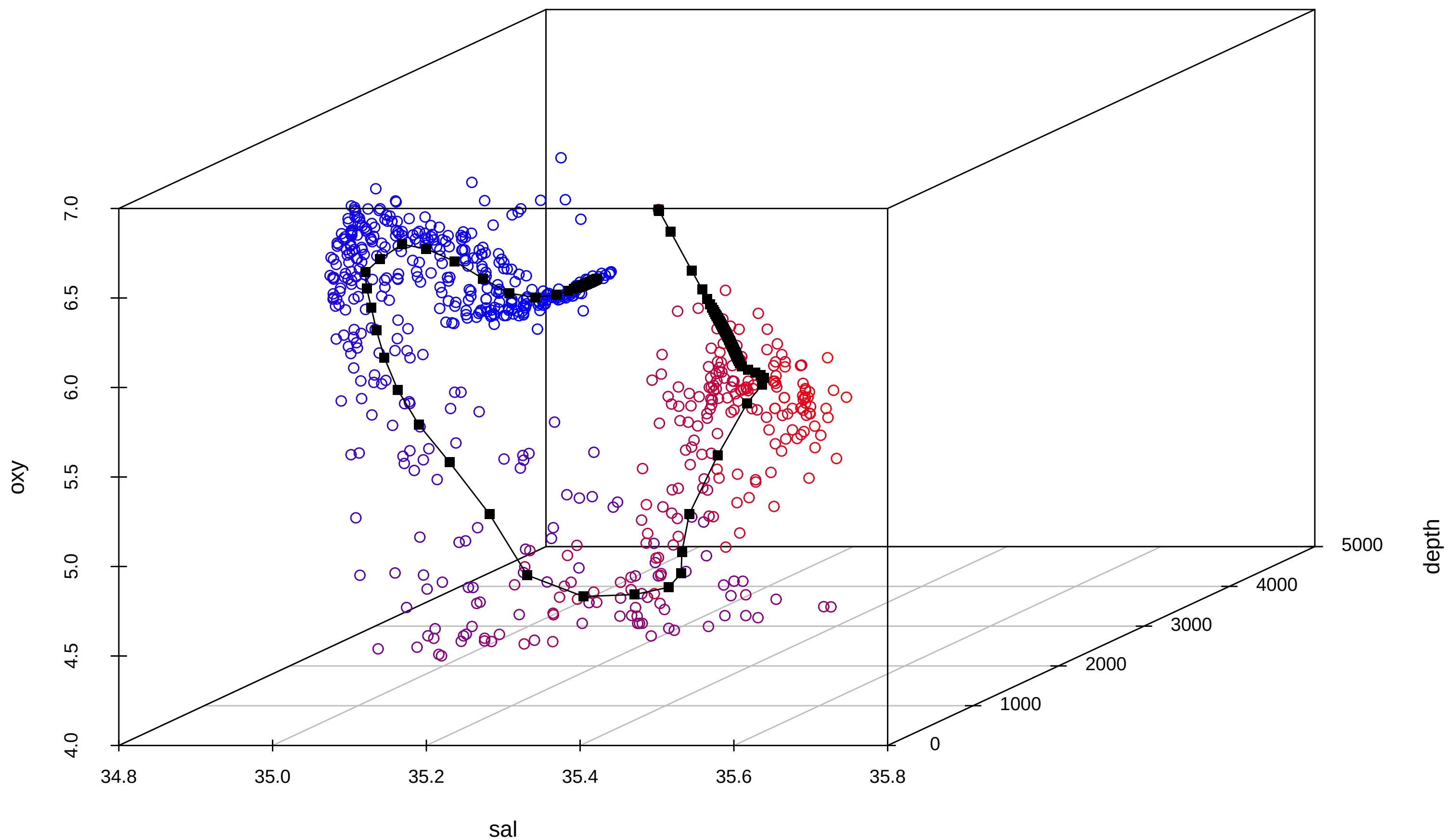
- Idea: Calculate alternately a local center of mass and a first localized principal component (Einbeck, Tutz, & Evers, 2005).



0: starting point,
$\boldsymbol{m}$: points of the LPC,
$1, 2, 3$: enumeration of steps.

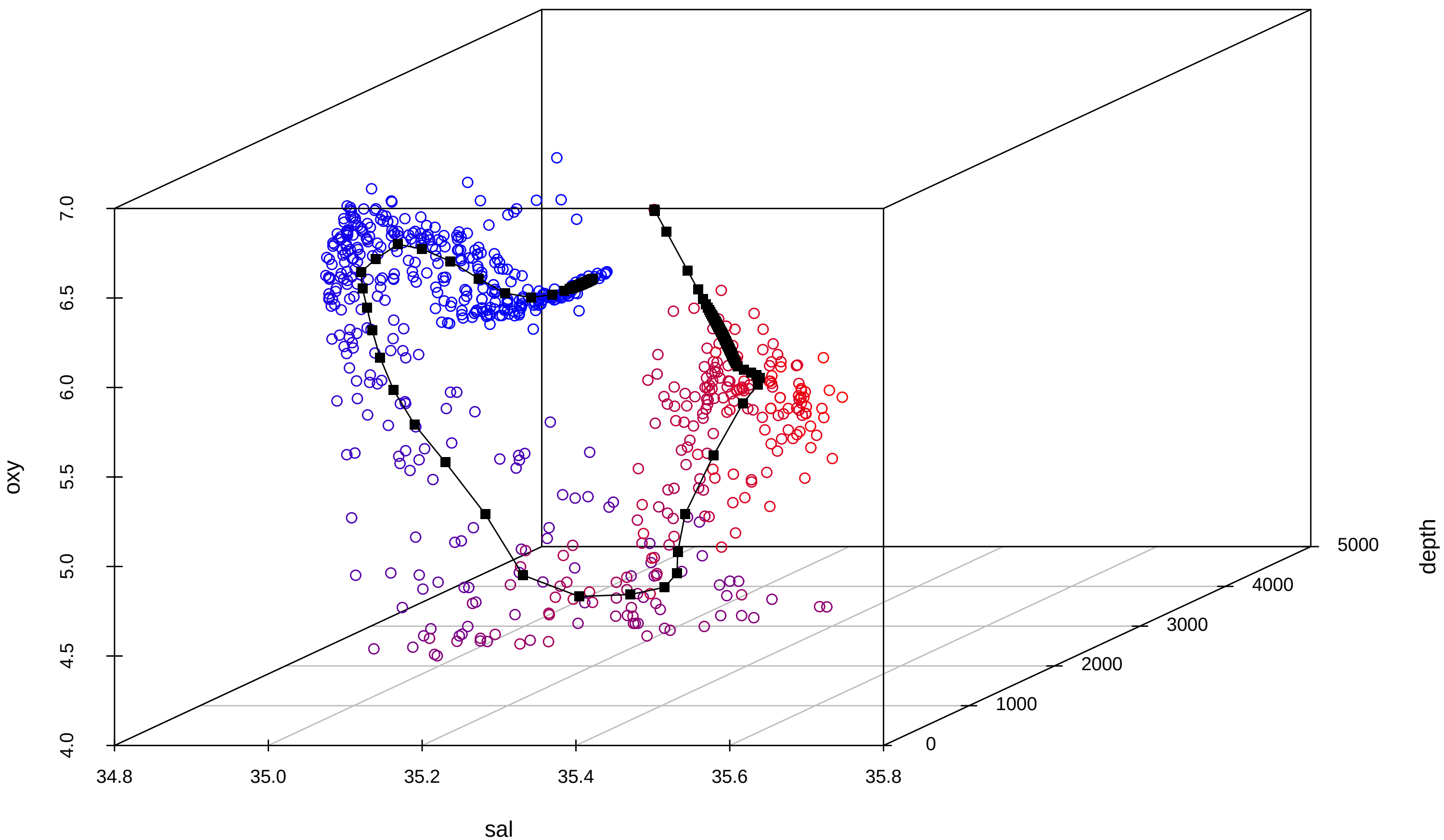
Fitting the LPC

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



Fitting the LPC

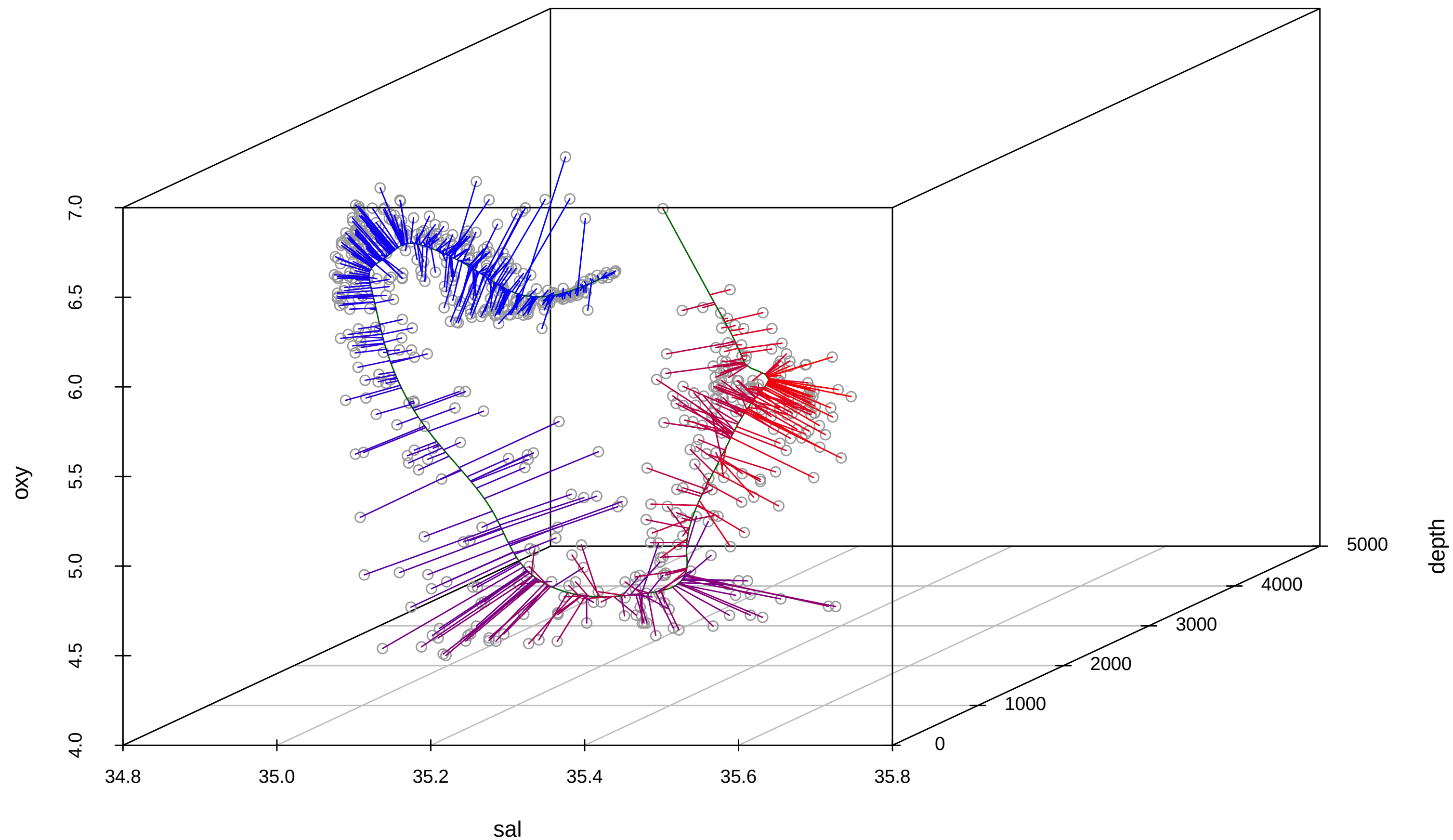
- LPC through oceanographic data set, with local centers of mass:



- The curve has yet to be parametrized, and one needs to be able to project the data points onto it.

Projecting onto the LPC

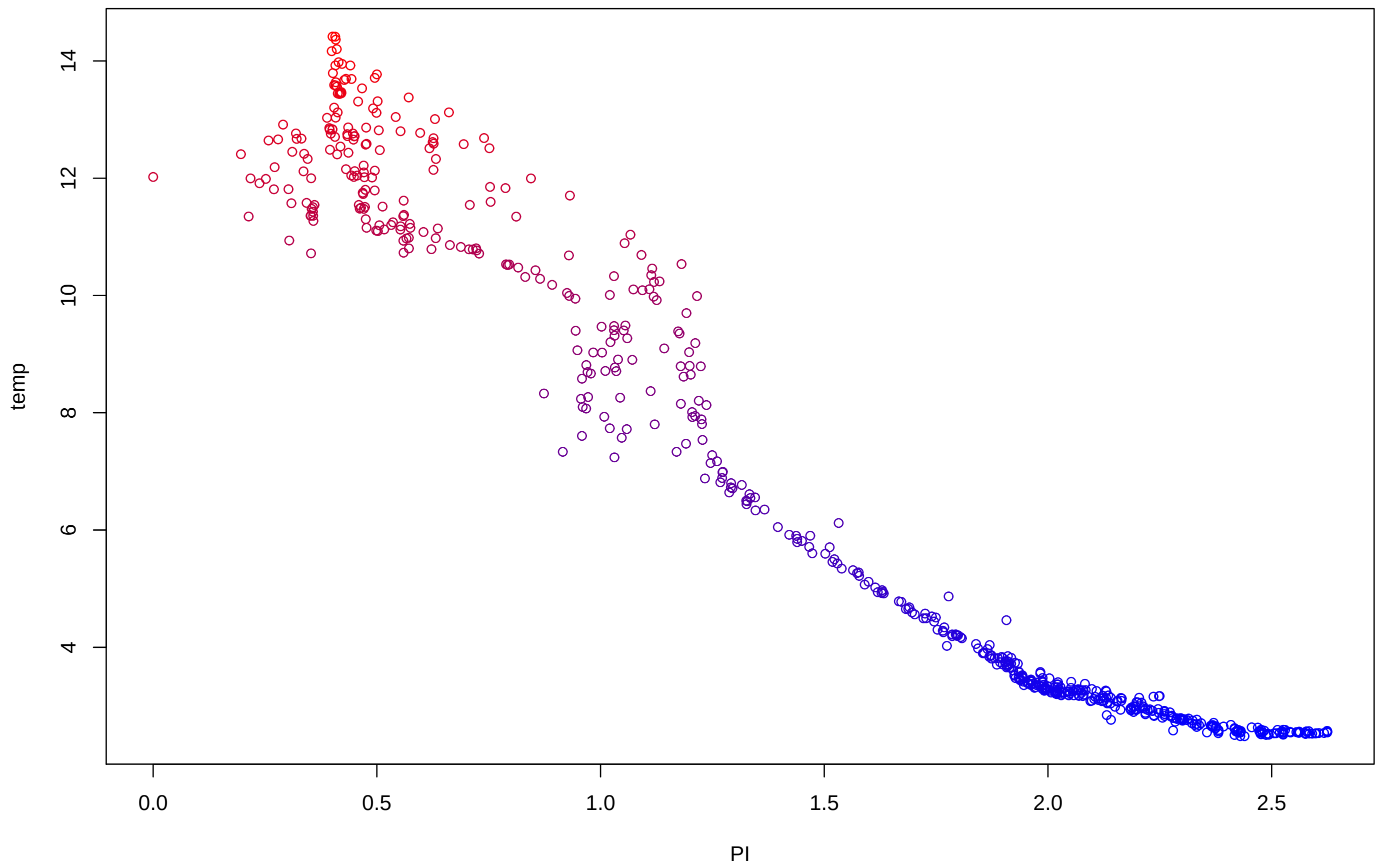
- We parametrize the LPC through the arc length of a cubic spline function laid through the local centers of mass, and project each data point $\boldsymbol{x}_i \in \mathbb{R}^d$ onto the nearest point on the curve, yielding a one-dimensional projection index $t_i \in \mathbb{R}$ (Einbeck, Evers, & Hinchliff, 2010).



Regression based on the LPC

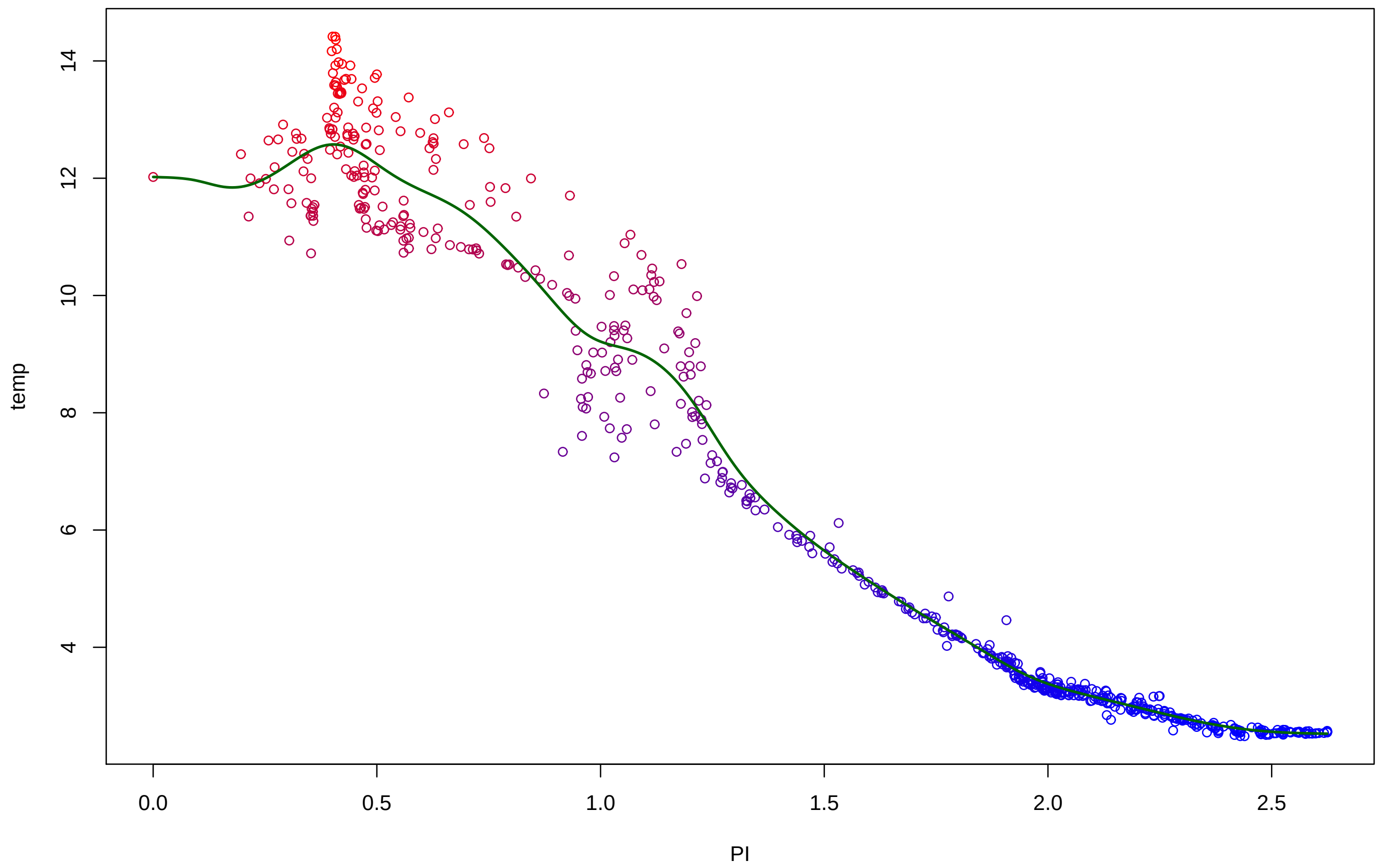
- It remains a simple univariate regression problem of type

$$y_i = g(t_i) + \varepsilon_i.$$



Regression based on the LPC

- It remains a simple univariate regression problem of type
$$y_i = g(t_i) + \varepsilon_i.$$



- This can be fitted any nonparametric smoother; for instance, a local linear smoother.

Manifolds of higher order?

- This may be considered as unsatisfactory: The data corresponding to "hot" temperatures show a branched structure, indicating that some information relevant for the response is orthogonal to the principal curve.
- If we deem the predictor space to be of intrinsic dimensionality 2 (rather than 1), we need to fit a principal surface (rather than a principal curve).
- Can we extend the local principal curve algorithm towards local principal surfaces (or manifolds of higher dimension)?

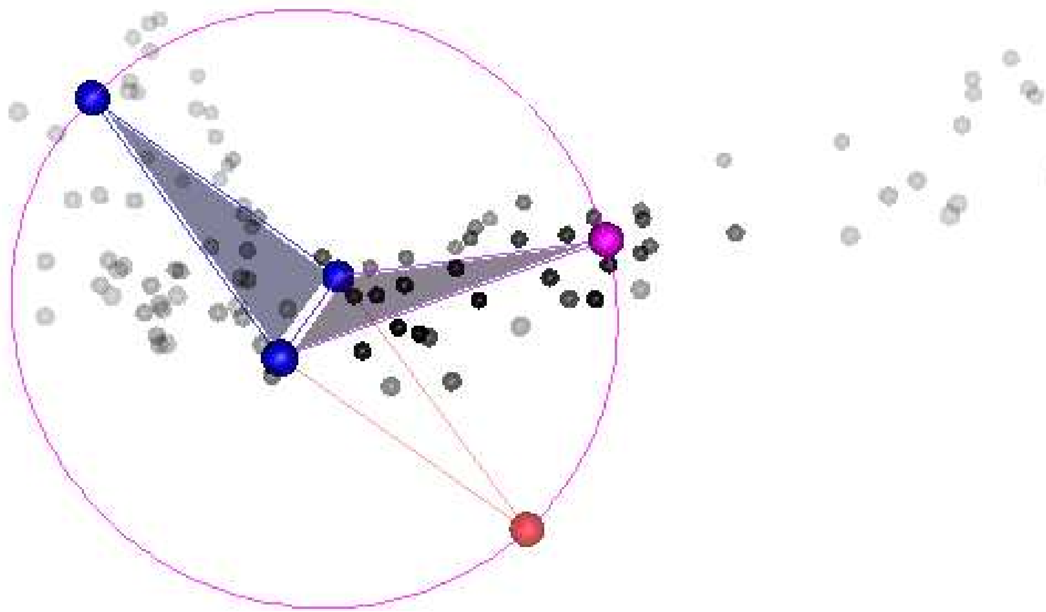
Local principal surfaces

- We are working now with the "building block" triangles Δ.
- Local PCA is only used to determine the initial triangle, say Δ_0.
- Then, the algorithm iterates
  (1) For a given triangle Δ, we glue further triangles at each of its sides $j = 1, 2, 3$.
  (2) For $j = 1, 2, 3$, adjust the free triangle vertex via the mean shift. We dismiss the new triangle if
    - the new vertex falls into a region of small density, or
    - the new vertex is too close to an existing one (Delaunay triangulation).

  until all sides of all triangles (including the new ones) have been considered.

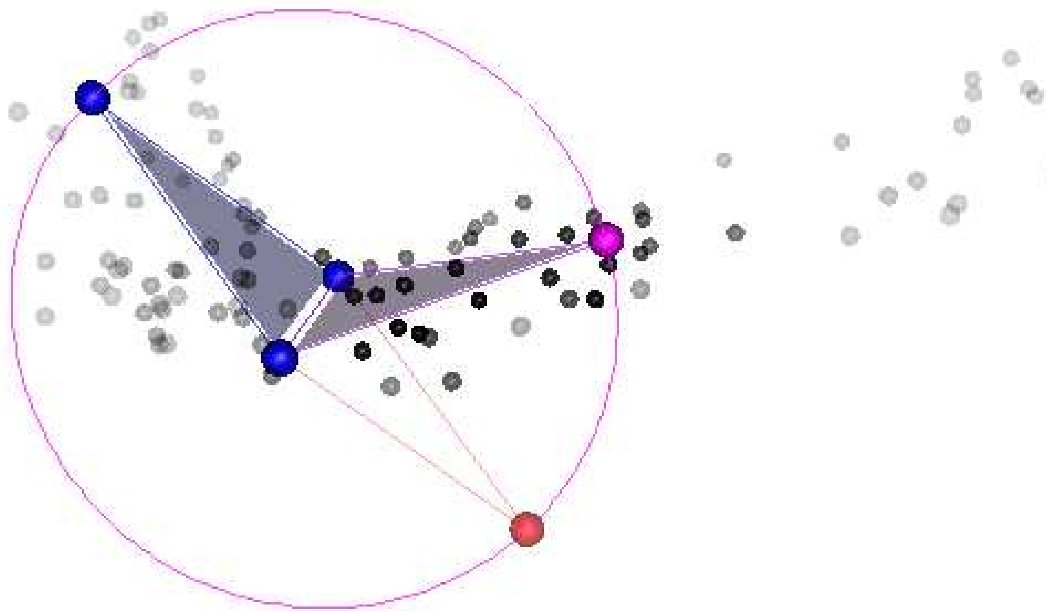
Local principal surfaces (cont.)

- Illustration: Constrained mean shift on a circle (enforcing equilateral triangles):



Local principal surfaces (cont.)

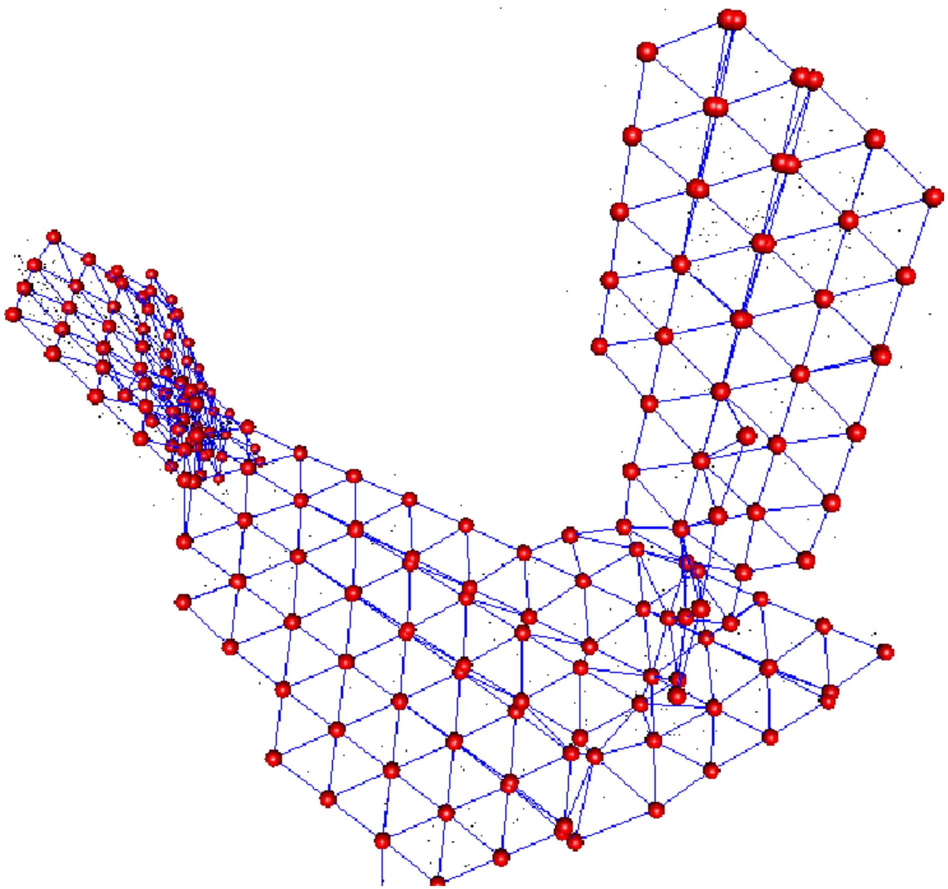
- Illustration: Constrained mean shift on a circle (enforcing equilateral triangles):



- Extendable to local principal manifolds (LPMs) of arbitrary dimension > 2 by replacing "triangles" with suitable "tetrahedrons" or "simplices".

Local principal surfaces (cont.)

- Local principal surface (LPS) for oceanographic data set:



Regression on the surface

- Then, how to use this surface for regression?
- It seems hard to define a meaningful 2-dim. parametrization on the surface.
- However, we may use *distances* instead: For each triangle, we can count the distance d to all other triangles through the smallest number of triangle borders that have to be crossed to walk from one to the other.
- Assign local weights via discrete distance-based kernel

$$\kappa(d) = e^{-d/\lambda}$$

The parameter $\lambda \in [0, \infty)$ steers the degree of smoothing on the manifold: the higher λ, the smoother.

Regression on the surface (cont.)

The entire fitting process is summarized as follows:

(I) Fit a LPS as explained above, yielding a surface with, say, R triangles.

(II) Assign each data point $\boldsymbol{x}_i, i = 1, \ldots, n$ to their nearest triangle.

(III) For each triangle $r = 1, \ldots, R$, compute the mean $\bar{y}_r$ over the response values of all data points assigned to it.

(IV) Compute all pairwise distances $d_{r,s}$ between all triangles on the surface.

(V) Use the discrete kernel $\kappa(\cdot)$ to smooth over the manifold. The smoothed response value g_r on triangle r is given by

$$g_r = \frac{\sum_s \kappa(d_{r,s})\bar{y}_s}{\sum_s \kappa(d_{r,s})}.$$

Simulation study

- We divide the data into a training and a test data set of size 500 and 143, respectively.
- Squared prediction errors for the additive model (AM), LPC– and LPS– based regression are given below.

		AM	LPC	LPS $\lambda = 0.2$	LPS $\lambda = 1$	LPS $\lambda = 2$
Training error	mean	0.089	0.326	0.043	0.073	0.144
	median	0.015	0.007	0.001	0.007	0.015
Test error	mean	0.155	0.310	0.111	0.116	0.175
	median	0.029	0.009	0.004	0.010	0.021

Simulation study

- We divide the data into a training and a test data set of size 500 and 143, respectively.
- Squared prediction errors for the additive model (AM), LPC– and LPS– based regression are given below.

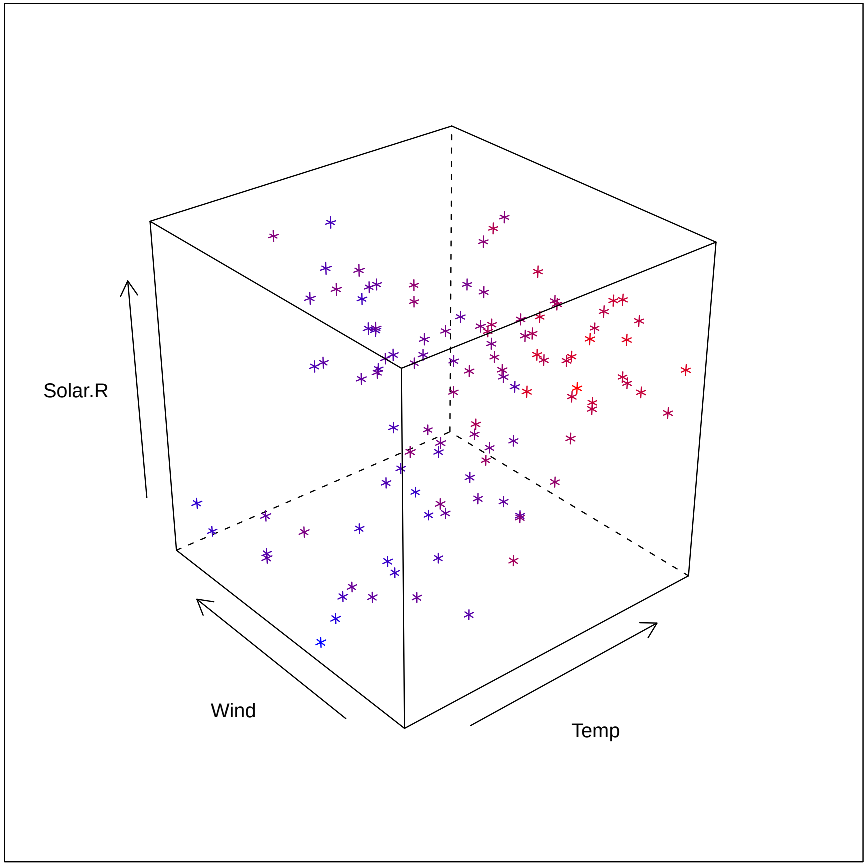
		AM	LPC	LPS $\lambda = 0.2$	LPS $\lambda = 1$	LPS $\lambda = 2$
Training error	mean	0.089	0.326	0.043	0.073	0.144
	median	0.015	0.007	0.001	0.007	0.015
Test error	mean	0.155	0.310	0.111	0.116	0.175
	median	0.029	0.009	0.004	0.010	0.021

- The LPS for $\lambda = 1$ performs superior to all other techniques.

New York Air Quality Measurements

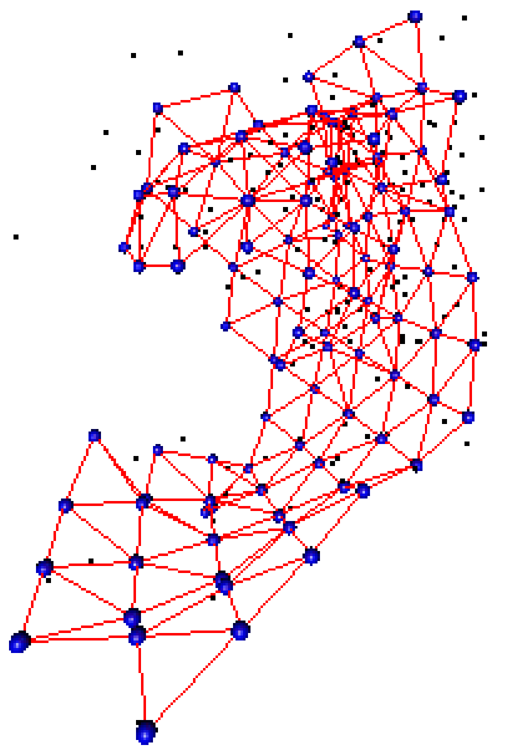
```
      Ozone Solar.R Wind Temp
1       41     190  7.4   67
...
24      32      92 12.0   61
25      NA      66 16.6   57
26      NA     266 14.9   58
27      NA      NA  8.0   57
28      23      13 12.0   67
29      45     252 14.9   81
30     115     223  5.7   79
31      37     279  7.4   76
32      NA     286  8.6   78
33      NA     287  9.7   74
34      NA     242 16.1   67
35      NA     186  9.2   84
36      NA     220  8.6   85
...
153     20     223 11.5   68
```

High ozone levels = red



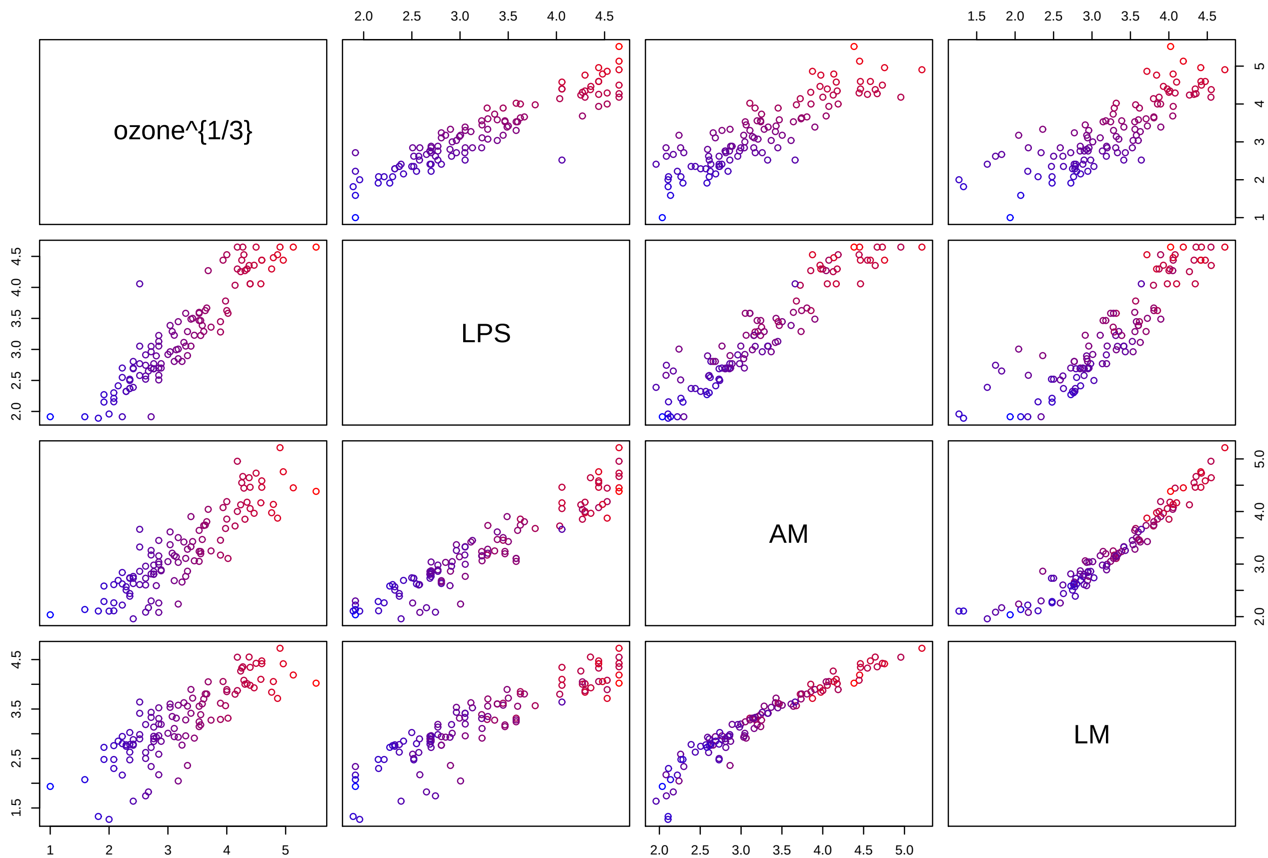
Air Quality Measurements (cont.)

- Definitely needs a surface rather than a curve!
- Special feature: Lots of missing values in the response (42 out of 153), but few missing predictors (only 7 out 153 rows).
- However, we can estimate the manifold using the complete 146 rows of the predictor space, and then use this manifold to predict the response.



Air Quality Measurements (cont.)

- True response, LPS-fitted ($\lambda = 1$), Additive Model (AM)-, and Linear Model (LM)- fitted values:



Remarks and Outlook

- The proposed techniques are neither thought to be "universal" nor "automatic", but may be useful in particular circumstances if there are strong nonlinear dependencies between the involved predictor variables.
- The technique unfolds its real power when considering predictor spaces of far higher dimension (for instance, spectral data).
- For high-dimensional predictor spaces a two-step strategy may be beneficial: Apply PCA on raw data, and approximate the scores through the manifold (Einbeck, Evers, & Powell, 2010).
- Retrospective post-processing (smoothing) of the manifold possible via the Elastic net algorithm (Gorban & Zinovyev, 2005).

Remarks and Outlook

- The proposed techniques are neither thought to be “universal” nor “automatic”, but may be useful in particular circumstances if there are strong nonlinear dependencies between the involved predictor variables.
- The technique unfolds its real power when considering predictor spaces of far higher dimension (for instance, spectral data).
- For high-dimensional predictor spaces a two-step strategy may be beneficial: Apply PCA on raw data, and approximate the scores through the manifold (Einbeck, Evers, & Powell, 2010).
- Retrospective post-processing (smoothing) of the manifold possible via the Elastic net algorithm (Gorban & Zinovyev, 2005).
- Desirable:
  - Smoothing "within" the triangle (or simplex).
  - More "Statistics"...

References

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

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

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.

LPCM: Local principal curves and manifolds. R package version 0.38-1, `http://www.maths.dur.ac.uk/~dma0je/software.html`.