Dimension reduction through local principal curves and manifolds

Jochen Einbeck

Durham University

jochen.einbeck@durham.ac.uk

Ambleside, 1st of September 2009

joint work with
Ludger Evers (University of Glasgow),

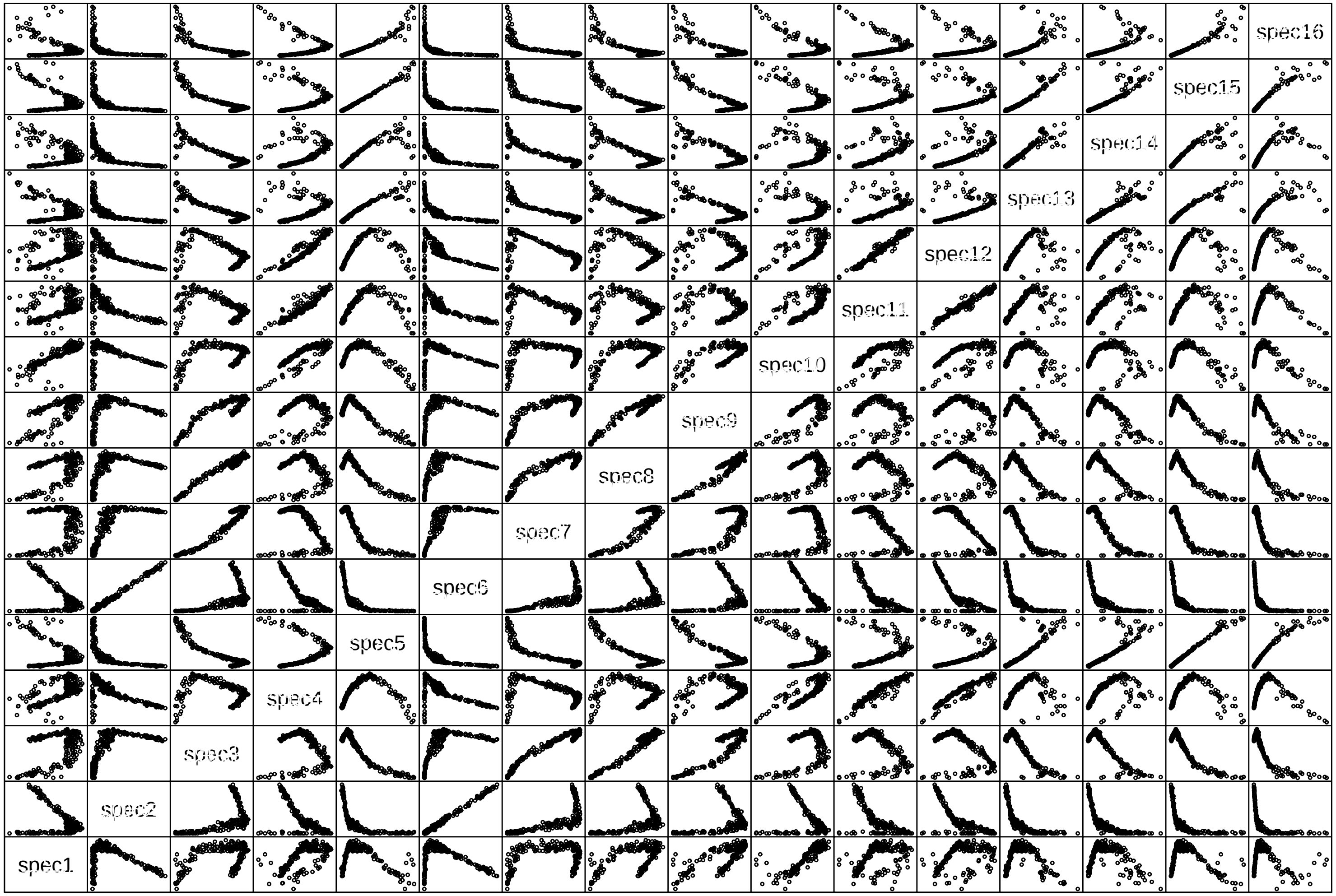
in collaboration with Coryn Bailer-Jones (MPIA Heidelberg).

Motivation: GAIA data

- GAIA is an astrophysics mission of the European Space Agency (ESA) which will undertake a detailed survey of over 10^9 stars in our Galaxy and extragalactic objects.
- Satellite to be launched in 2011.
- Aims of the mission (among others)
    - Classify objects (star, galaxy, quasar,...)
    - Determine astrophysical parameters ("APs": temperature, metallicity, gravity) from spectroscopic data (photon counts at certain wavelengths).
- Work is led by the group "Astrophysical parameters" based at MPIA Heidelberg, being part of the DPAC (Data Processing and Analysis Consortium) which is responsible for the general handling of data from the GAIA mission.
- Yet, one has to work with simulated data generated through complex computer models.

GAIA data

- Photon counts $(n = 8286)$ simulated from APs:



GAIA data: Estimation of APs

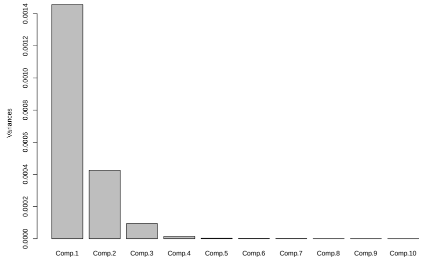
- For the actual estimation problem, the photon counts form the *predictor space* and the AP's form the *response space* (this is opposite to the direction of simulation!)
- Hence, the regression problem may be degenerate (i.e., one set of photon counts may be associated to two different APs).
- Try linear model for the temperature, using training sample of size $n = 1000$:

```
> lm(temperature˜ spec1 +...+ spec16, data= gaia)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -14033286   21104764  -0.665    0.506
spec1        14065842   21104812   0.666    0.505
.                   .          .       .        .
spec16       13886697   21106076   0.658    0.511
Residual standard error: 1978 on 983 degrees of freedom
```

Does not seem to be a useful model for this data.

Dimension reduction

- Usual remedies:
  - Model/ variable selection procedures
  - Dimension reduction techniques
- Look at scree plot:



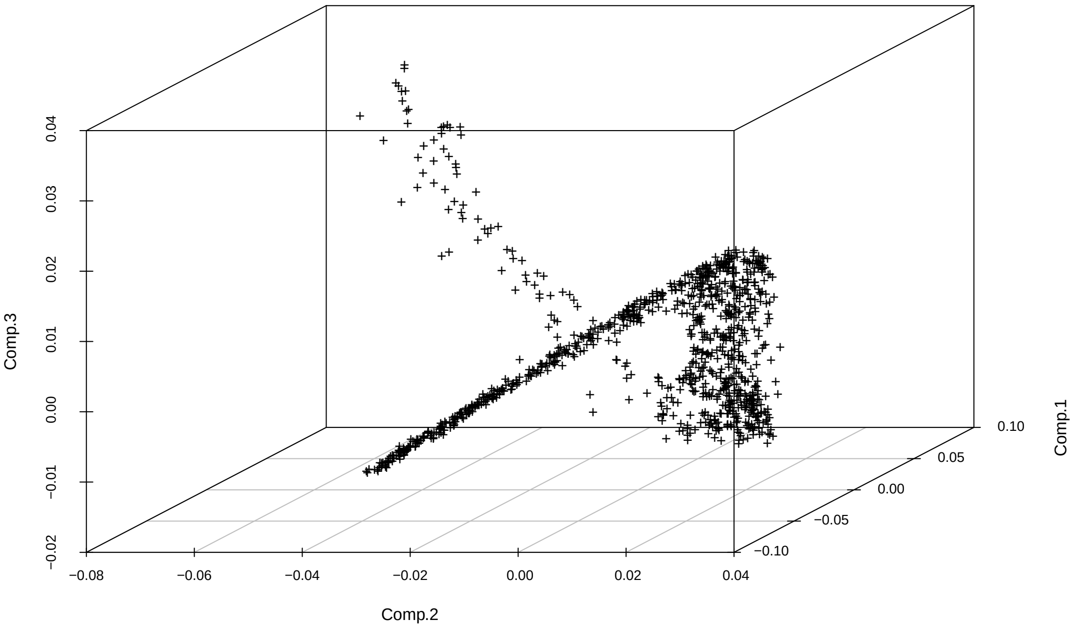
- Three principal components appear to be sufficient.

```
> lm(temperature ˜ Comp1 + Comp2 + Comp3, data = gaiapc)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   10835.90      65.14  166.34   <2e-16 ***
Comp1       -187339.39    1706.85 -109.76   <2e-16 ***
Comp2       -173967.35    3157.61  -55.09   <2e-16 ***
Comp3       -155314.86    6726.19  -23.09   <2e-16 ***
Residual standard error: 2060 on 996 degrees of freedom
```

looks better than LM, but...

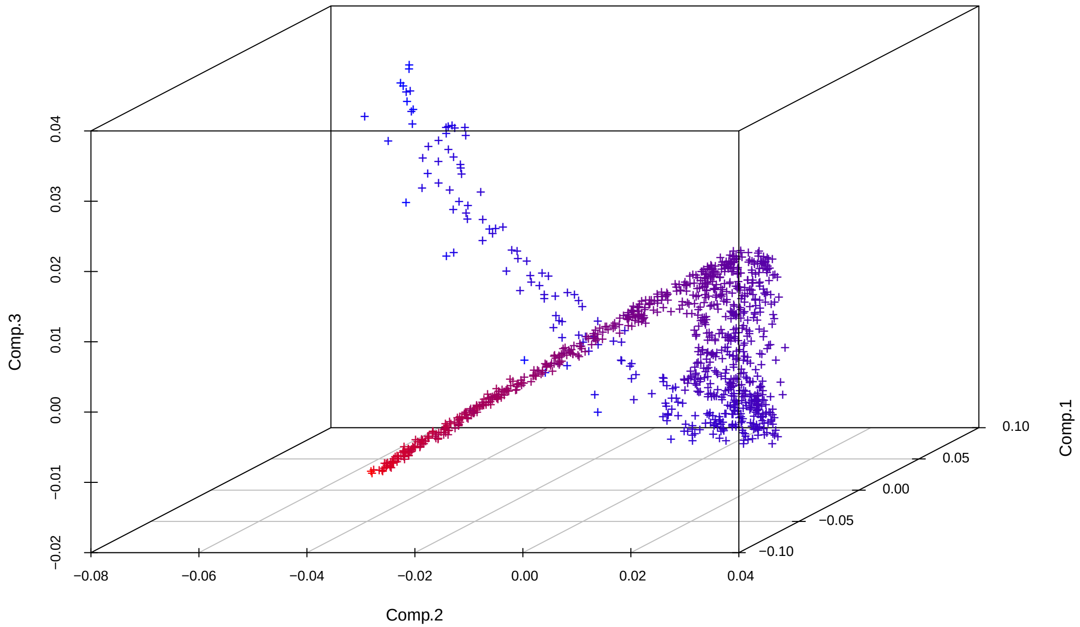
Principal component scores

- We plot the the first three principal component scores.



Principal component scores (cont.)

- We plot the the first three principal component scores and shade higher temperatures red.



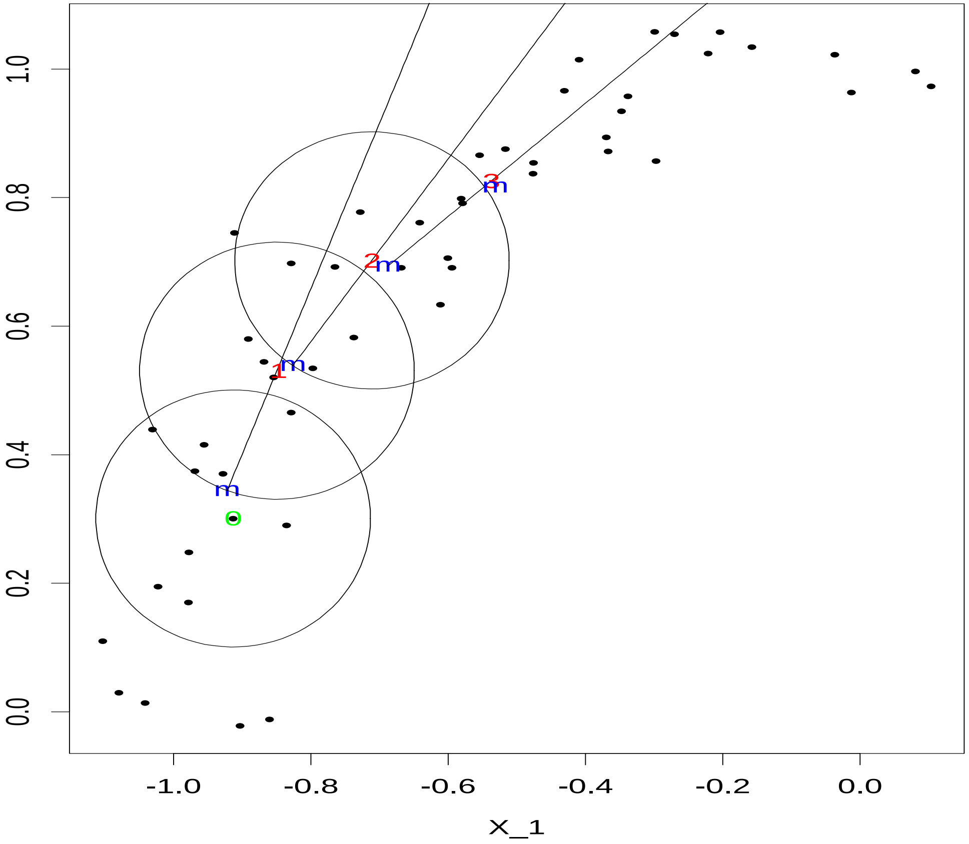
- Actually, we seem to need only *one* parameter if we were able to lay a smooth curve through the data cloud.
- The parametrization along such a curve would be informative w.r.t. to the target variable, temperature.

GAIA data and principal curves

- Hence, the following is to do:
  1. (1) Estimate the smooth curve capturing the structure of the (3-dim/16-dim) predictor space.
  2. (2) Parametrize this curve.
  3. (3) Project all data points onto it.
  4. (4) Fit temperature (or other APs) against the (1-dim.) projections.
- Step (1) is a task for principal curves. There are a couple of principal curve algorithms available, but not all of them are suitable for tasks (2)-(4).
- We concentrate here on local principal curves (LPC, Einbeck, Tutz & Evers, 2005).

Step 1: Fitting the LPC

- Idea: Beginning at some starting point, calculate alternately a local center of mass and a first local principal component.

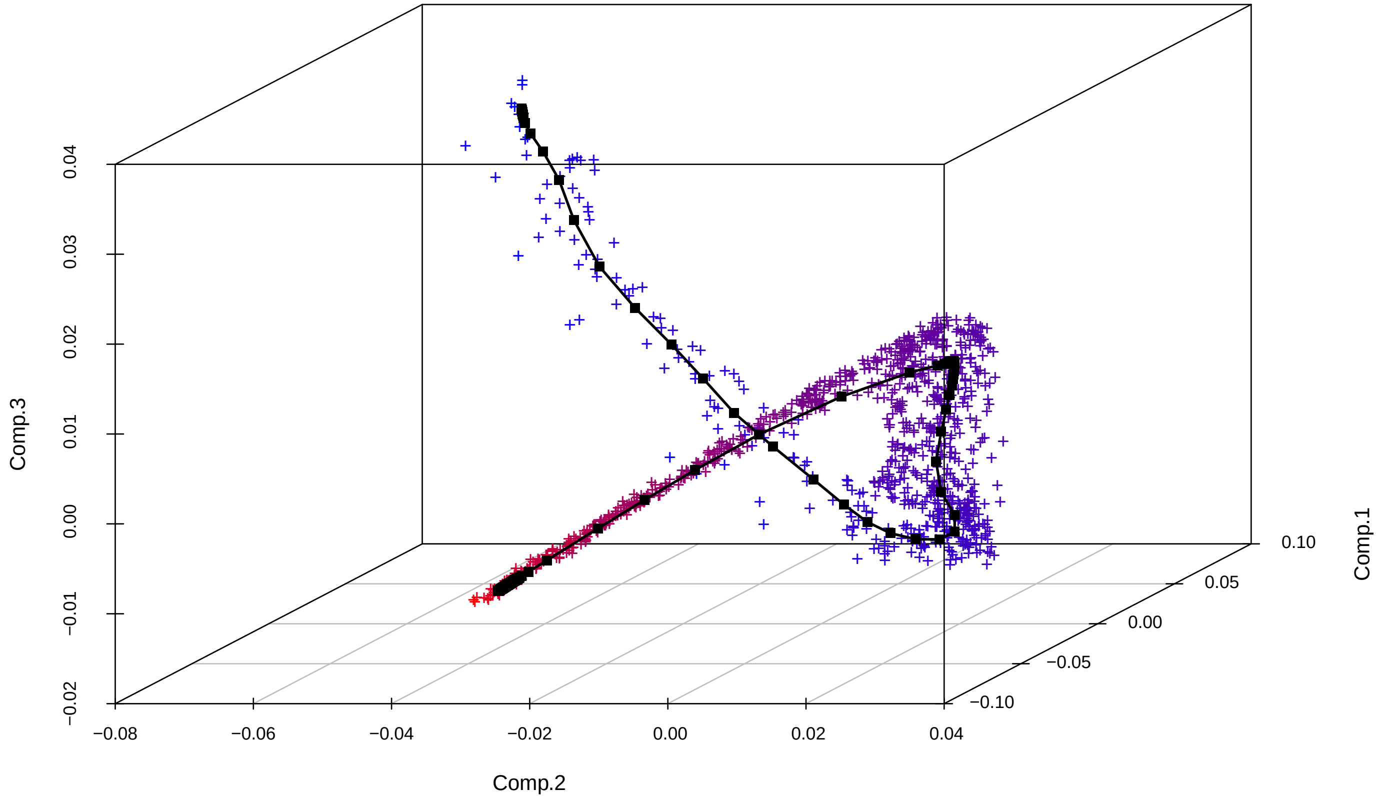


0: starting point,
m: points of the LPC,
$1, 2, 3$: enumeration of steps.

Step 1: Fitting the LPC (cont.)

- LPC through principal component scores of photon counts, with local centers of mass μ (■) :

```
> gaia.lpc <- lpc(gaia.pc$scores)
```

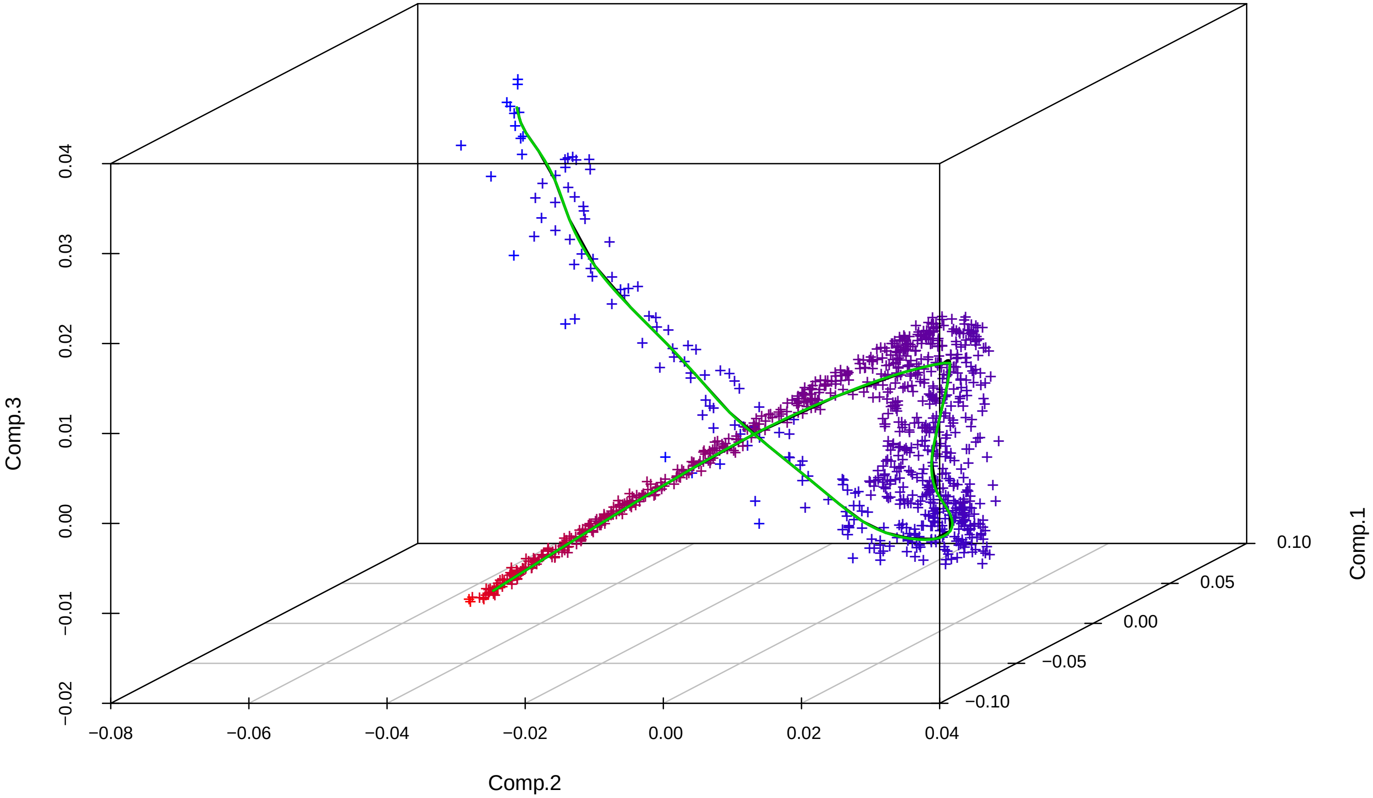


Step 2: Parametrization

- Unlike HS curves, LPCs do not have a natural parametrization, so it has to be computed retrospectively.
- Define a preliminary parametrization $s \in \mathbb{R}$ based on Euclidean distances between neighboring $\mu \in \mathbb{R}^d$.
- For each component μ_j, $j = 1, \ldots, d$, use a natural cubic spline to construct functions $\mu_j(s)$, yielding together a function $(\mu_1, \ldots, \mu_d)(s)$ representing the LPC (no smoothing involved here!).
- Recalculate the parametrization along the curve through the arc length of the spline function.

Step 2: Parametrization (cont.)

```
> lpc.spline(gaia.lpc)
```

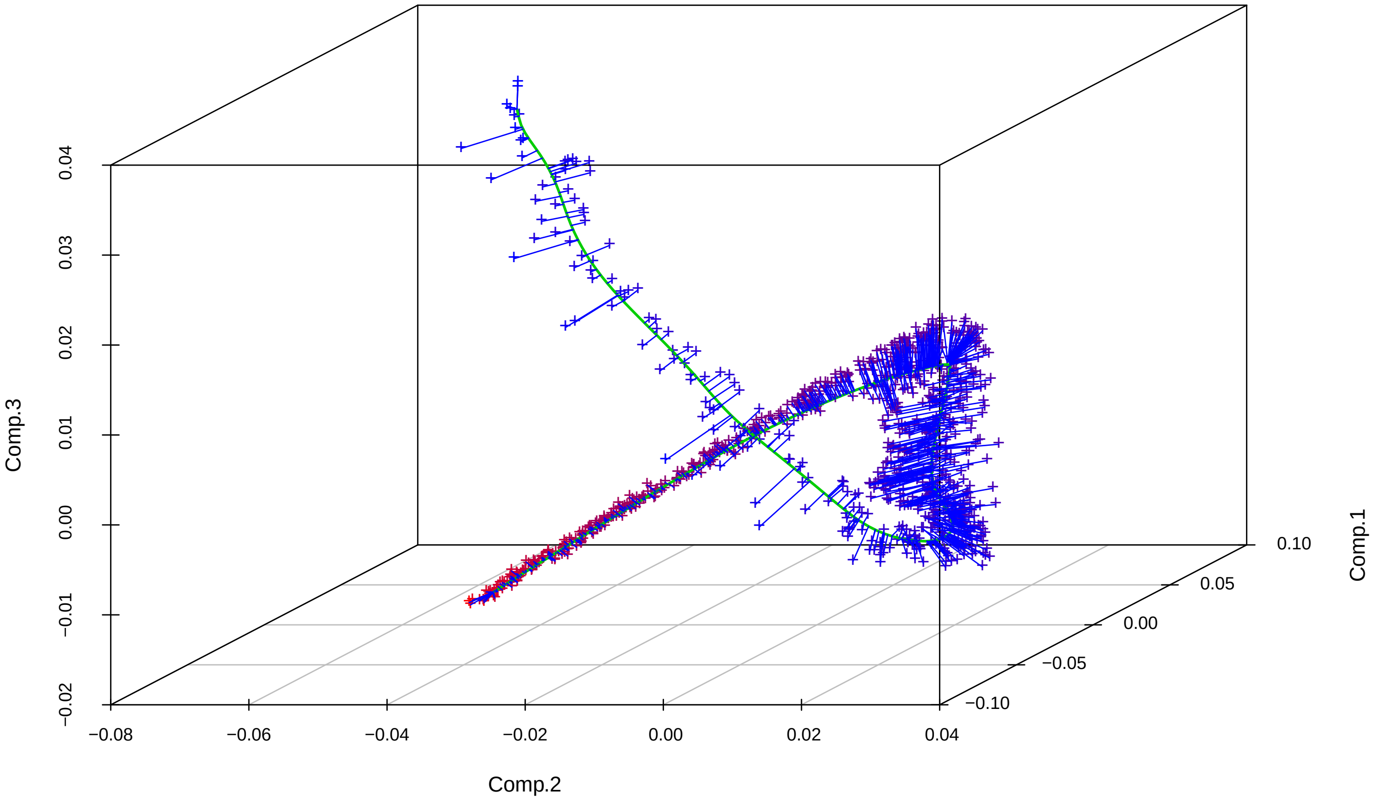


- The spline function (—) is almost indistinguishable from the original LPC (—).

Step 3: Projection

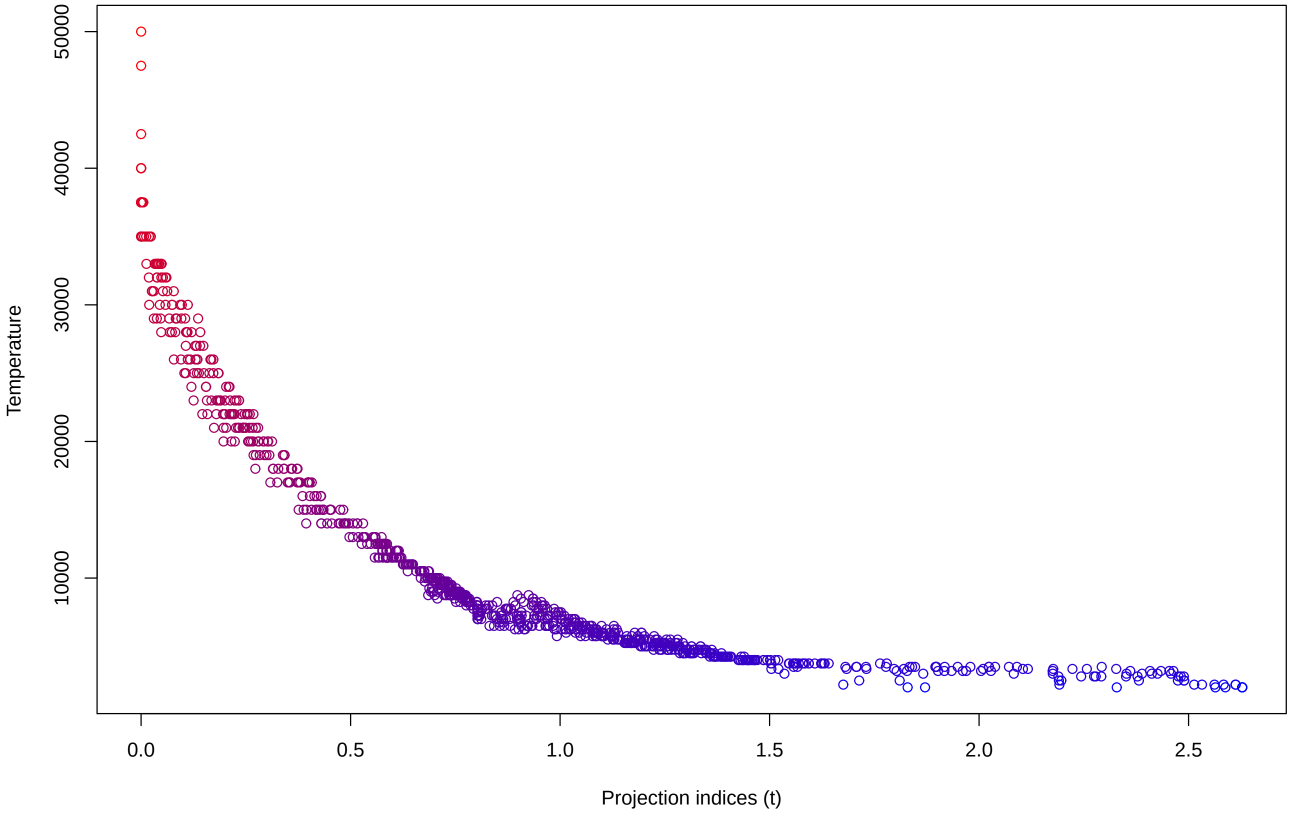
- Each point $x_i \in \mathbb{R}^d$ is projected on the point of the curve nearest to it, yielding the corresponding projection index t_i

```
> lpc.spline(gaia.lpc, project=TRUE)
```



Step 4: Regression

- We want to predict stellar temperature from 16-d spectral data, using the projection indices of the spectra as predictors.

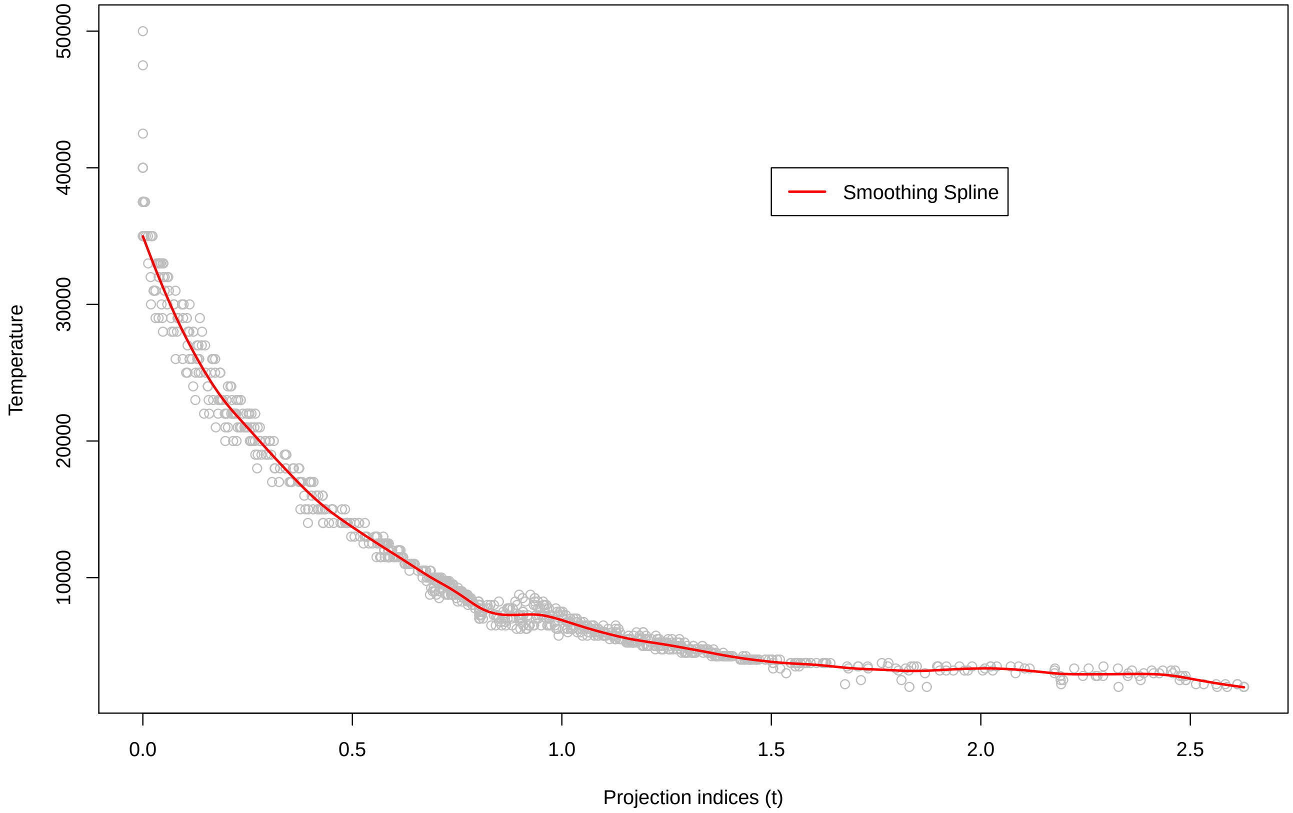


Step 4: Regression (cont.)

- This is now a simple one-dimensional regression problem.

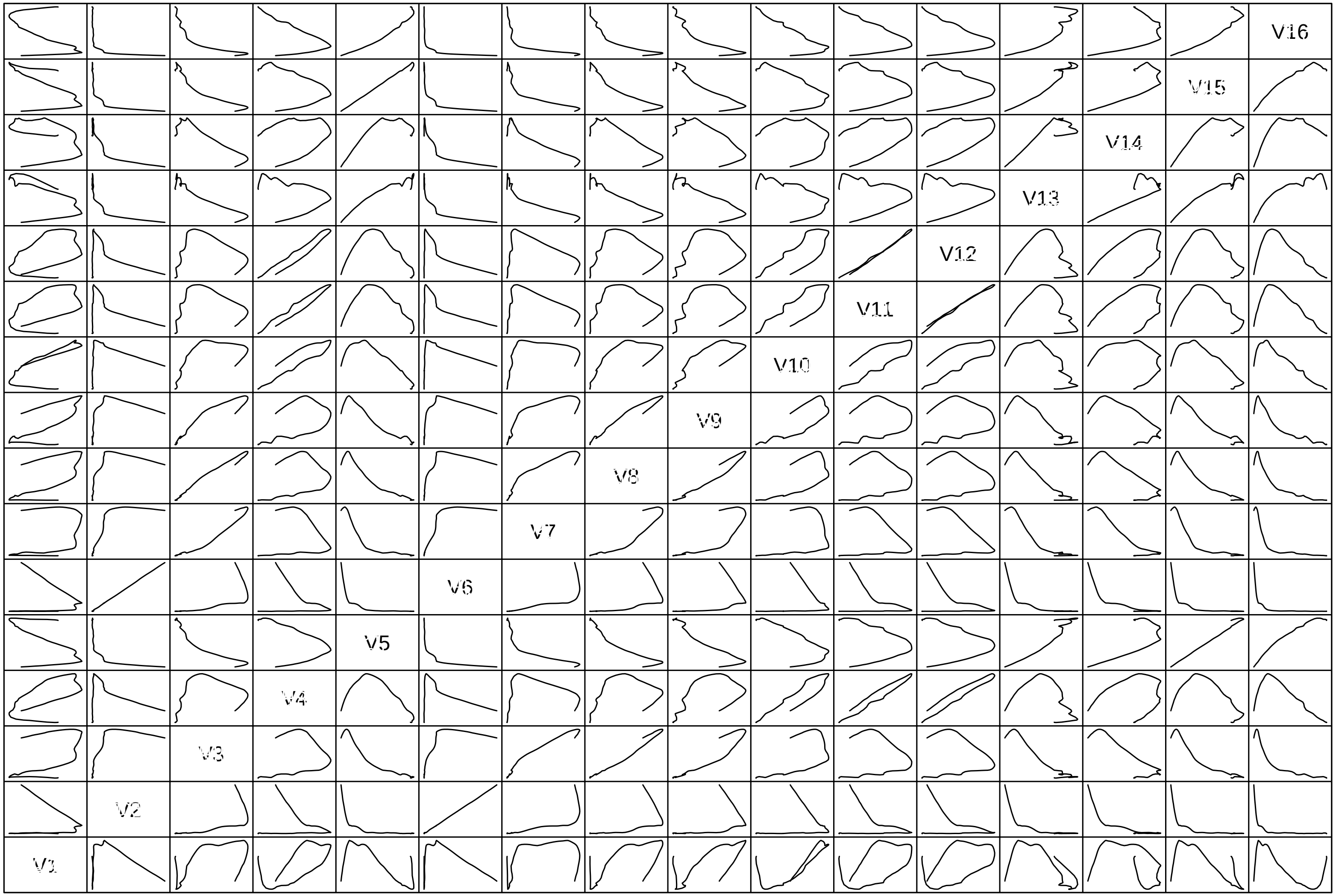
$$y_i = m(t_i) + \varepsilon_i$$

- Using penalized smoothing splines:



Shortcut

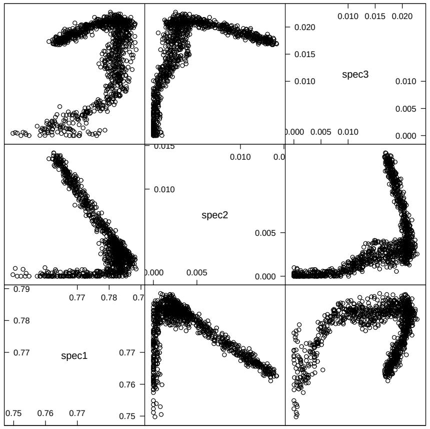
- LPC fitted *directly* through 16- dimensional space:



Direct data compression with LPCs

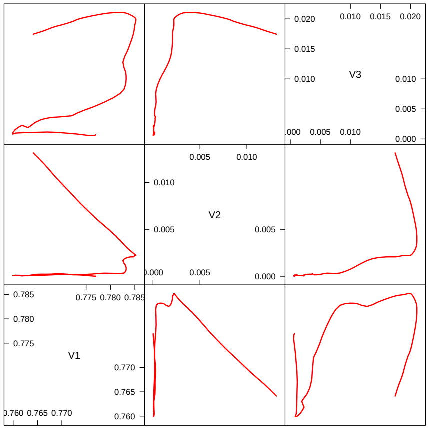
- Zoom into the the first three dimensions:

Data



Scatter Plot Matrix

LPC



Scatter Plot Matrix

Prediction

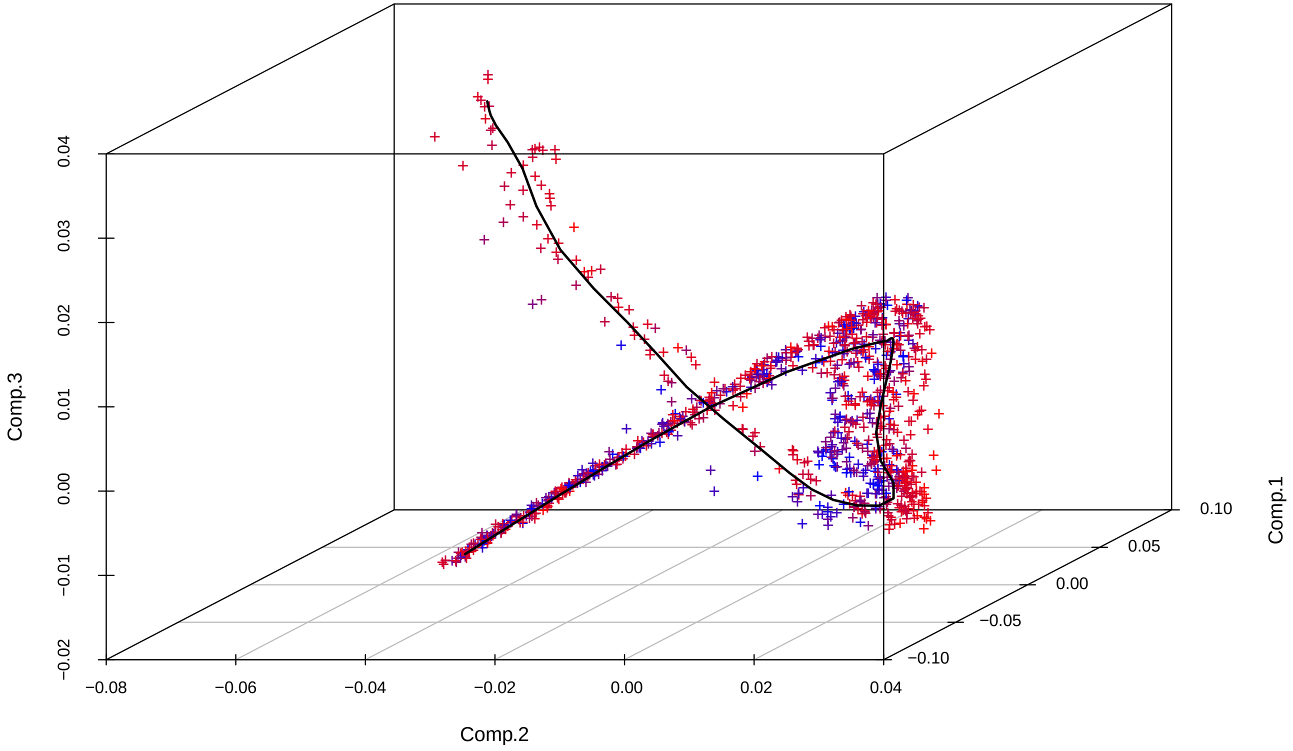
- For a new observation x_{new} (i.e., here, a new set of spectra), prediction proceeds as follows:
  - Project x_{new} onto the LPC, giving t_{new}.
  - Compute $\hat{y}_{new} = \hat{m}(t_{new})$ from the fitted regression model.
- Comparison: We sample $n' = 1000$ test data from the remaining $8286 - 1000$ observations and observe the prediction error:

	LM	PC+LM	PC+AM	PC+LPC	LPC (2nd run)
average($\hat{\varepsilon}_i^2$)	4'593	4'967	1'732	1'447	1'044 (2'025)
median($\hat{\varepsilon}_i^2$)	1'049	1'124	104	51	69 (71)

where $\hat{\varepsilon}_i$ is the difference between true and predicted temperature.

Limits of one-dimensional data summaries

- Look at "metallicity"



- The relevant information seems to be orthogonal to the principal curve!

Local principal surfaces and manifolds

- To handle this and more complex data, the extension to *local principal surfaces* and *manifolds* should be considered.
- To this end, firstly observe that, from the two components of the LPC algorithm, namely

  (1) local center of mass (mean shift)

  (2) localized first principal component

  the more important is (rather surprisingly) (1).
- Instead of (2), any other movement "roughly in the direction of the data cloud" can be made, and step (1) will shift it back to the data cloud.
- We exploit this observation for the extension to local principal surfaces.

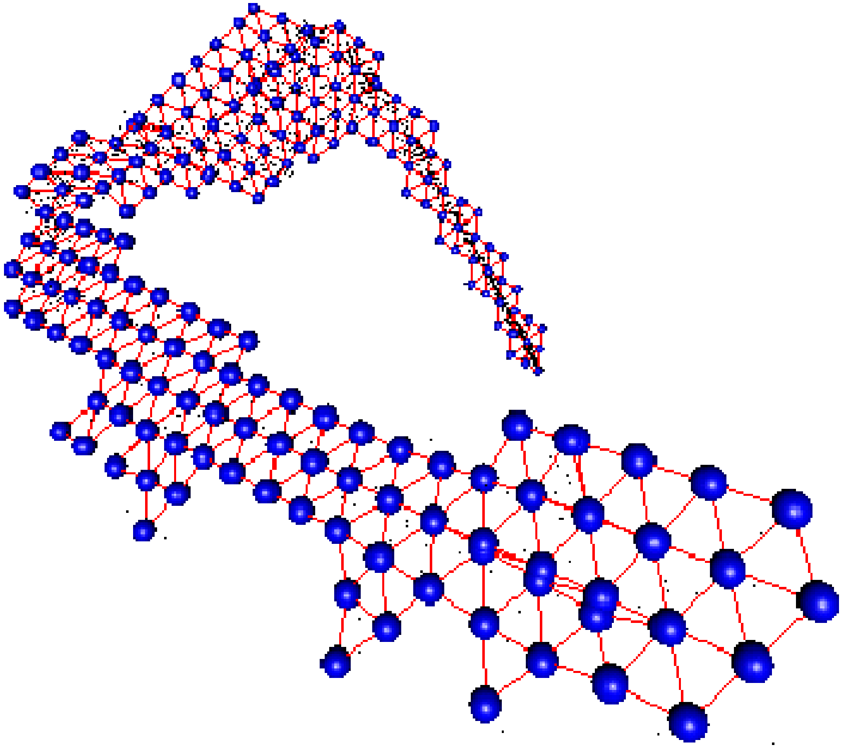
Local principal surfaces and manifolds (cont.)

- Instead of points x, we work 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.
- Straightforward extension to local principal manifolds (LPMs) of higher dimensions by using tetrahedrons instead of triangles.

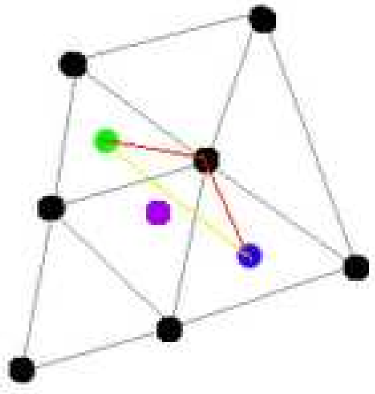
Local principal surface for GAIA data

- Local principal surface for PC scores based on training data set with $n = 1000$:



Regression with LPMs

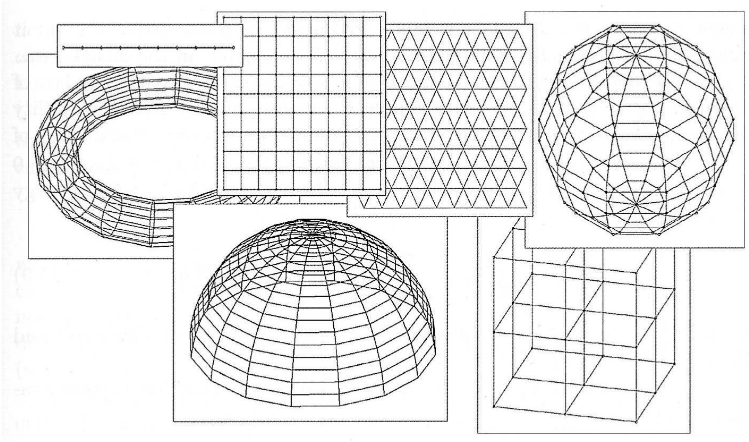
- The next question is how to parametrize the manifold in order to use it for regression.
- Actually, we do not need a full parametrization, just *distances*.
- What is the distance between the green and the blue point?



  - MDS/ISOMAP ?
  - Local regression ?
- work in progress....

LPMs and the elastic net

- Gorban & Zinovyev (2005) developed the elastic net algorithm, based on a physical analogy with elastic membranes.
- Roughly, a graph structure is defined consisting of nodes, edges, and ribs, and an *elastic energy* is assigned to the graph depending on its complexity.
- The elastic net is that graph minimizing the sum of energy and approximation error.
- The elastic net algorithm requires to define some sort of "proposal net" of type:

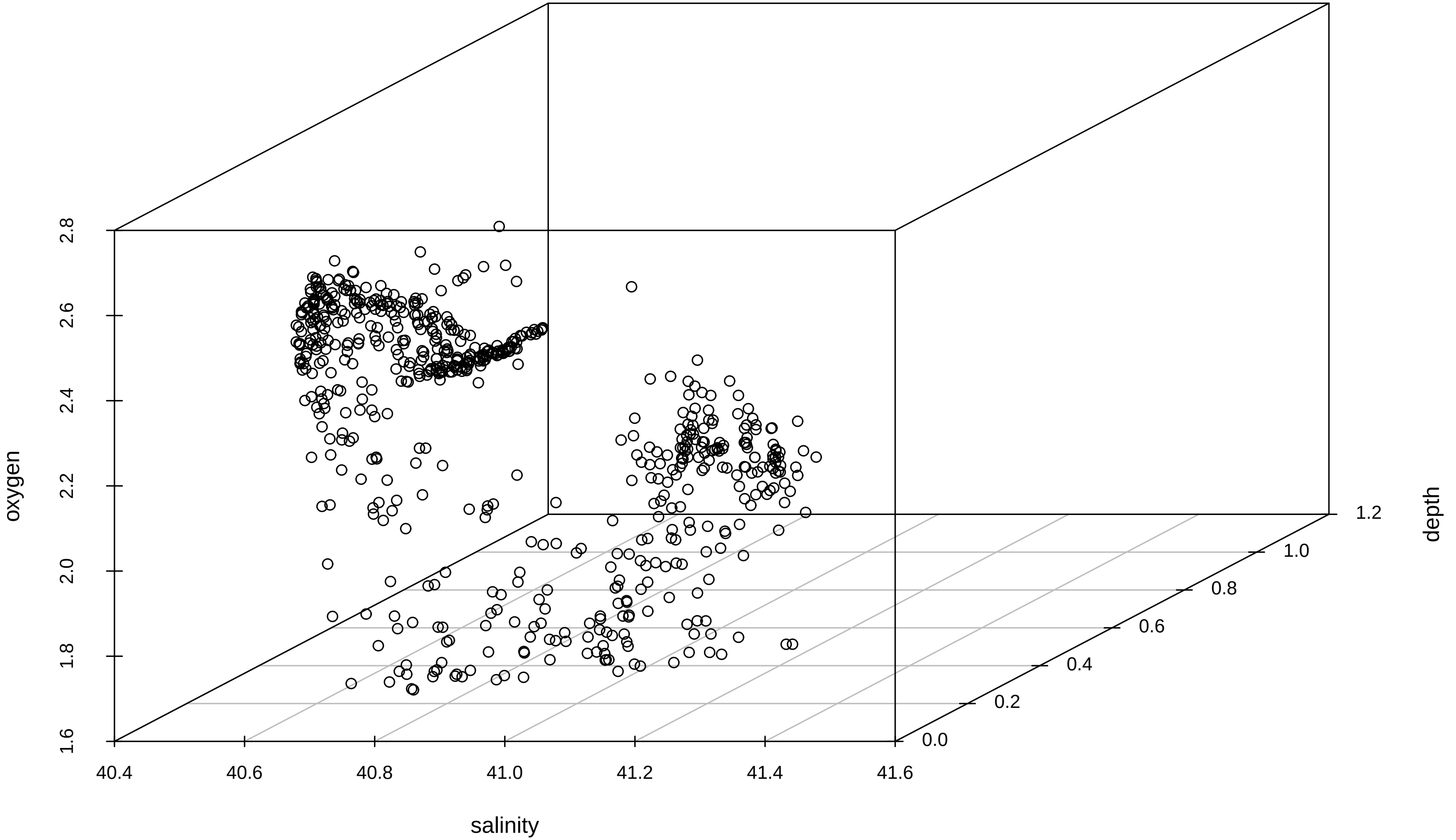


LPMs and the elastic net (cont.)

- This is where LPMs may be useful:
  - The elastic net gives a very smooth and stable estimate, which is, however, based on some initial proposal net which may be difficult to specify if the structure of the data could is unknown.
  - LPMs, in turn, are very flexible, do not need prior knowledge, but can be quite unsmooth and unstable (in the sense that they depend on some initial starting point/ triangle).
- Idea: Why not use the LPC/LPM as the proposal net for the elastic net?

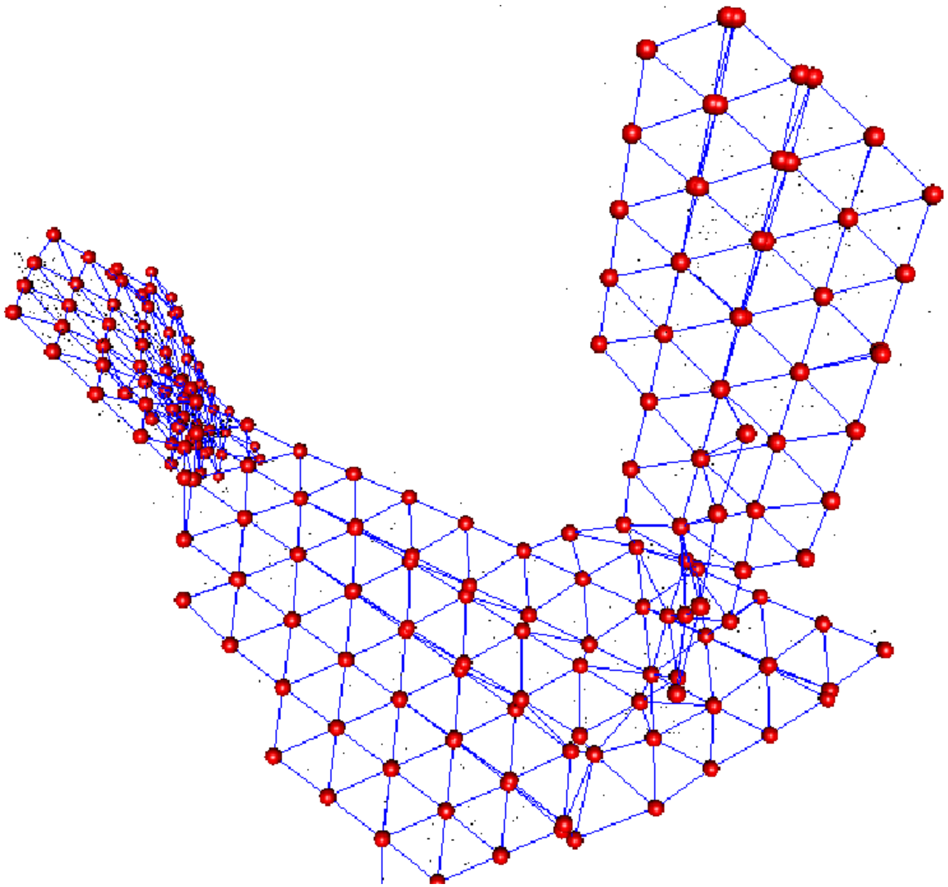
LPMs and the elastic net (cont.)

- For illustration: Oceanographic data taken by the German research vessel “Gauss” in May 2000 southwest of Ireland:



LPMs and the elastic net (cont.)

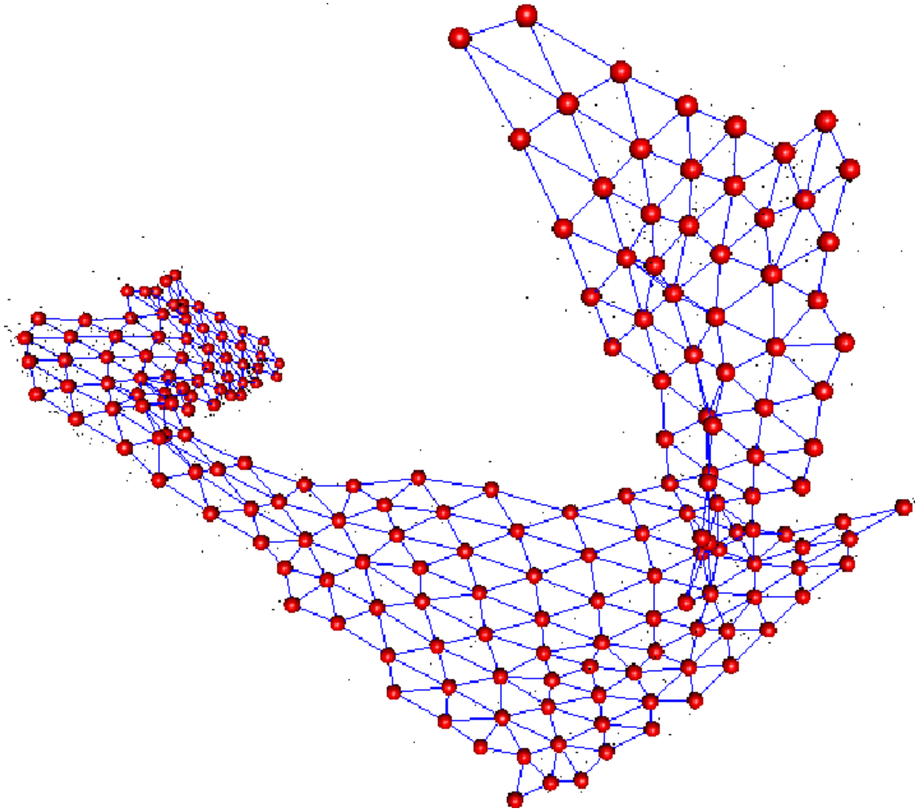
- Fit a local principal manifold:



- looks a little bit unsmooth....

LPMs and the elastic net (cont.)

- Postprocess via elastic net:



- Final result neater than the LPM!

Conclusion

- After parametrization through cubic splines, LPCs can be used for dimension reduction provided that
  - the intrinsic (topological) dimensionality of the data cloud is close to 1, or, at least,
  - the projections on the curve are informative for the target variable.
- Extension of LPCs to LPMs works by considering the building block “triangles”.
- Regression against LPMs still to do!
- LPMs may be useful as “proposal manifold” for the elastic net algorithm.
- R package “LPCM” in development, available on request from authors.

Literature

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

Einbeck, Evers & Bailer-Jones (2008): Representing complex data using localized principal components with application to astronomical data. In: Gorban, Kegl, Wunsch, & Zinovyev: Principal Manifolds for Data Visualization and Dimension Reduction; *Lecture Notes in Computational Science and Engineering* **58**, 180–204.

Einbeck & Evers (2009): **LPCM** – Local principal curves and manifolds (R package version 0.36, available from authors).

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

Powell (2009): An Introduction to Smoothers. *4H Project Report, Durham University*.