

---

# Data compression and regression based on local principal curves

Jochen Einbeck



jochen.einbeck@durham.ac.uk

*GfKI, Hamburg, 18th of July 2008*

joint work with  
Ludger Evers (University of Bristol),

in collaboration with Gerhard Tutz (LMU Munich) and Coryn Bailer-Jones (MPIA Heidelberg).

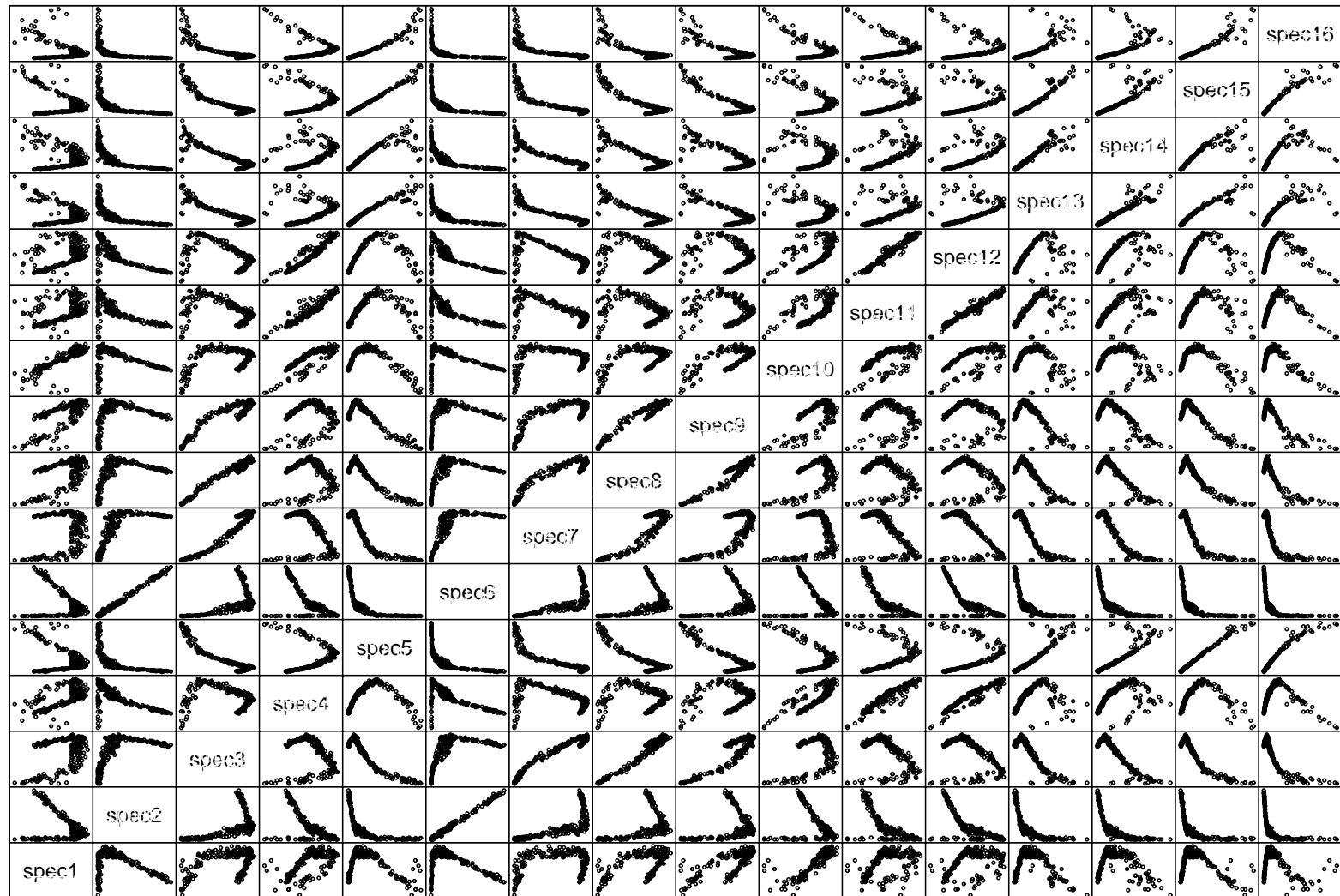
# Motivation: GAIA data

---

- 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 on these aims 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 simulated from APs through computer models:



# GAIA data: Estimation of APs

---

- Note that, 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!)
- As a consequence, the regression problem may be degenerate (i.e., one set of photon counts may be associated to two different APs). We focus here on the temperature, which features the least amount of degeneration.
- Can one use a linear model here?
  - Fit temperature against 16-dim photon counts ( $n = 1000$ ).

```
> gaia.lm <- lm(temperature ~ spec1 + spec2 +  
               ... + spec16, data= gaia)
```

# GAIA data: Estimation of APs (cont.)

---

```
> summary(gaia.lm)
```

```
Coefficients:
```

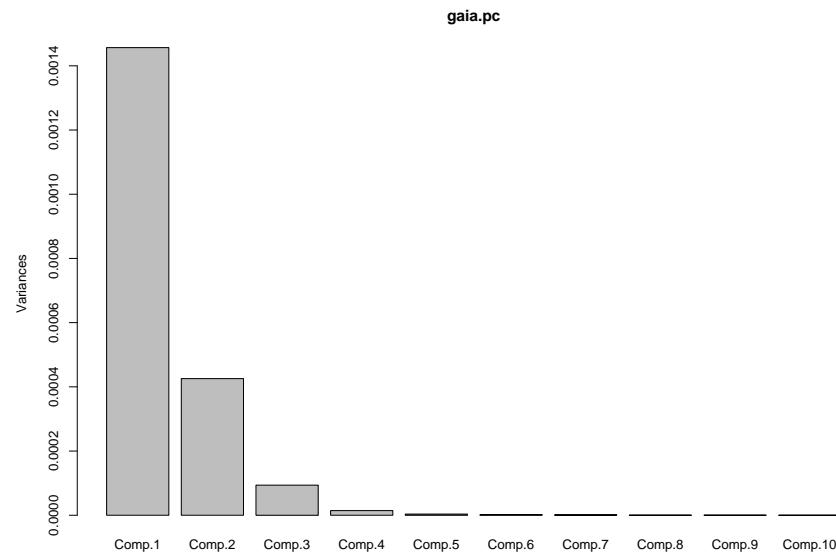
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-14033286	21104764	-0.665	0.506
spec1	14065842	21104812	0.666	0.505
spec2	14216977	21107526	0.674	0.501
spec3	13982281	21106961	0.662	0.508
spec4	13987405	21109664	0.663	0.508
.	.	.	.	.
.	.	.	.	.
spec16	13886697	21106076	0.658	0.511

```
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
- The second one is obviously the more promising here.
- Look at scree plot:



- Two (or maximal three) components appear to be sufficient.

# Principal component regression

- Fit temperature against the first three PC scores:

```
gaia.pclm <- lm(temperature ~ Comp1 + Comp2 +  
               + Comp3, data = gaiapc)
```

Coefficients:

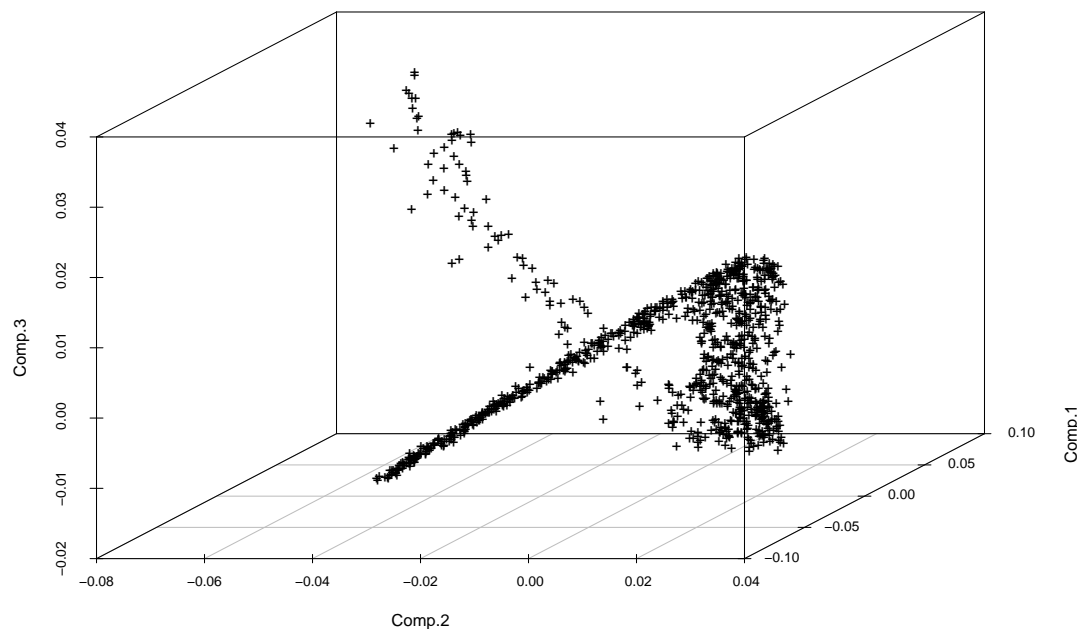
	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	10835.90	65.14	166.34	<2e-16	**
Comp1	-187339.39	1706.85	-109.76	<2e-16	**
Comp2	-173967.35	3157.61	-55.09	<2e-16	**
Comp3	-155314.86	6726.19	-23.09	<2e-16	**

Residual standard error: 2060 on 996 degrees of freedom

- This is somewhat more appropriate than the full linear model, but....

# Principal component scores

- We plot the the first three principal component scores.

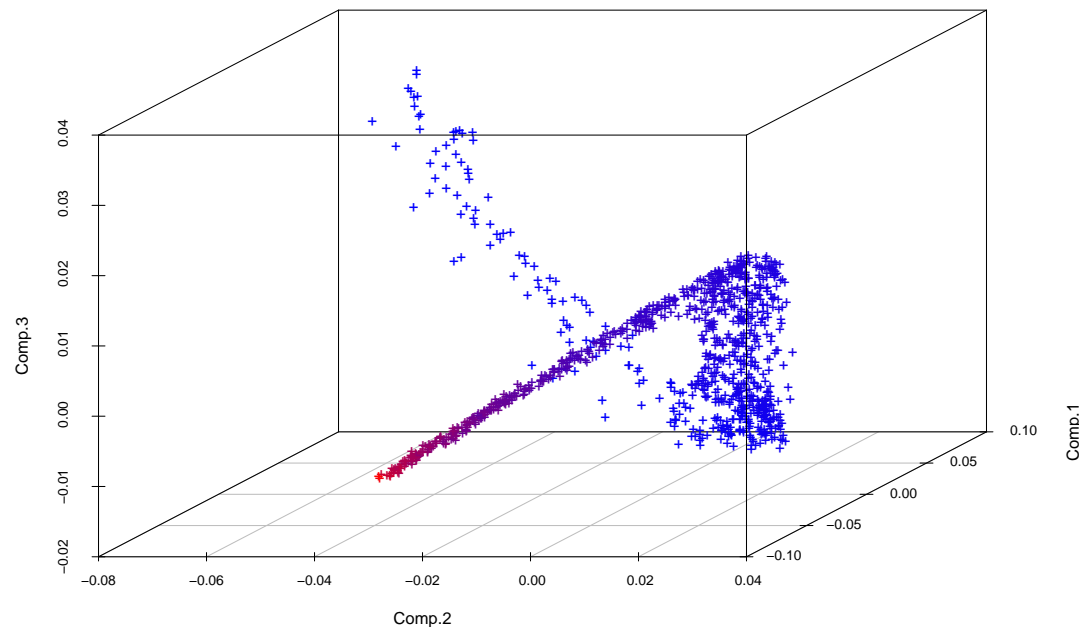


- Actually, we seem to need only *one* parameter if we were able to lay a smooth curve through the data cloud.



# Principal component scores

- 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 parameterization along such a curve would be informative w.r.t. to the target variable, temperature.

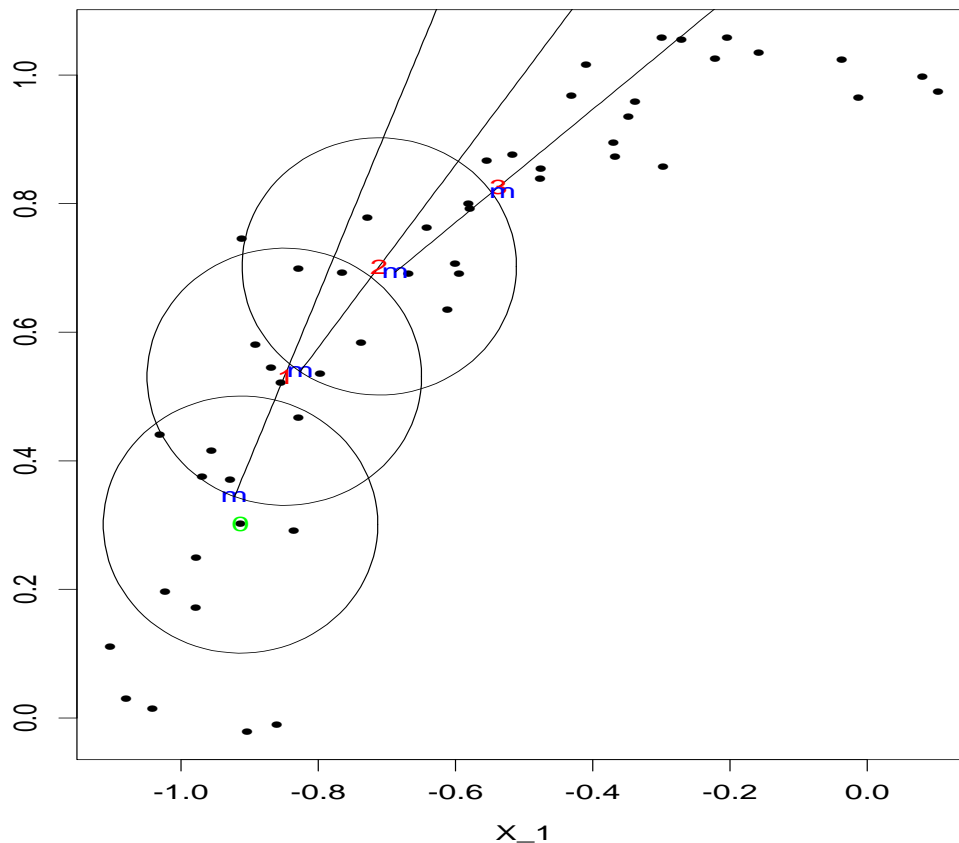
# GAI A data and principal curves

---

- Hence, the following is to do:
  - (1) Estimate the smooth curve capturing the structure of the (3-dim/16-dim) predictor space.
  - (2) Parameterize this curve and project all data points onto it.
  - (3) 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 task (2).
- We concentrate here on **local principal curves**.

# Local principal curves (LPCs)

- Einbeck, Tutz & Evers (2005)
- Idea: 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.

# Algorithm for LPCs

Given: A data cloud  $X = (X_1, \dots, X_n)$ , where  $X_i = (X_{i1}, \dots, X_{id})$ .

1. Choose a starting point  $x_0$ . Set  $x = x_0$ .
2. At  $x$ , calculate the local center of mass  $\mu^x = \sum_{i=1}^n w_i X_i$ , where  $w_i = K_H(X_i - x)X_i / \sum_{i=1}^n K_H(X_i - x)$ , with bandwidth matrix  $H$ .
3. Compute the 1<sup>st</sup> local eigenvector  $\gamma^x$  of  $\Sigma^x = (\sigma_{jk}^x)_{(1 \leq j, k \leq d)}$ , where
$$\sigma_{jk}^x = \sum_{i=1}^n w_i (X_{ij} - \mu_j^x)(X_{ik} - \mu_k^x).$$
4. Step from  $\mu^x$  to  $x := \mu^x + t_0 \gamma_1^x$ .
5. Repeat steps 2. to 4. until the  $\mu^x$  remain constant. Then set  $x = x_0$ , set  $\gamma^x := -\gamma^x$  and continue with 4.

The sequence of the local centers of mass  $\mu^x$  makes up the local principal curve (LPC).

# Parametrization

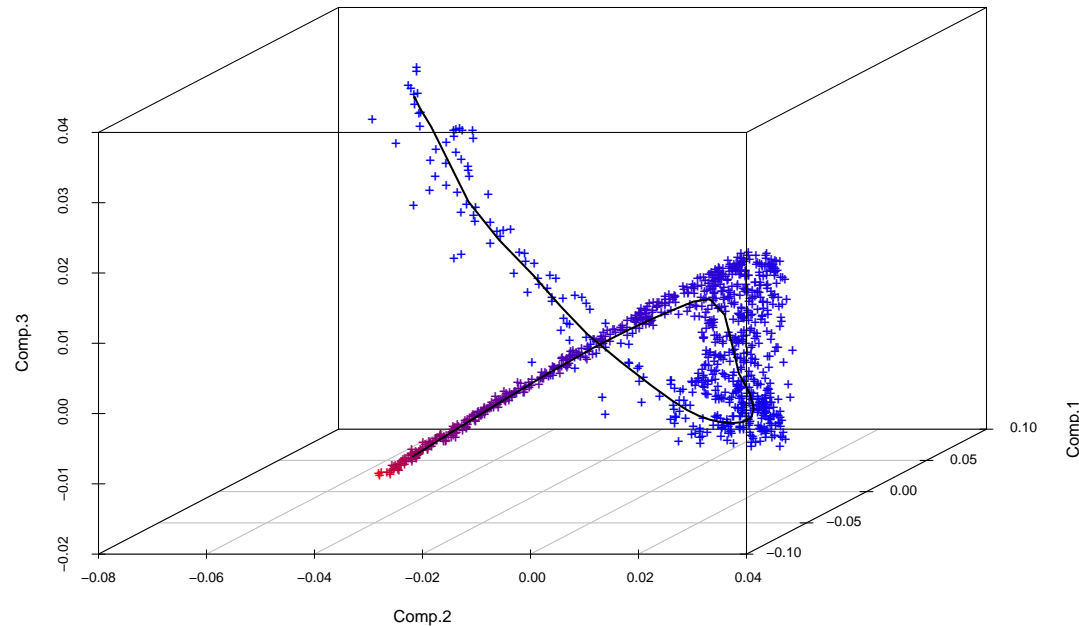
- After the LPC is fitted, a cubic spline is laid through the points  $\mu^x$  and the origin is fixed to  $t = 0$  at one of the two ends. E.g., for  $d = 2$ , we consider the LPC as a curve

$$f : \mathbb{R} \longrightarrow \mathbb{R}^2, t \mapsto \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}$$

and the parametrization along the curve is defined through the arc length of the spline w.r.t. this origin.

- Each point  $(x_i, y_i)$  can then be projected on the point of the curve nearest to it, yielding the corresponding projection index  $t_i$ , which could, for example, be used as one-dimensional summary of the predictor space in a multivariate regression problem.

# LPCs through GAIA data

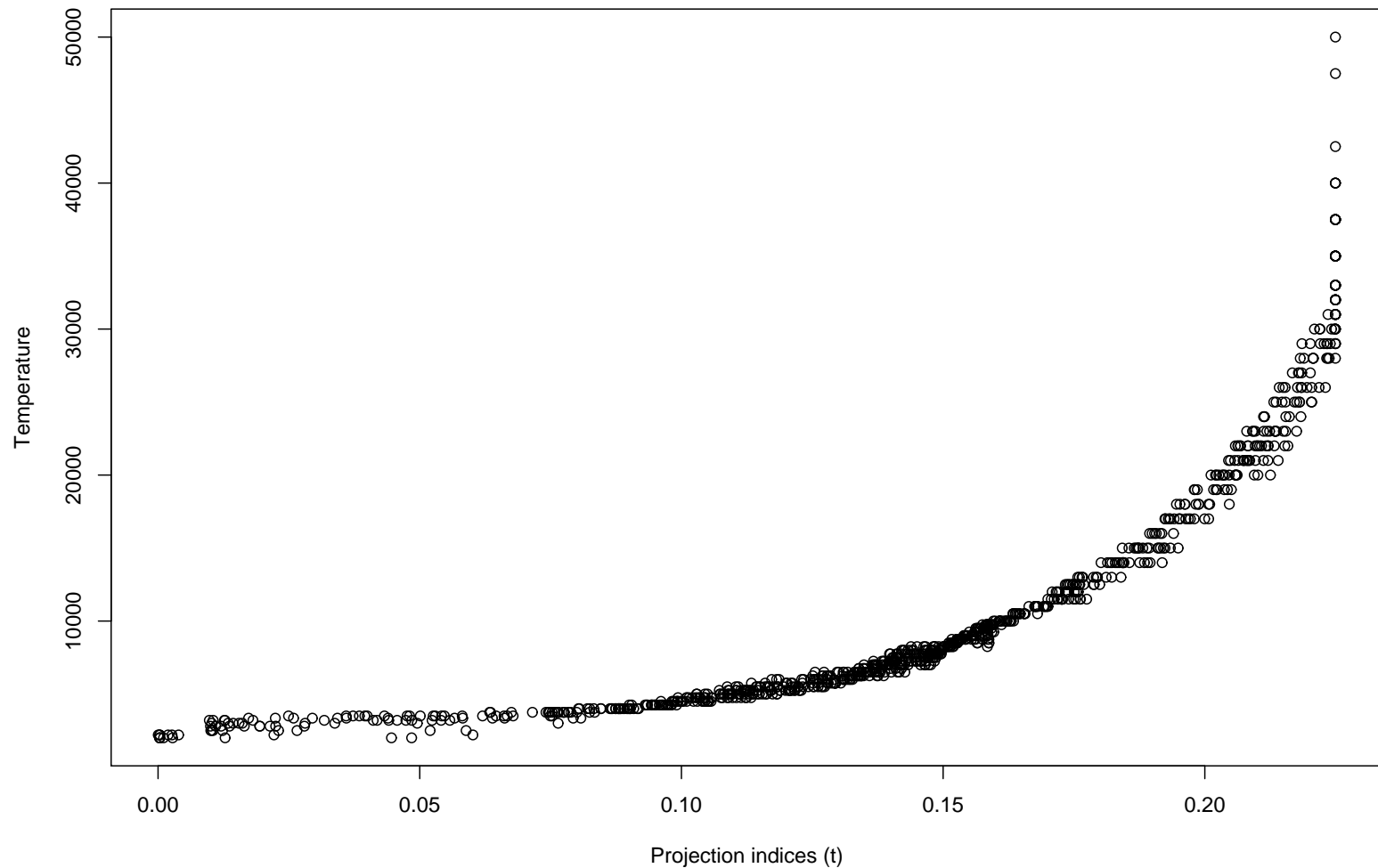


- Local principal curves seem to capture the structure of the data cloud adequately.
- We want to predict stellar temperature from spectral data, using the projection indices of the spectra as predictors.

# LPC Regression with GAIA data

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

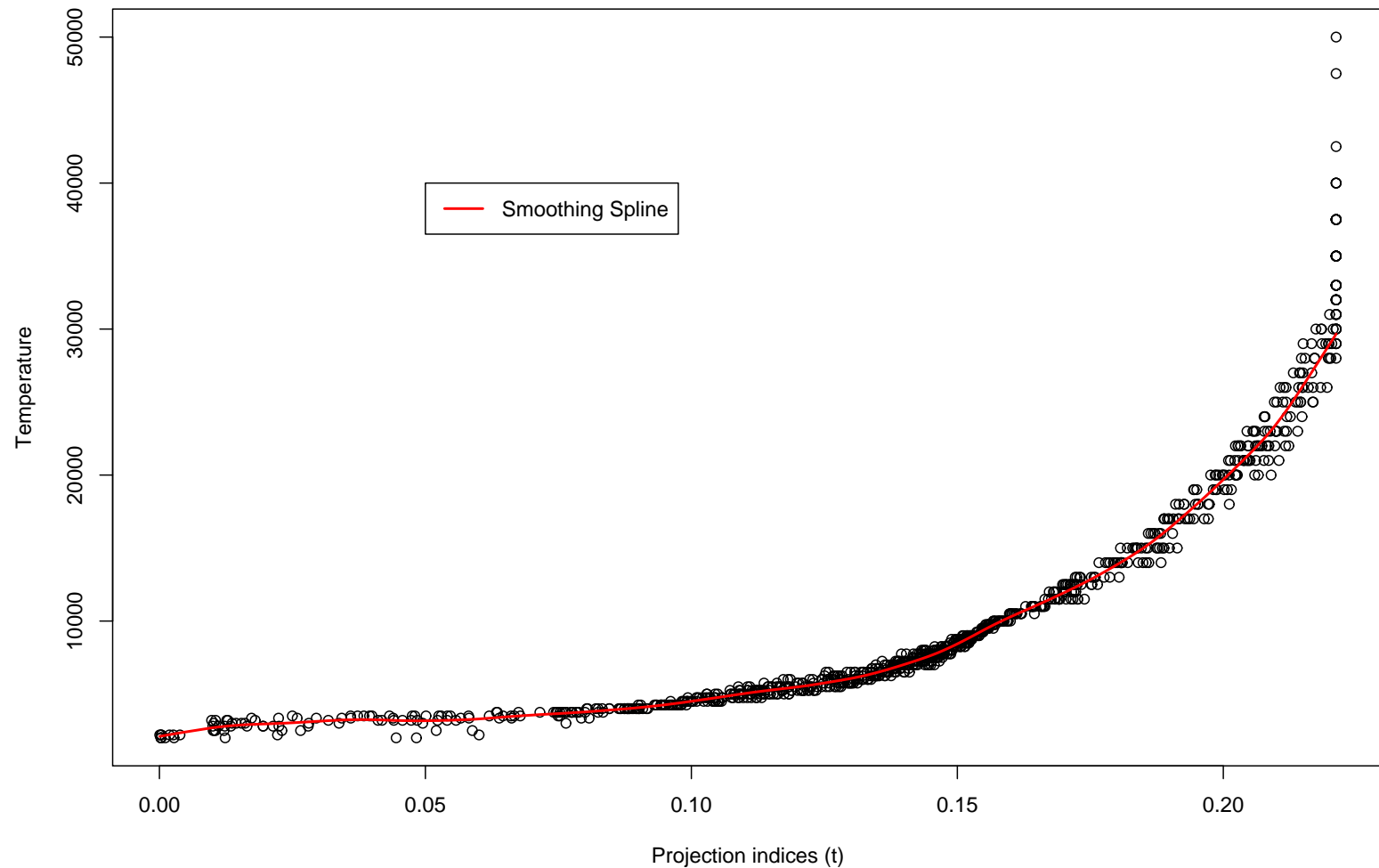
$$y_i = f(t_i) + \varepsilon_i :$$



# LPC Regression with GAIA data

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

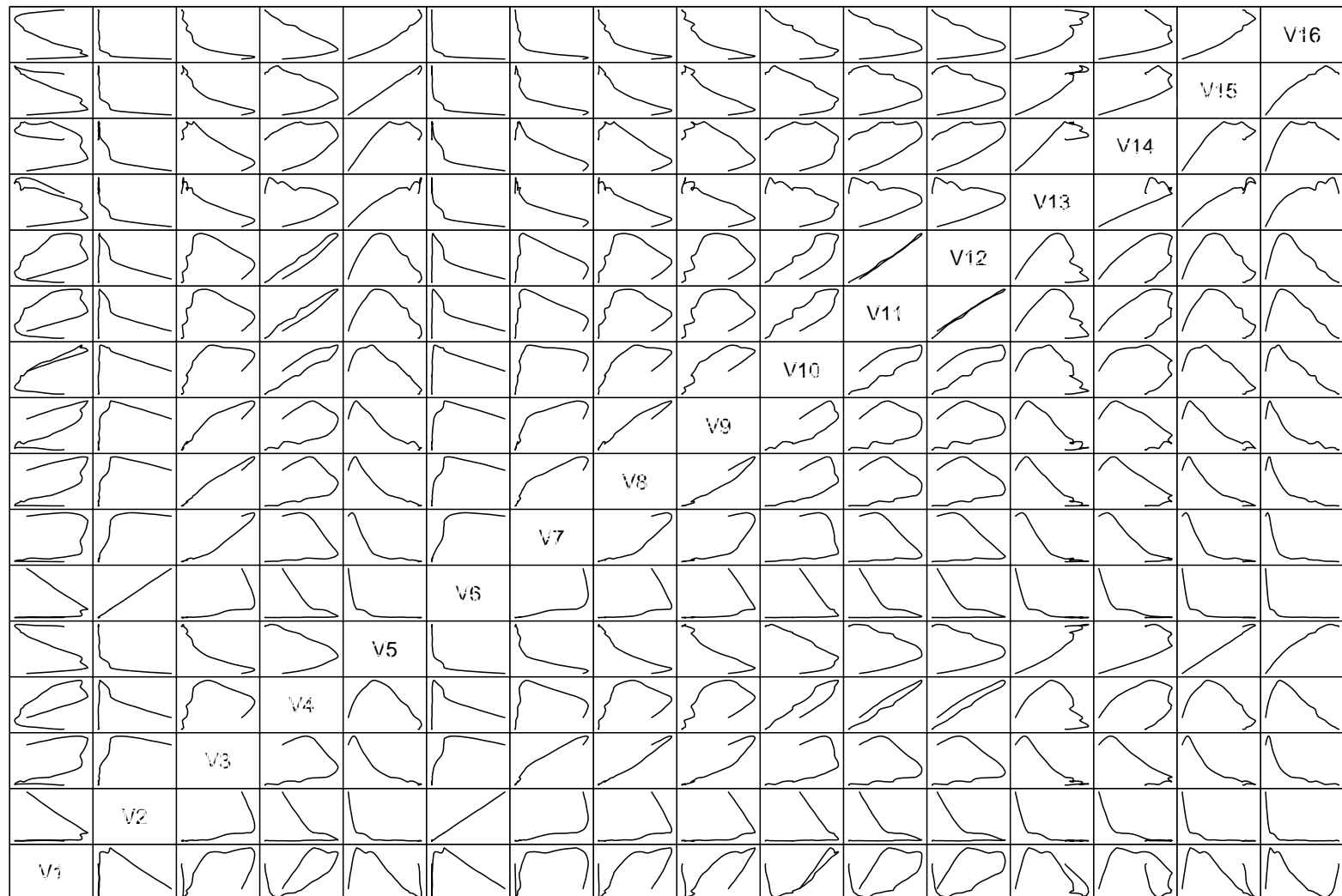
$$y_i = f(t_i) + \varepsilon_i :$$





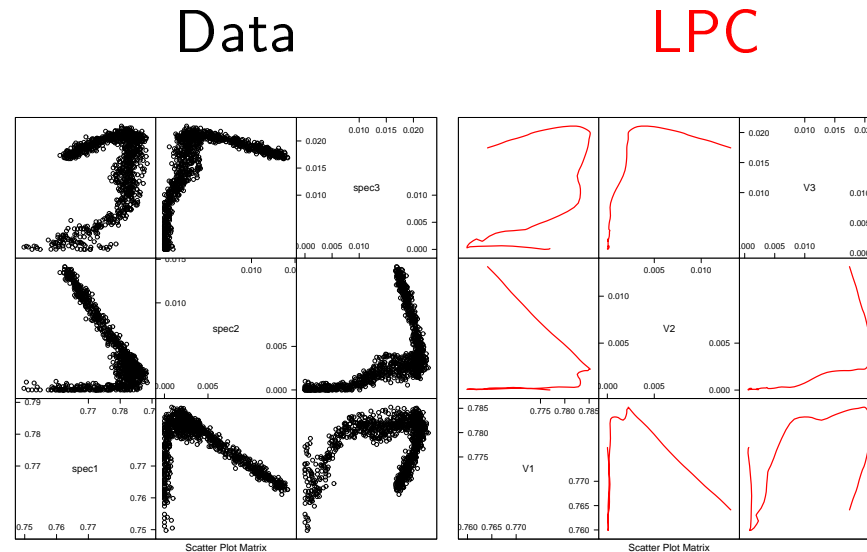
# Shortcut

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



# Direct data compression with LPCs

- Zoom into the the first three dimensions:



- Approximating the data cloud directly by a LPC works in principle, but is potentially “dangerous”: As data gets sparse in high dimensions, the LPC may miss remote parts of the predictor space and its performance will depend on the choice of the starting point.

# Prediction

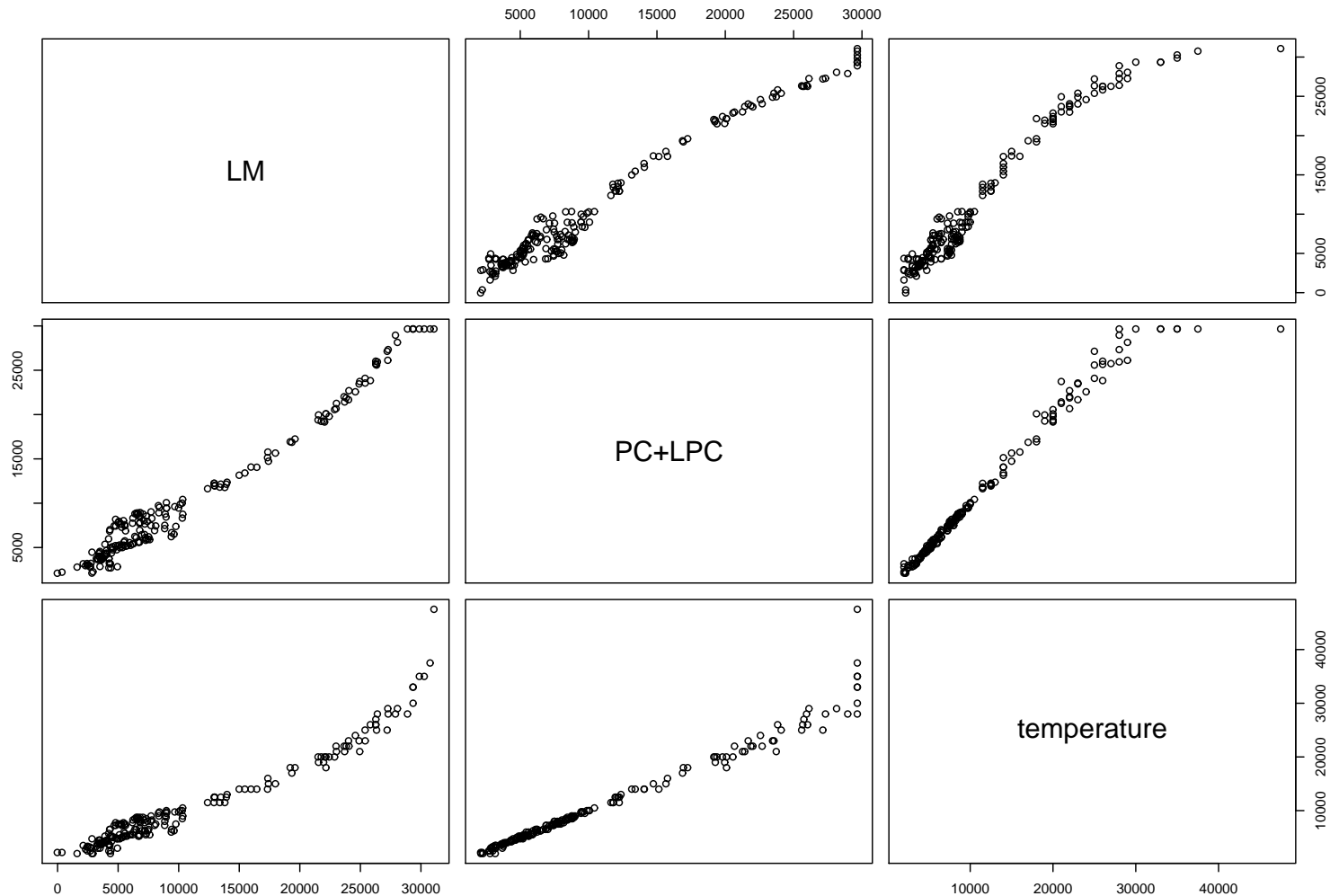
- For a new observation  $x_{new}$  (i.e., here, a new set of spectra), prediction proceeds as follows:
  - Project  $x_{new}$  onto the LPC, giving  $t_{new}$ .
  - Compute  $\hat{y}_{new} = f(t_{new})$  from the fitted regression model.
- Comparison: Prediction error ( $*10^3$ ) for 200 observations sampled from the training data set:

	LM	PC+Regr	LPC+Regr	PC+LPC+Regr
average( $\hat{\varepsilon}_i^2$ )	4,119	4,395	2,215	2,633
median( $\hat{\varepsilon}_i^2$ )	1,035	1,300	66	51

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

# Take care with boundaries!

- Compare predicted values through LM and PC+LPC with true temperature values:



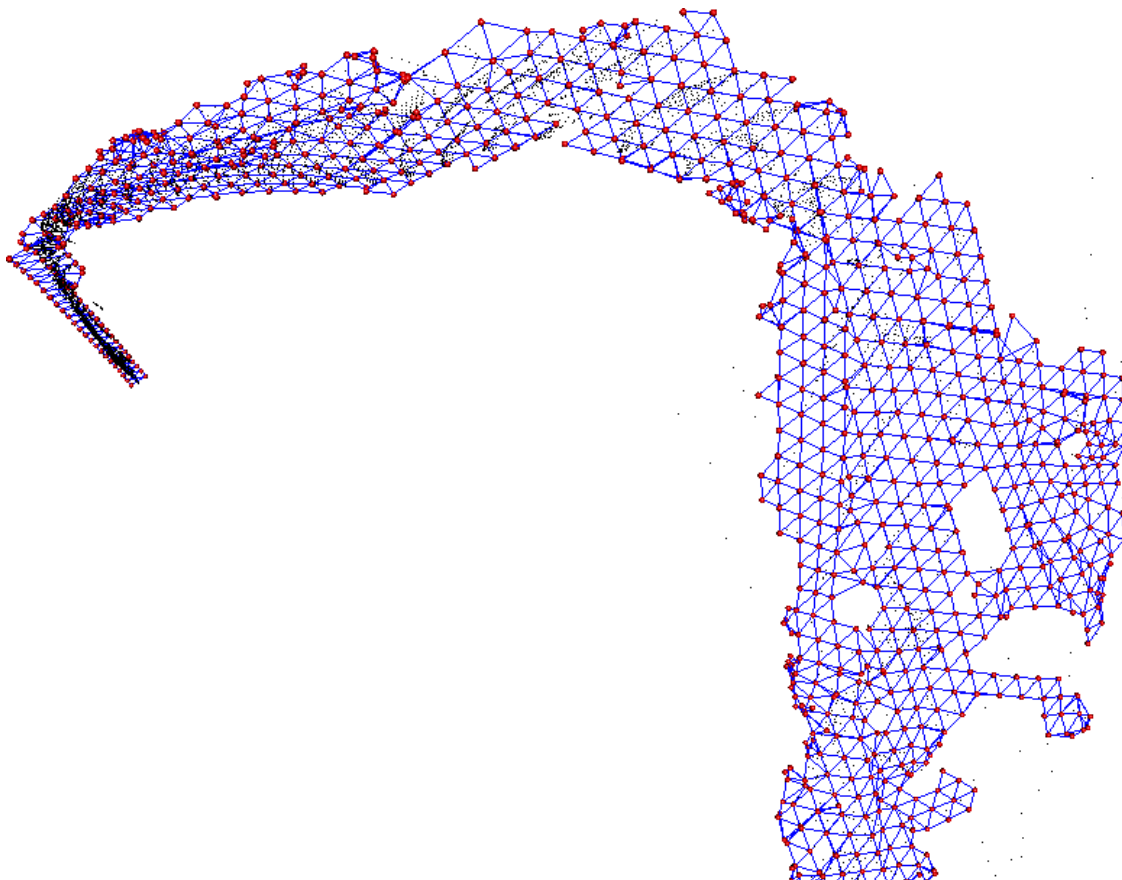
# Conclusion and Outlook

---

- Local principal curves are well suited to compress complex high-dimensional data structures, as long as the intrinsic dimensionality of the data cloud is close to one.
- When the intrinsic dimensionality is two or larger, the extension to *local principal manifolds* should be considered.

# Conclusion and Outlook

- Local principal curves are well suited to compress complex high-dimensional data structures, as long as the intrinsic dimensionality of the data cloud is close to one.
- When the intrinsic dimensionality is two or larger, the extension to *local principal manifolds* should be considered.



...work in progress!

# Literature

---

- Einbeck, Tutz & Evers** (2005): Local principal curves. *Statistics and Computing* **15**, 301–313.
- Einbeck, Tutz & Evers** (2005b): Exploring multivariate data structures with local principal curves. In: Weihs, C. and Gaul, W. (Eds.): *Classification - The Ubiquitous Challenge*. Springer, Heidelberg.
- Einbeck, Evers & Bailer-Jones** (2007): 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.