Principal curves and surfaces: Data visualization, compression, and beyond

Jochen Einbeck

Department of Mathematical Sciences, Durham University

jochen.einbeck@durham.ac.uk

Canterbury, 3rd of October 2013

Durham University
Durham Energy Institute

in collaboration with:

Ludger Evers (University of Glasgow)

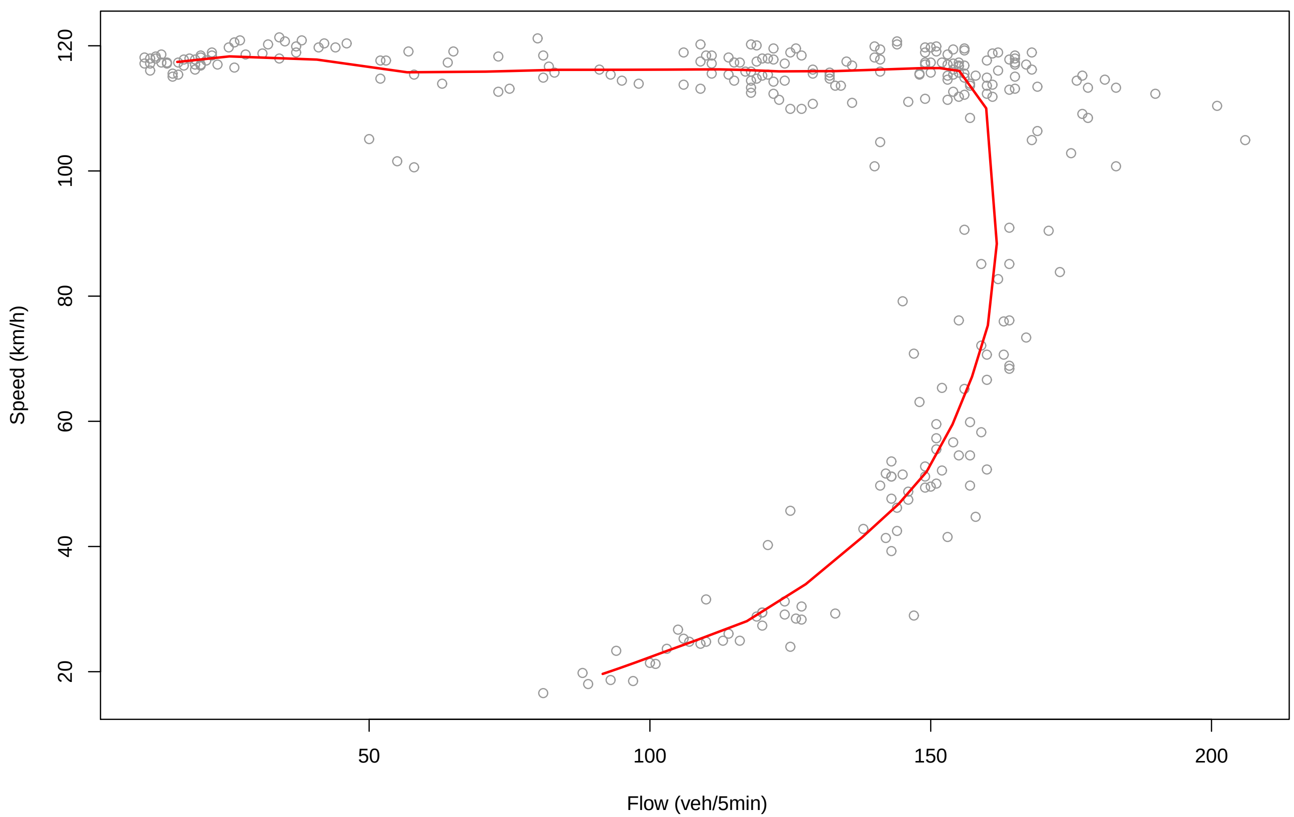
Alessandro Parente and Ben Isaac (University Libré de Bruxelles)

Brian Caffo (Johns Hopkins University)

Principal curves

Principal Curves are smooth curves passing through the 'middle' of a multivariate data cloud.

Example: Speed-Flow data from a Californian 'freeway'.



Types of principal curves

Today exist a variety of different notions of principal curves, which vary essentially in how the "middle" of the data cloud is defined/found:

- Global ('top-down') algorithms start with an initial line (usually the 1st PC line) and bend this line or concatenate other lines to it until some convergence criterion is met (Hastie & Tibshirani (HS) 1989, Tibshirani 1992, Kégl et al. 2002).
  - Allows theoretical analysis (based on global optimization criterion).
  - Problems with strongly twisted, branched, or disconnected data clouds.

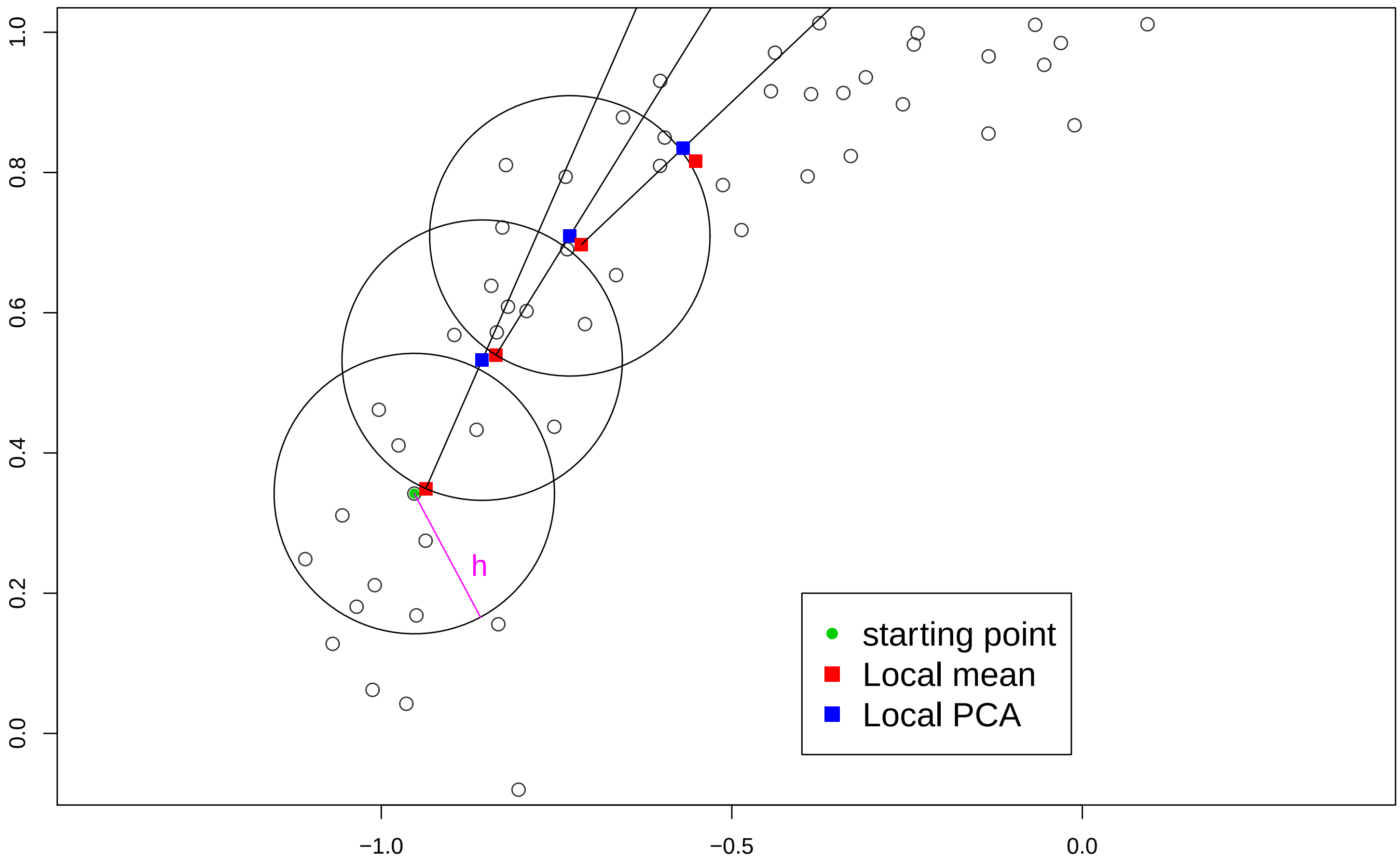
Types of principal curves

Today exist a variety of different notions of principal curves, which vary essentially in how the "middle" of the data cloud is defined/found:

- Global ('top-down') algorithms start with an initial line (usually the 1st PC line) and bend this line or concatenate other lines to it until some convergence criterion is met (Hastie & Tibshirani (HS) 1989, Tibshirani 1992, Kégl et al. 2002).
    - Allows theoretical analysis (based on global optimization criterion).
    - Problems with strongly twisted, branched, or disconnected data clouds.
- Local ('bottom-up') algorithms estimate the principal curve locally moving step by step through the data cloud (Delicado 2001, Einbeck et al. 2005).
    - More flexible, but far more variable.
    - Extend straightforwardly to branched and disconnected data.

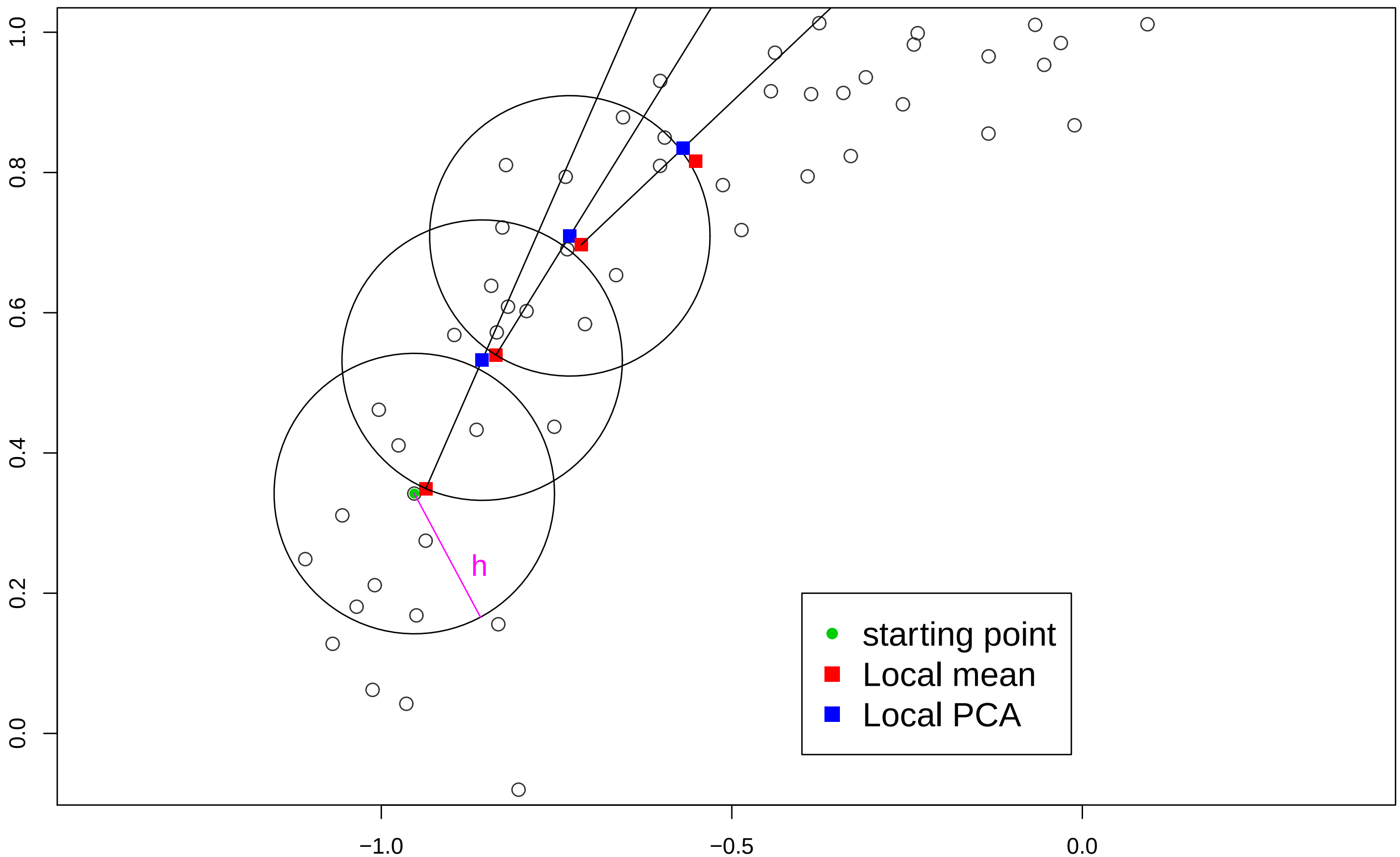
Local principal curves (LPC)

- 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.

Algorithm for LPCs

- Given: A data cloud $X = (X_1, \dots, X_n)$, where $X_i \in \mathbb{R}^d$.

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* $w_i = K_H(X_i - x) X_i / \sum_{i=1}^n K_H(X_i - x)$.
3. *Compute the* 1^{st} *local eigenvector* γ^x *of*
$$\Sigma^x = \sum_{i=1}^n w_i (X_i - \mu^x)(X_i - \mu^x)^T$$
4. *Step from* μ^x *to* $x := \mu^x + t_0 \gamma^x$*.*
5. *Repeat steps 2. to 4. until the* μ^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 μ^x makes up the local principal curve (LPC).

Algorithm for LPCs

- Given: A data cloud $X = (X_1, \ldots, X_n)$, where $X_i \in \mathbb{R}^d$.

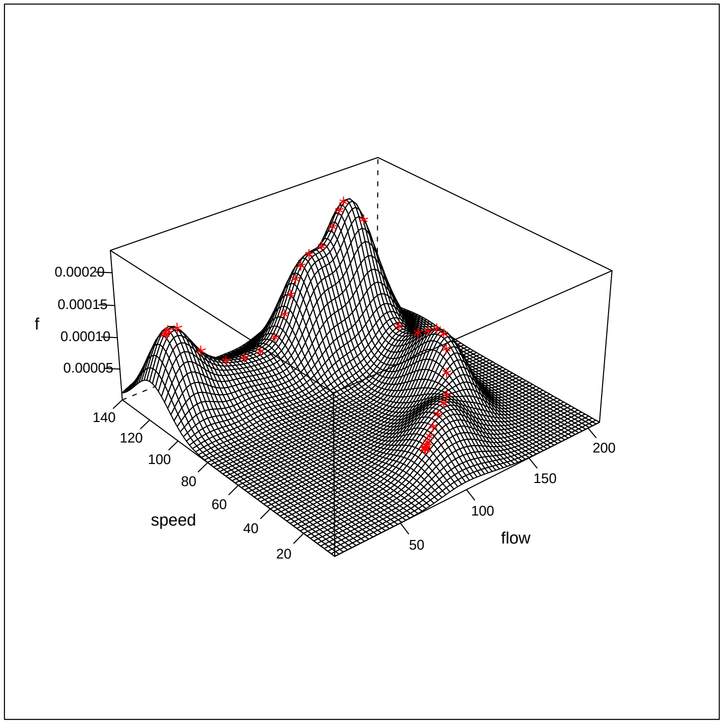
1. *Choose a starting point* x_0. *Set* $x = x_0$.
2. *At* x, *calculate the local center of mass* $\mu^x = \sum_{i=1}^n w_i X_i$, *where* $w_i = K_H(X_i - x)X_i / \sum_{i=1}^n K_H(X_i - x)$.
3. *Compute the* 1^{st} *local eigenvector* γ^x *of*
$$\Sigma^x = \sum_{i=1}^n w_i (X_i - \mu^x)(X_i - \mu^x)^T$$
4. *Step from* μ^x *to* $x := \mu^x + t_0 \gamma^x$.
5. *Repeat steps 2. to 4. until the* μ^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 μ^x makes up the local principal curve (LPC).

- Need "signum flipping" of γ^x at every loop in order to maintain direction of curve.

Which feature do LPCs extract?

- A local principal curve approximates the density ridge.



Kernel density estimate:

$$\hat{f}(x) = \frac{1}{n|H|^{\frac{1}{2}}} \sum_{i=1}^{n} K\left(H^{-1/2}(X_i - x)\right)$$

For $H = \text{diag}(h^2)$, at ℓ-th iteration,

$$\mu^{x_{\ell+1}} - \mu^{x_\ell} \approx \left[\frac{1}{f(x_\ell)}h^2 \pm \frac{1}{||\nabla f(x_\ell)||}t_0\right] \nabla f(x_\ell)$$

(Einbeck & Zayed, 2013).

Some theory and simulation

- The LPC stops when
$$f(x) = \frac{h^2}{t_0}||\nabla f(x)||$$
- Special case: $X \sim N(0, \sigma^2 \boldsymbol{I})$. Then $f(x) = c||\nabla f(x)||$ iff $||x|| = \frac{1}{c}\sigma^2$.

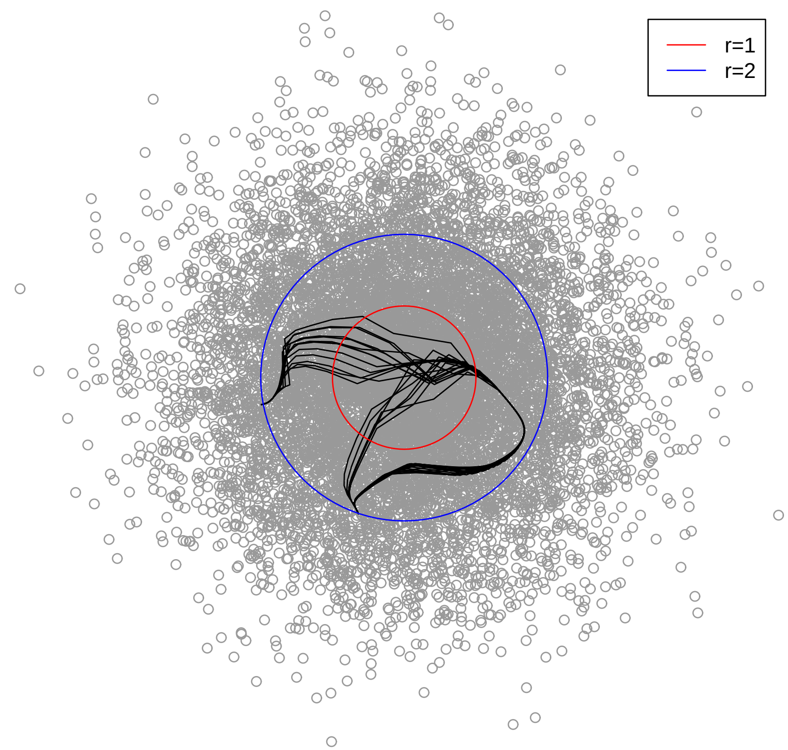
Some theory and simulation

- The LPC stops when

$$f(x) = \frac{h^2}{t_0} ||\nabla f(x)||$$

- Special case: $X \sim N(0, \sigma^2 \boldsymbol{I})$. Then $f(x) = c||\nabla f(x)||$ iff $||x|| = \frac{1}{c}\sigma^2$.

- Simulation: BVN with $\sigma^2 = 2$.
- 20 LPCs with $h = 1$, $t_0 = 1$ started within circle of radius $r = 1$.
- All of them converge to blue circle $r = \sigma^2 = 2$.



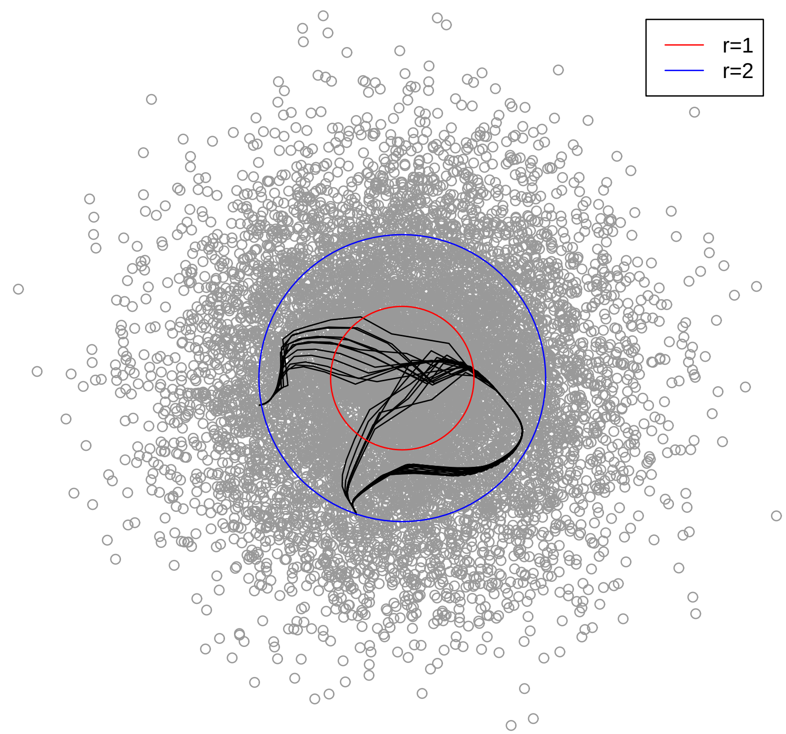
Some theory and simulation

- The LPC stops when

$$f(x) = \frac{h^2}{t_0} ||\nabla f(x)||$$

- Special case: $X \sim N(0, \sigma^2 \boldsymbol{I})$. Then $f(x) = c||\nabla f(x)||$ iff $||x|| = \frac{1}{c}\sigma^2$.

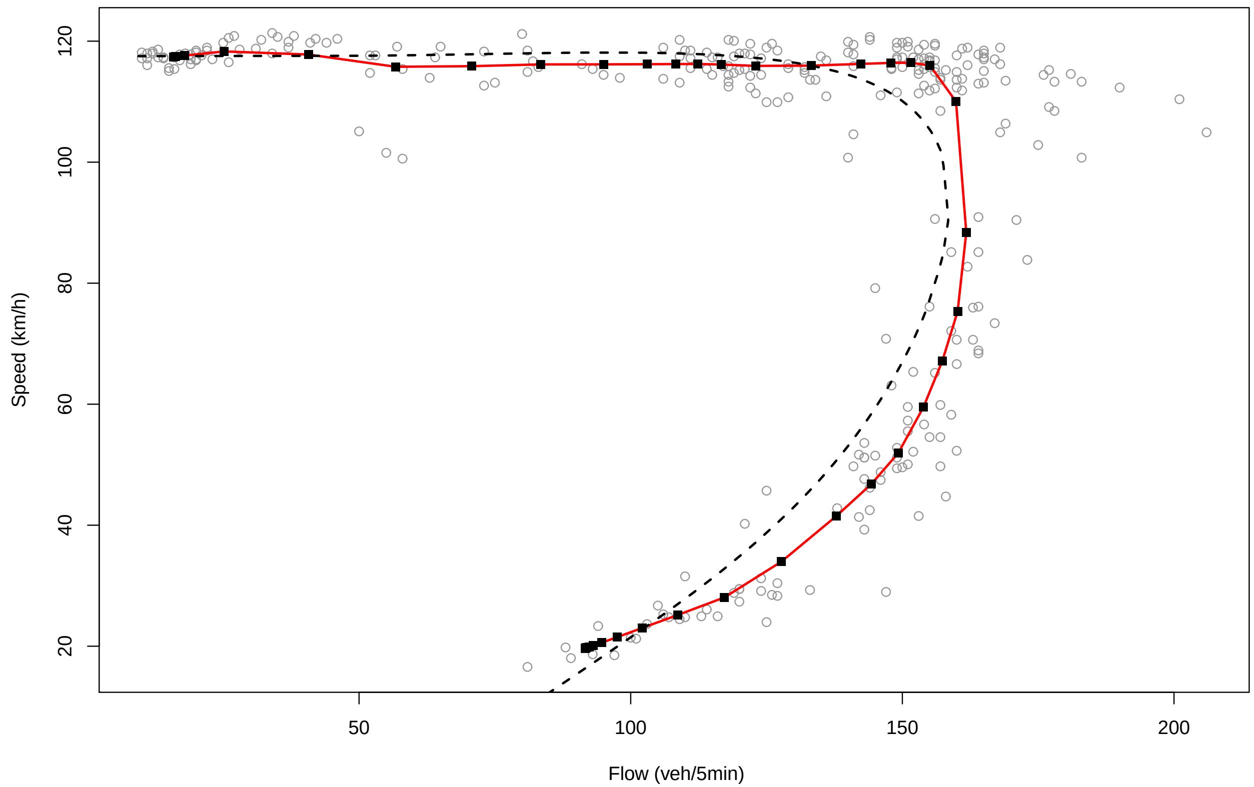
- Simulation: BVN with $\sigma^2 = 2$.
- 20 LPCs with $h = 1$, $t_0 = 1$ started within circle of radius $r = 1$.
- All of them converge to blue circle $r = \sigma^2 = 2$.



- Can be exploited for boundary extension (Einbeck & Zayed, 2013).

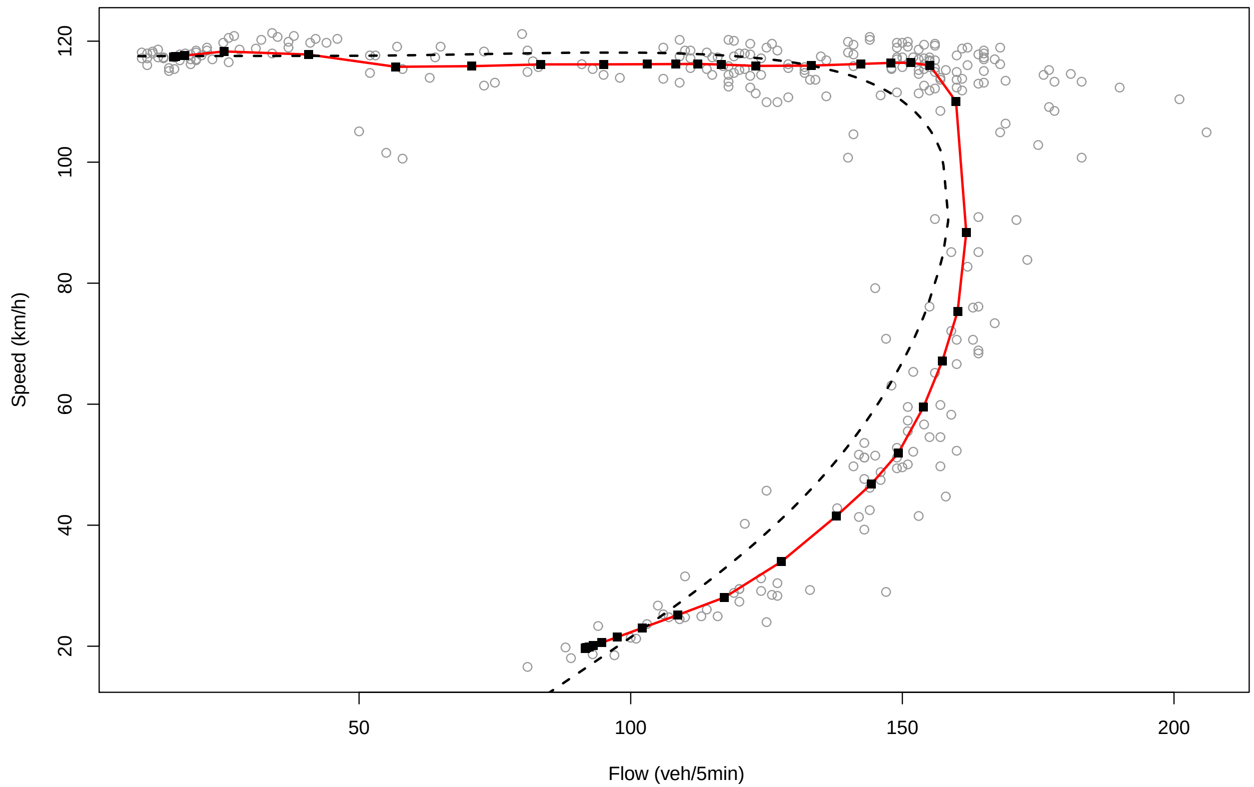
Application on traffic data

- LPC (red curve, $h = t_0 = 12$) with local centers of mass μ^x (black squares). A HS curve is shown for comparison (black, dashed).



Application on traffic data

- LPC (red curve, $h = t_0 = 12$) with local centers of mass μ^x (black squares). A HS curve is shown for comparison (black, dashed).



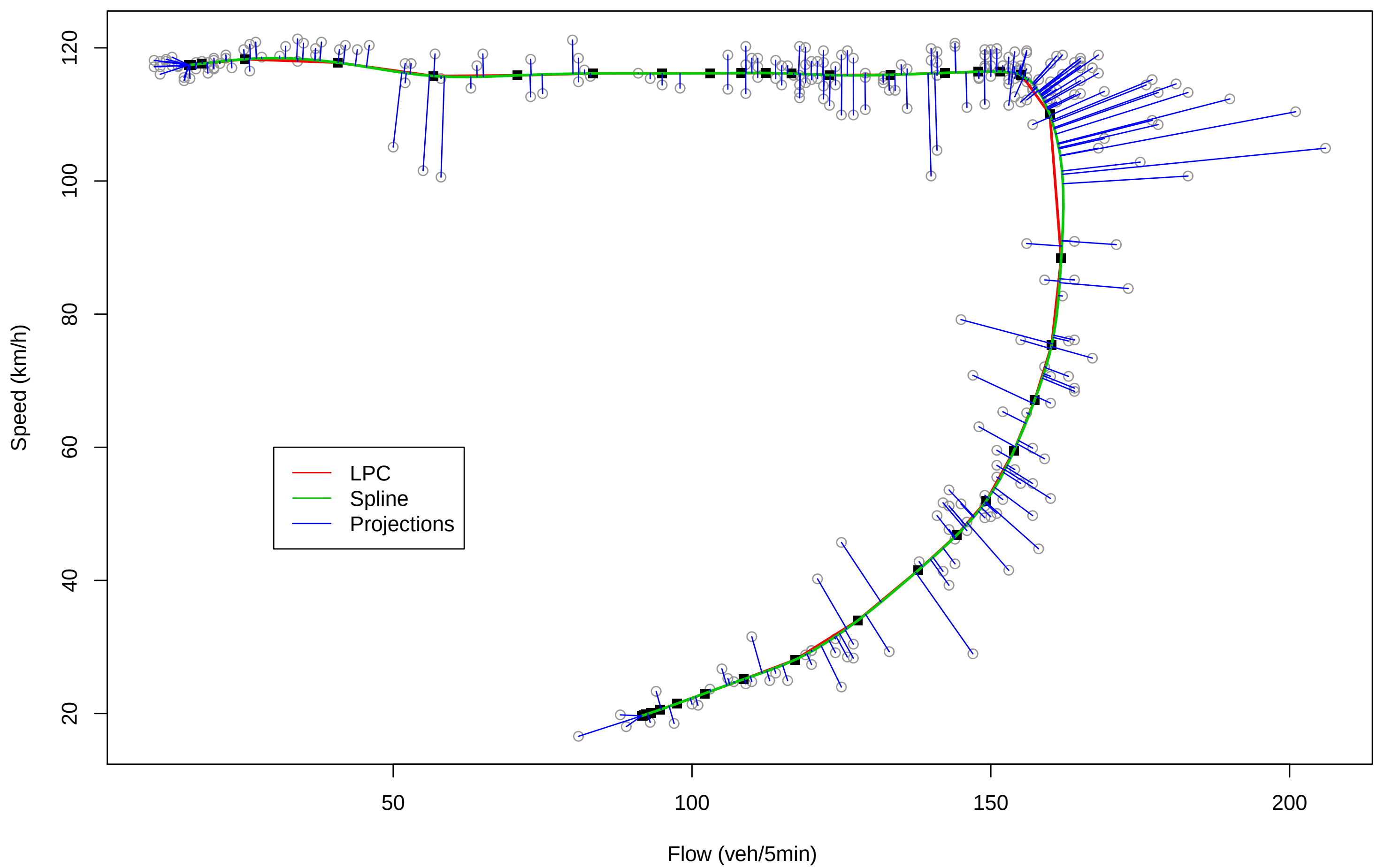
- To make further use of this curve, we need to be able to project data onto it.

Parametrization and Projection

- 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 local means $\mu \in \mathbb{R}^d$.
- 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, t, along the curve through the arc length of the spline function.
- 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.

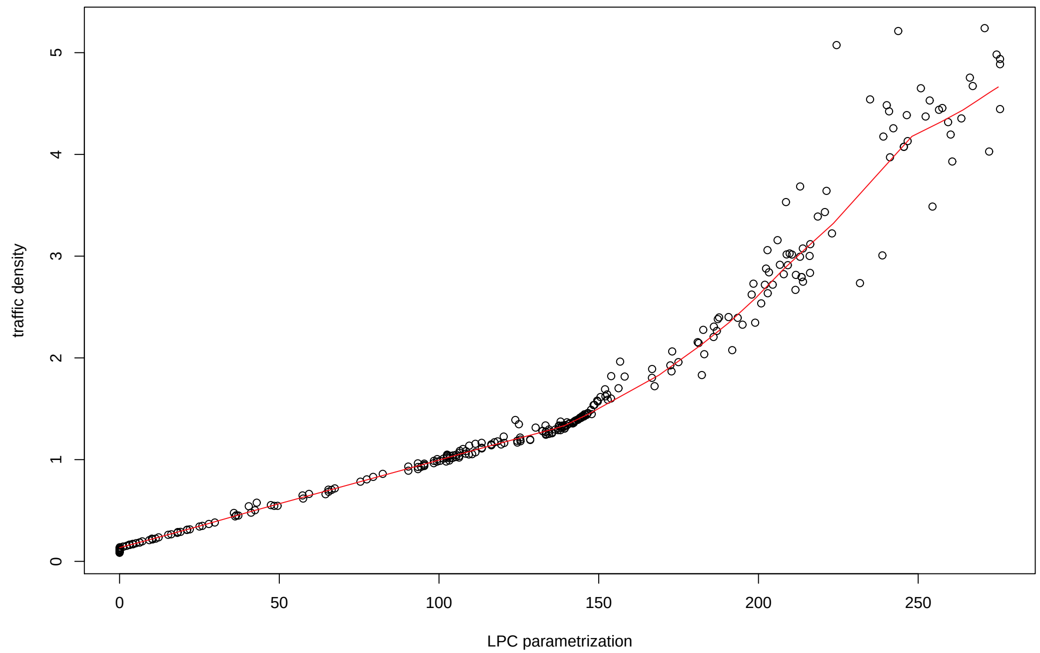
Illustration: traffic data

Original LPC, spline, and projections for speed-flow data:



Interpreting the curve parametrization

- Plotted are traffic density (flow/speed) versus the curve parametrization t.

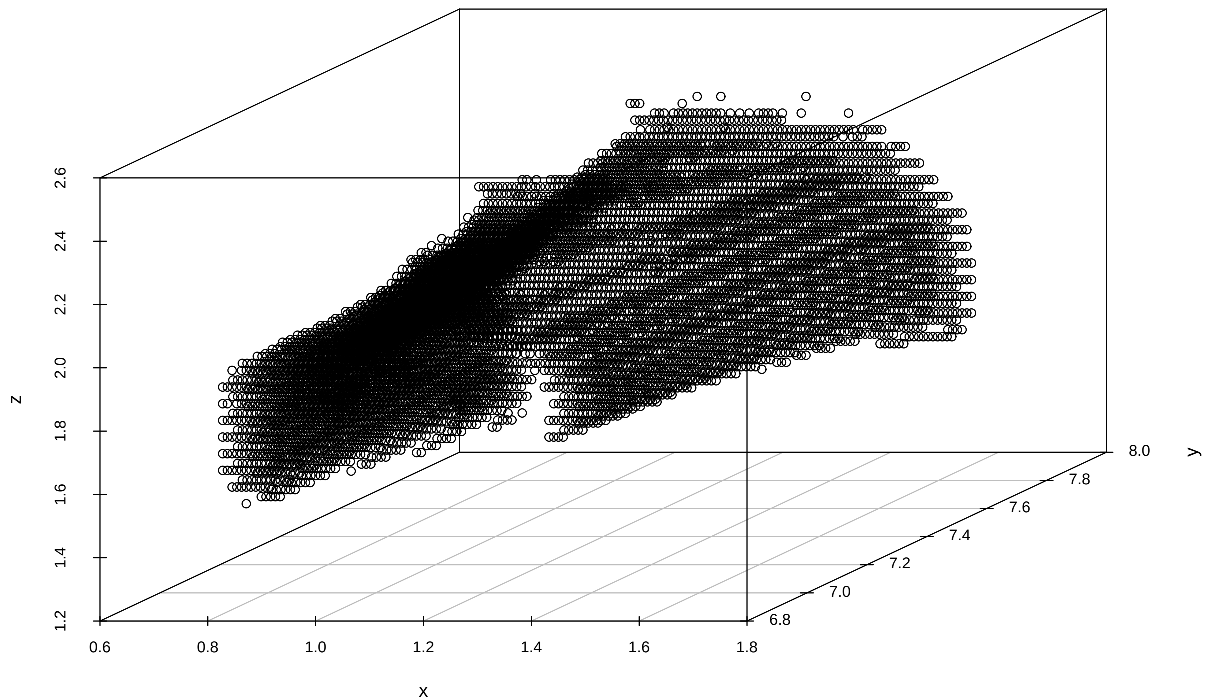


- A calibration curve can be used to link the parametrization to physical variables.

(Einbeck & Dwyer, 2011)

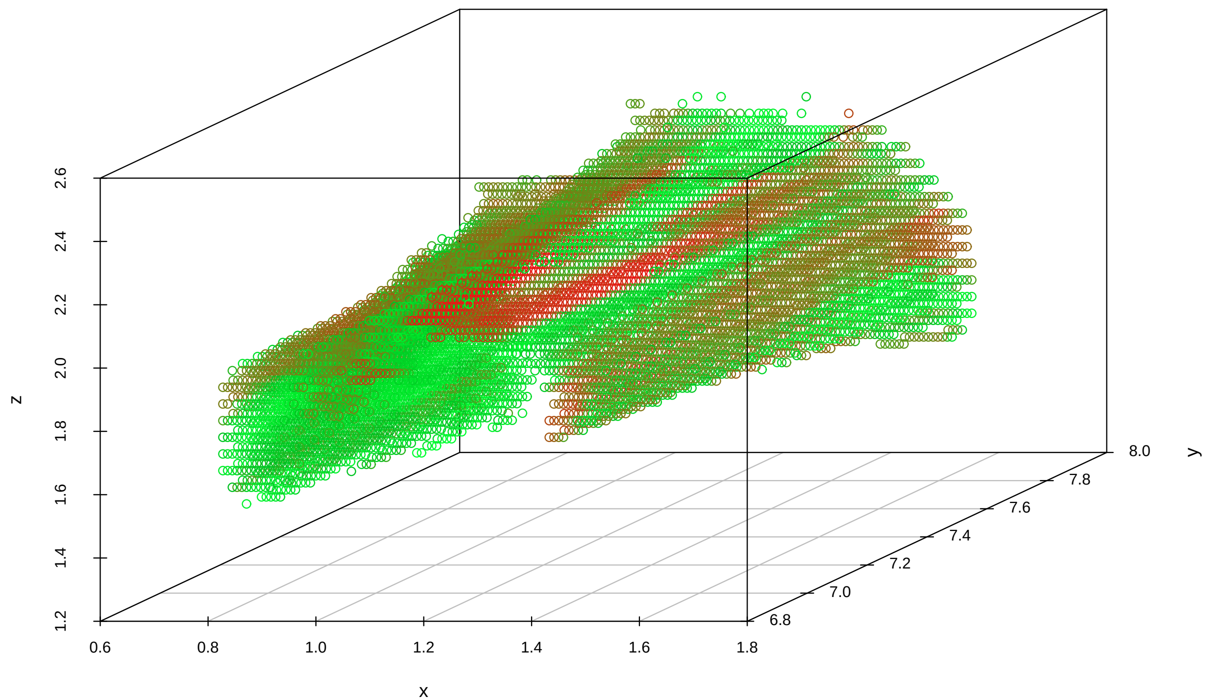
From curves to surfaces

- Example from neuroscience: FMRI scan (3d coordinates) of the ‘corpus callosum’ for a ‘healthy volunteer’



From curves to surfaces

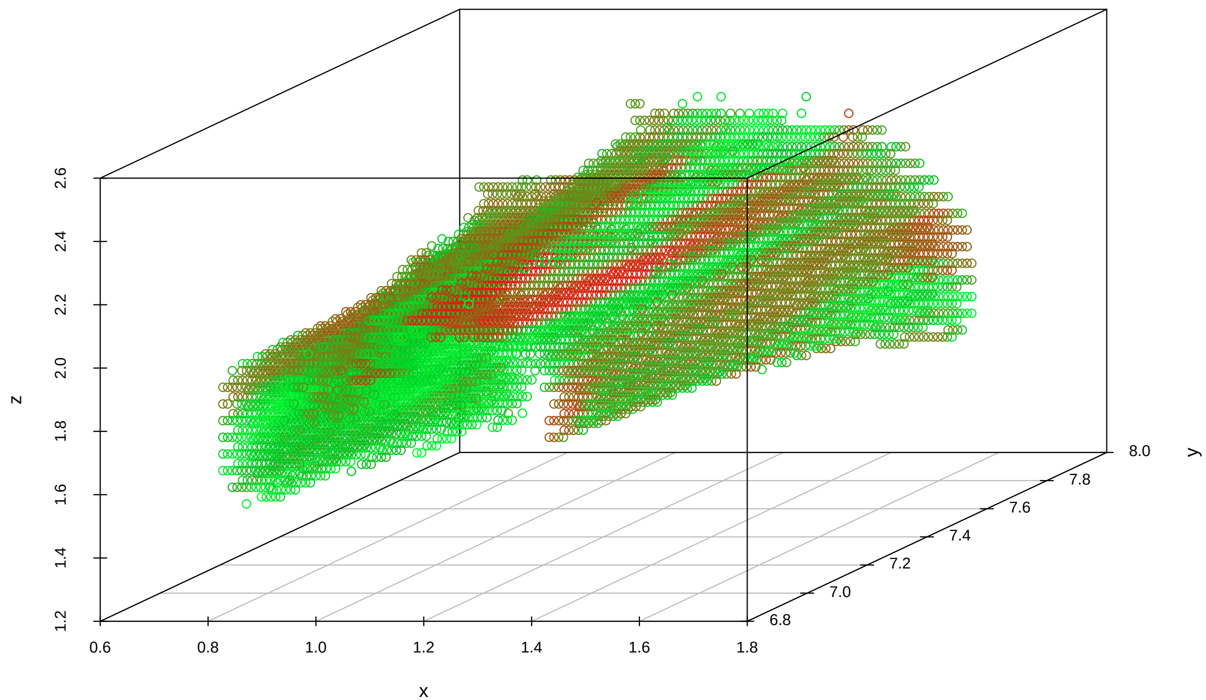
- Example from neuroscience: FMRI scan (3d coordinates) of the 'corpus callosum' for a 'healthy volunteer'



with associated intensities of 'fractional anisotropy' (red=high, green=small).

From curves to surfaces

- Example from neuroscience: FMRI scan (3d coordinates) of the 'corpus callosum' for a 'healthy volunteer'



with associated intensities of 'fractional anisotropy' (red=high, green=small).

- Can we provide a 'map' of intensities on the surface?

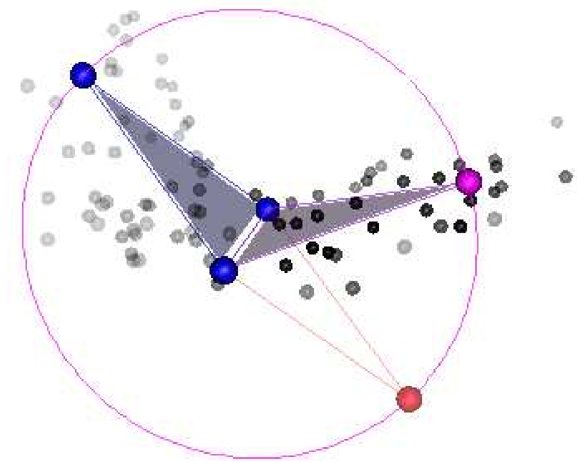
Principal surfaces

- 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.

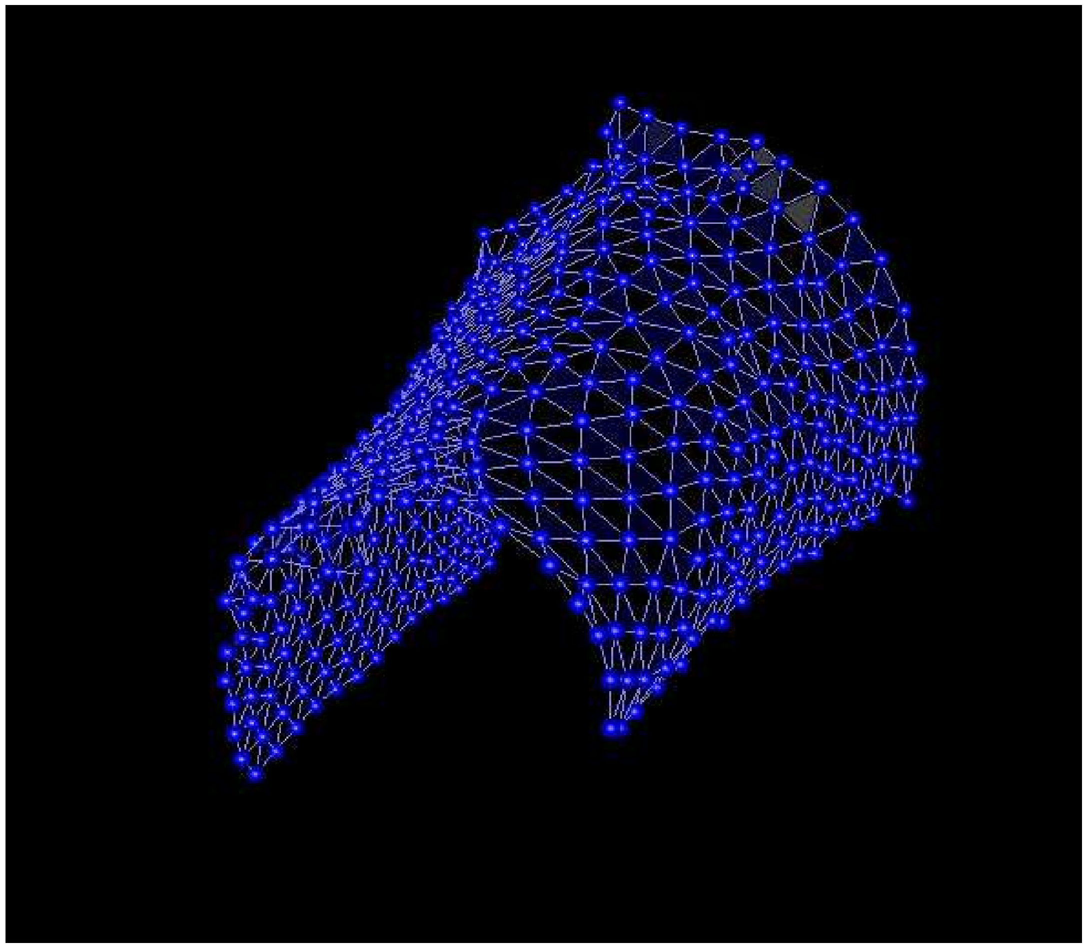


... until all triangles have been considered.

(Einbeck, Evers & Powell, 2010)

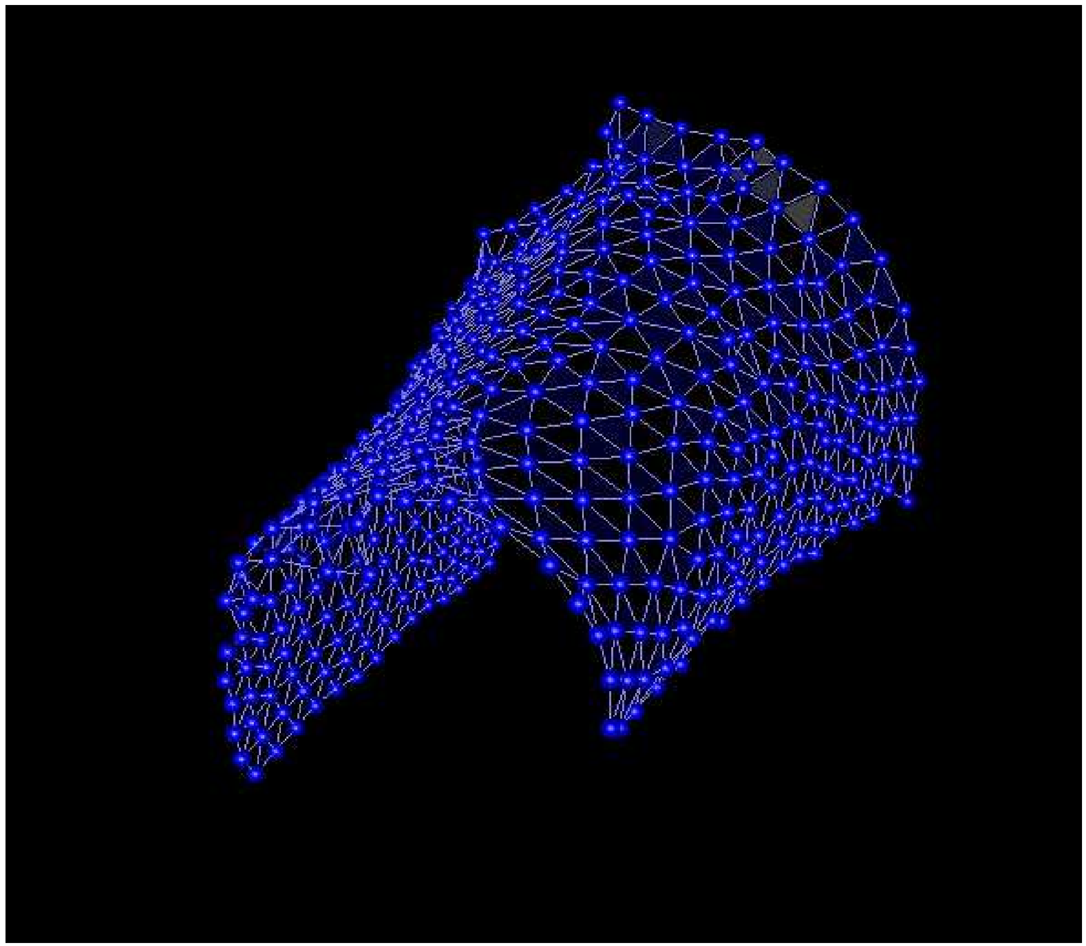
Principal surfaces (cont.)

- Local principal surface fitted to FMRI scan:



Principal surfaces (cont.)

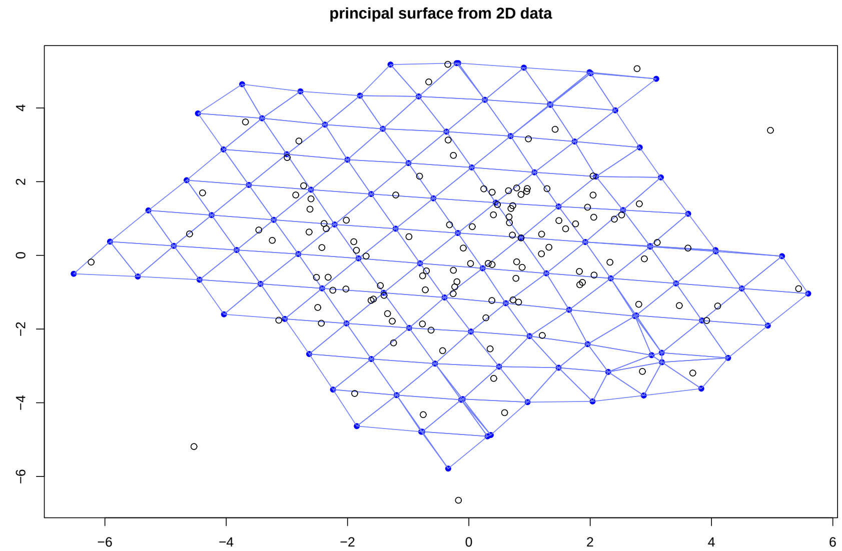
- Local principal surface fitted to FMRI scan:



- Then, how to use this for regression?

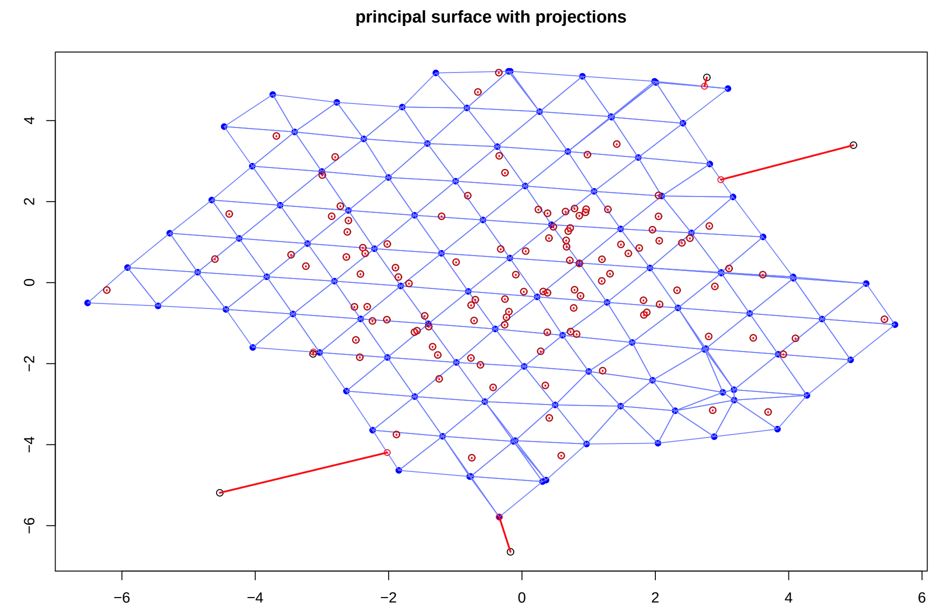
Regression on principal surface

- Toy example: A principal surface for bivariate data.



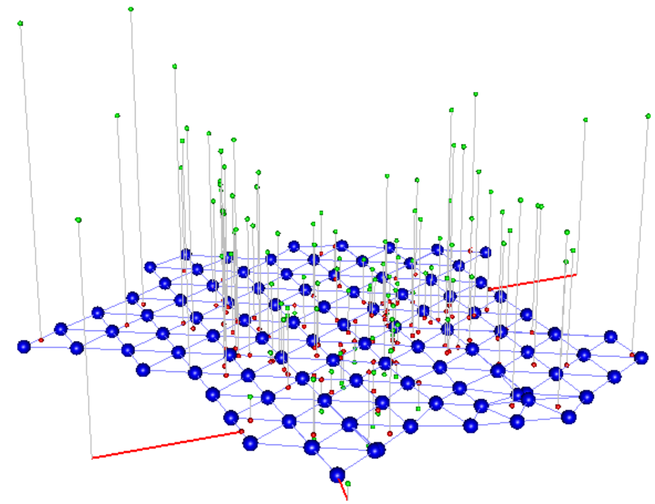
Regression on principal surface

- 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.



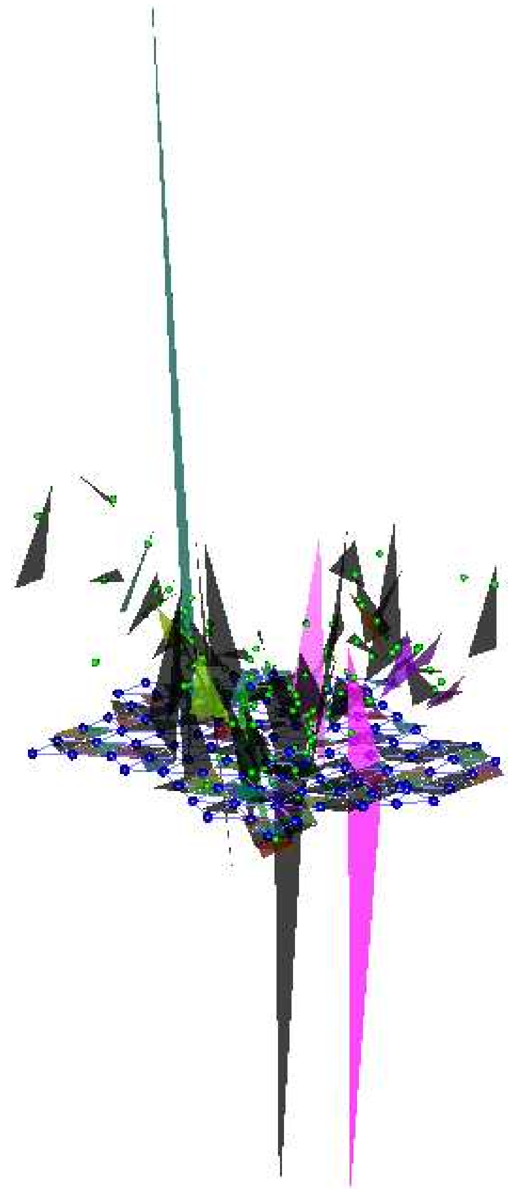
- Next, consider a response y_i.
- Assume separate regression models for each triangle j

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

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.

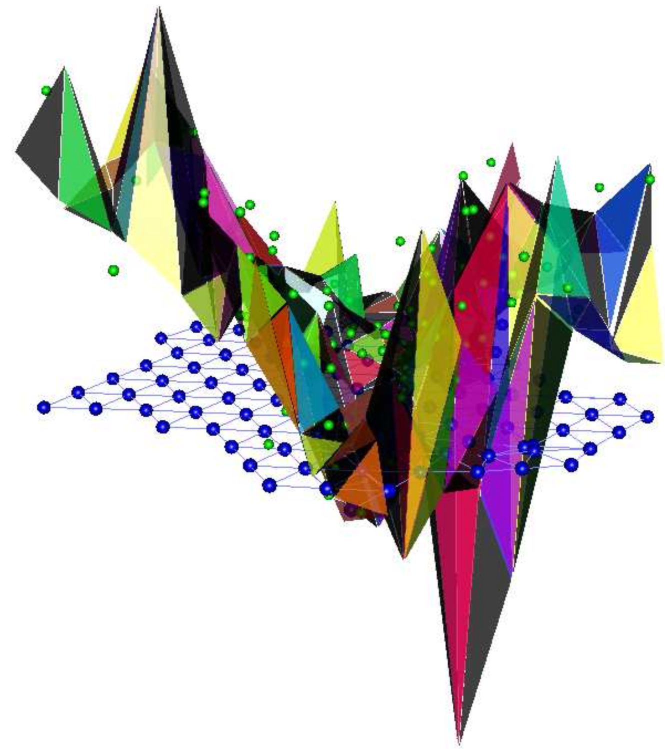
Penalized regression

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



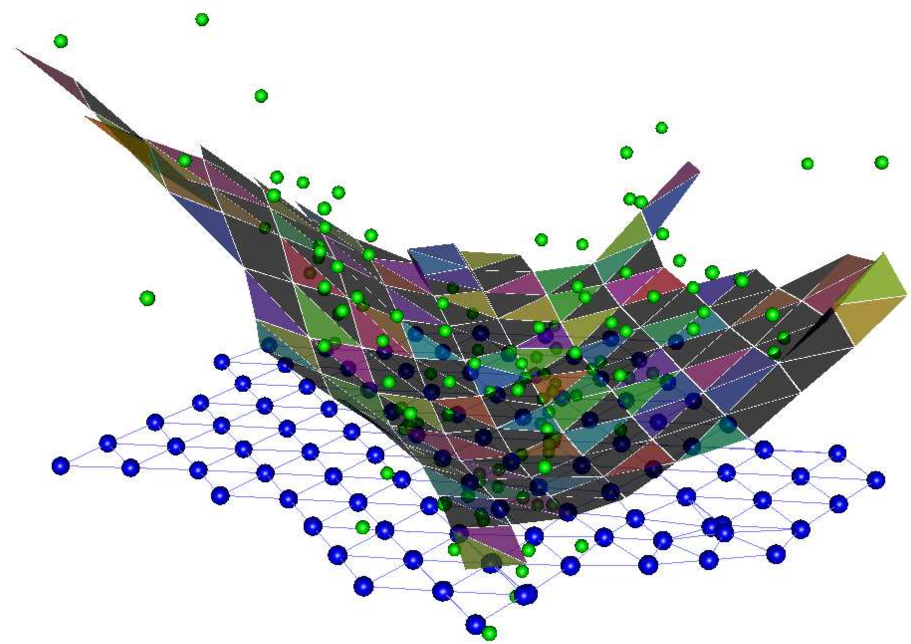
Penalized regression

- 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

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

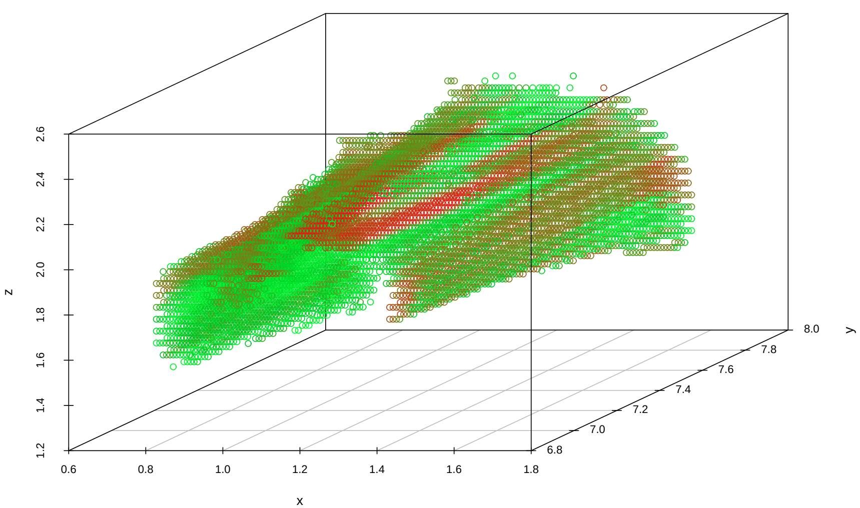
- Though the matrices $\mathbf{Z}$, $\mathbf{D}$ and $\mathbf{E}$ can be very large, they are also very sparse, which allows for quick computations.
- The solution is given by

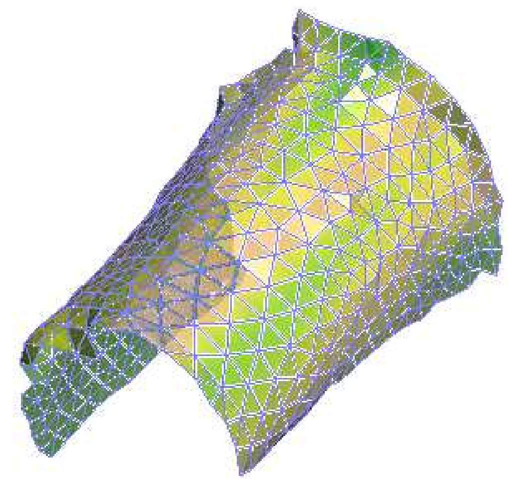
$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z} + \lambda\mathbf{D}'\mathbf{D} + \mu\mathbf{E}'\mathbf{E}\ )^{-1}\mathbf{Z}'\mathbf{y}.$$

(Evers & Einbeck, 2013)

Regression on the corpus callosum

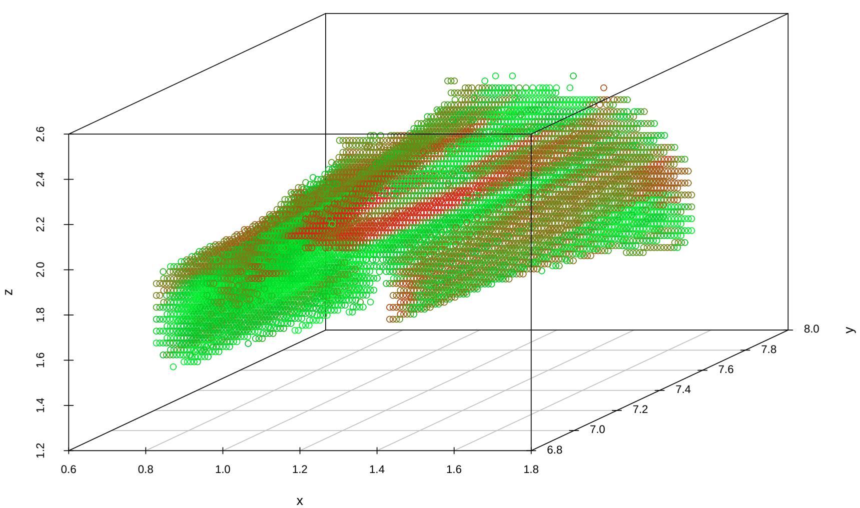
- Raw data (left), with estimated principal surface (right), shaded according to fitted intensities:

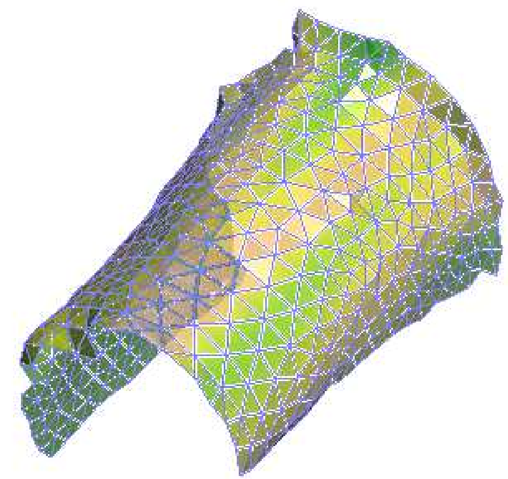




Regression on the corpus callosum

- Raw data (left), with estimated principal surface (right), shaded according to fitted intensities:





- Future goal: Relate fitted (ideally flattened) surface to scalar disability scores...

Case study: Combustion

- Combustion is a sequence of exothermic chemical reactions between a fuel and an oxidant
- accompanied by the production of heat (light, flames)
- Most simple example: combustion of hydrogen and oxygen to water vapor

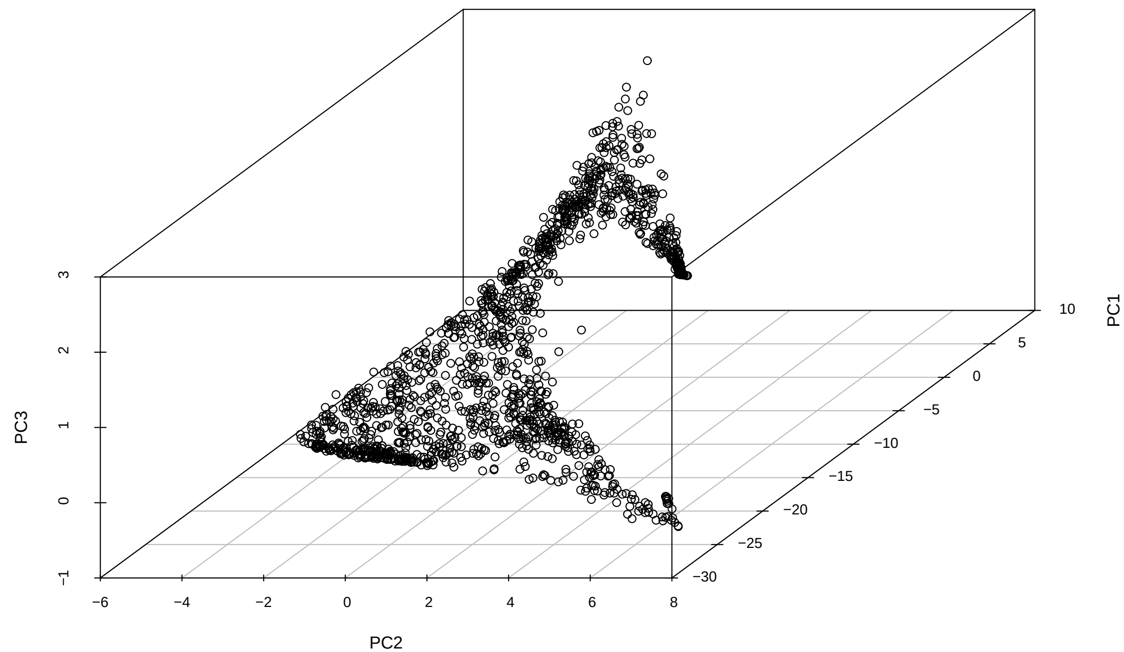
$$2H_2 + O_2 \longrightarrow 2H_20$$

Case study: Combustion

- Combustion is a sequence of exothermic chemical reactions between a fuel and an oxidant
- accompanied by the production of heat (light, flames)
- Most simple example: combustion of hydrogen and oxygen to water vapor
$$2H_2 + O_2 \longrightarrow 2H_20$$
- A combustion system involving p chemical species is described by its thermochemical state
$$\mathbf{\Phi} = [z_1, \ldots, z_{p-1}, T],$$
with $p-1$ chemical mass fractions $z_1, \ldots, z_{p-1}$, and temperature T.
- The (space/time) behavior of $\mathbf{\Phi}$ is governed by a set of p highly coupled transport equations.
- For large p, this system of equations is usually intractable.

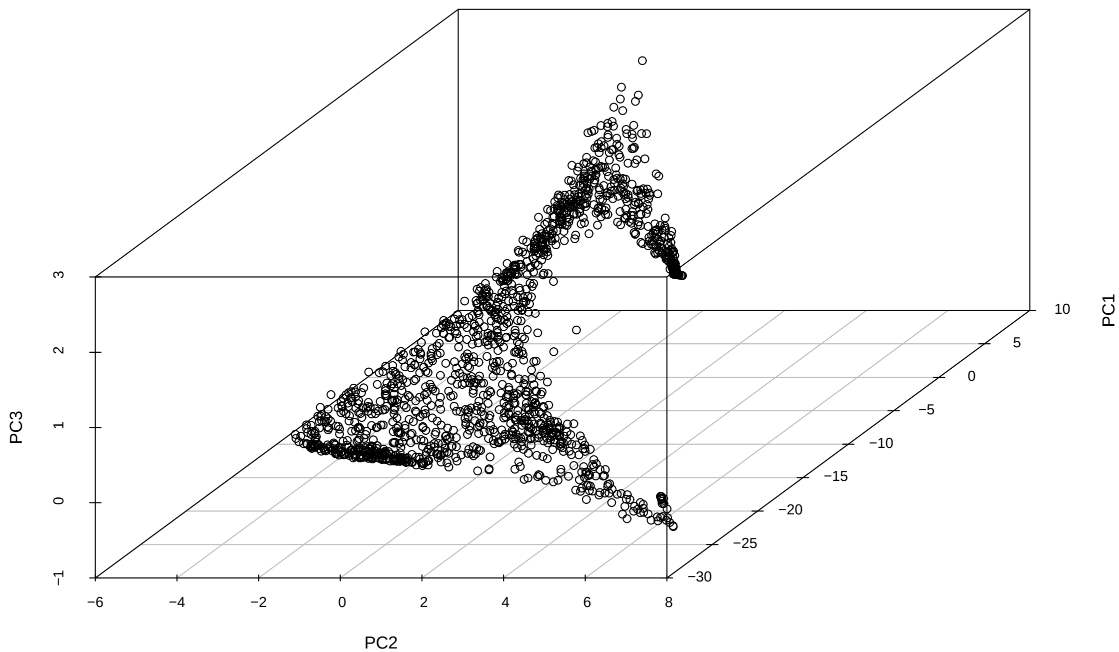
Combustion data

- Simulated combustion system with 11 chemical species
  H_2, O_2, O, OH, H_2O, H, HO_2, H_2O_2, CO, CO_2, HCO
- First three principal components of state space Φ ($n = 4000$):



Combustion data

- Simulated combustion system with 11 chemical species
  H_2, O_2, O, OH, H_2O, H, HO_2, H_2O_2, CO, CO_2, HCO
- First three principal components of state space Φ ($n = 4000$):



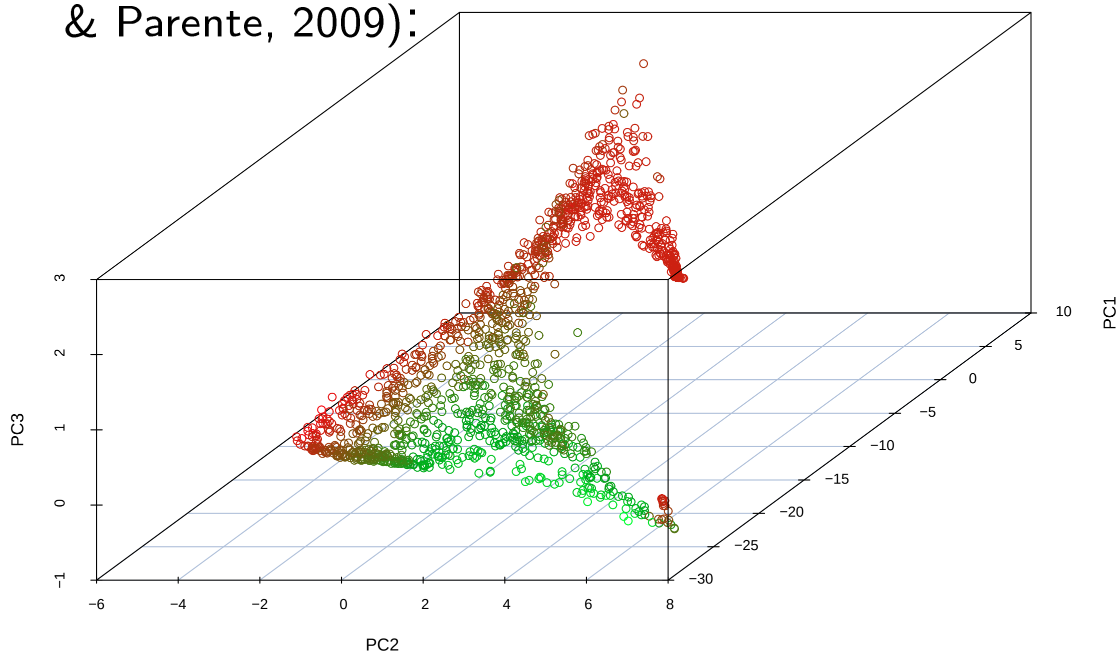
- It is well-known that the thermochemical state space of combustion systems resides on low–dimensional manifolds.
- This is convenient, as the transport equations based on the reduced system of, say, 3 principal components *are* tractable.

Combustion data

- Complication: The rates of production ('source terms') of the principal components are unknown.
- In practice, they have to be found by regression on the principal components.

Combustion data

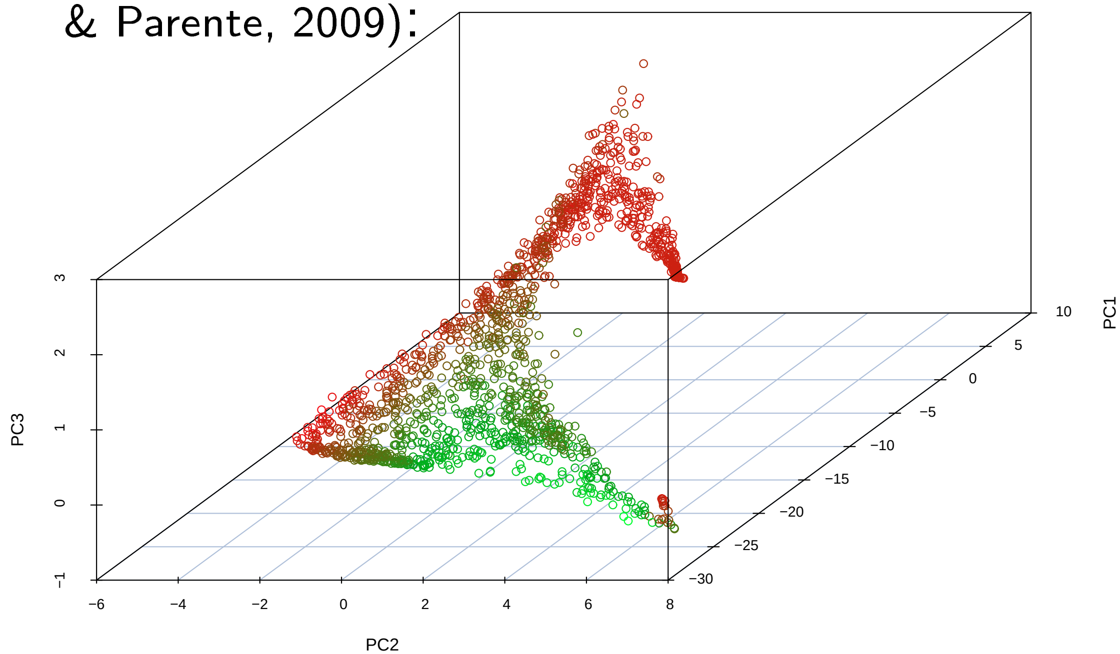
- Complication: The rates of production ('source terms') of the principal components are unknown.
- In practice, they have to be found by regression on the principal components.
- Requires 'high–fidelity' data with tabulated source terms (Sutherland & Parente, 2009):



red=high
green=low
first PC source terms.

Combustion data

- Complication: The rates of production ('source terms') of the principal components are unknown.
- In practice, they have to be found by regression on the principal components.
- Requires 'high–fidelity' data with tabulated source terms (Sutherland & Parente, 2009):



red=high
green=low
first PC source terms.

- Clearly, the position on the surface (=2D manifold) is informative for the source terms.

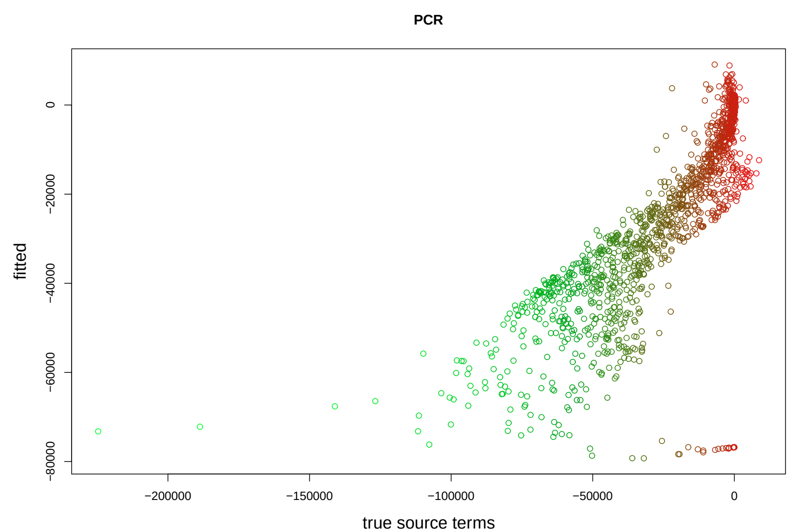
Principal component regression

- A simple approach is to use Principal component regression, where the first three principal component scores serve as predictors, and the source terms, s, as response:

$$s = \beta_0 + \beta_1 \mathrm{PC}_1 + \beta_2 \mathrm{PC}_2 + \beta_3 \mathrm{PC}_3 + \epsilon$$

  (Sutherland & Parente, 2009).

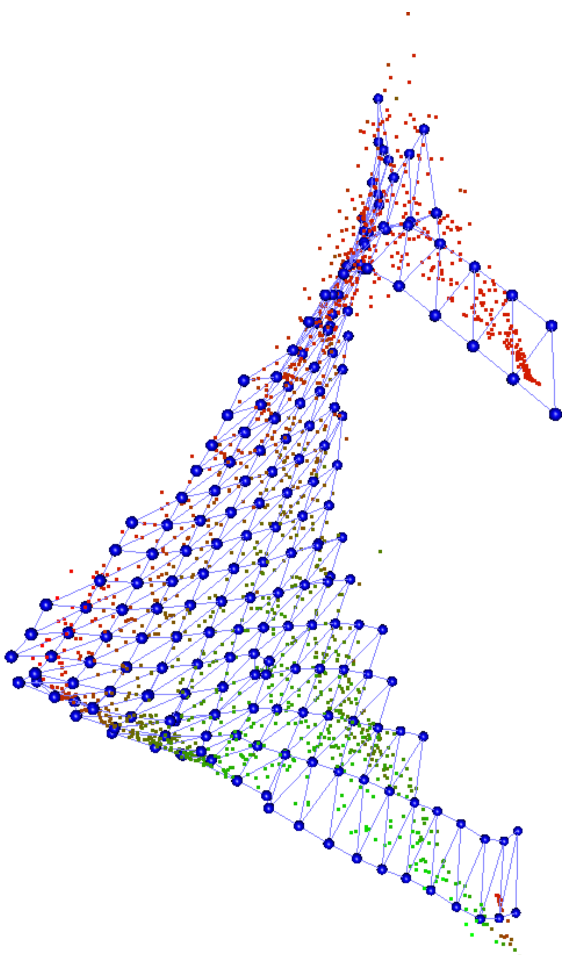
- Fitted versus true values ($R^2 = 0.77$):

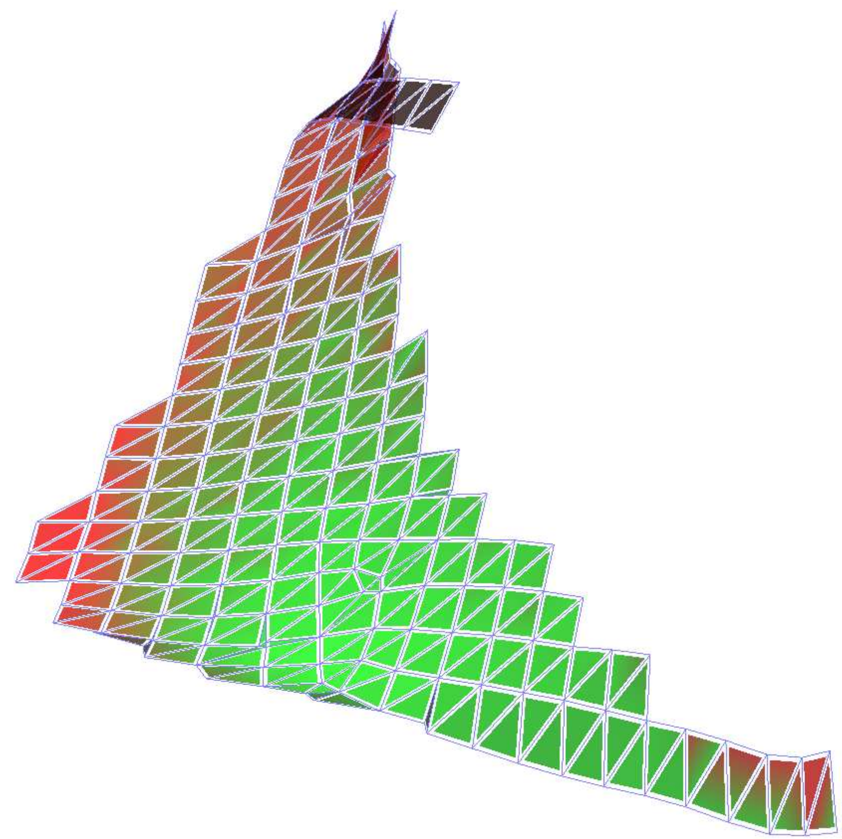


- ... turns out to be not good enough!

Principal surface regression

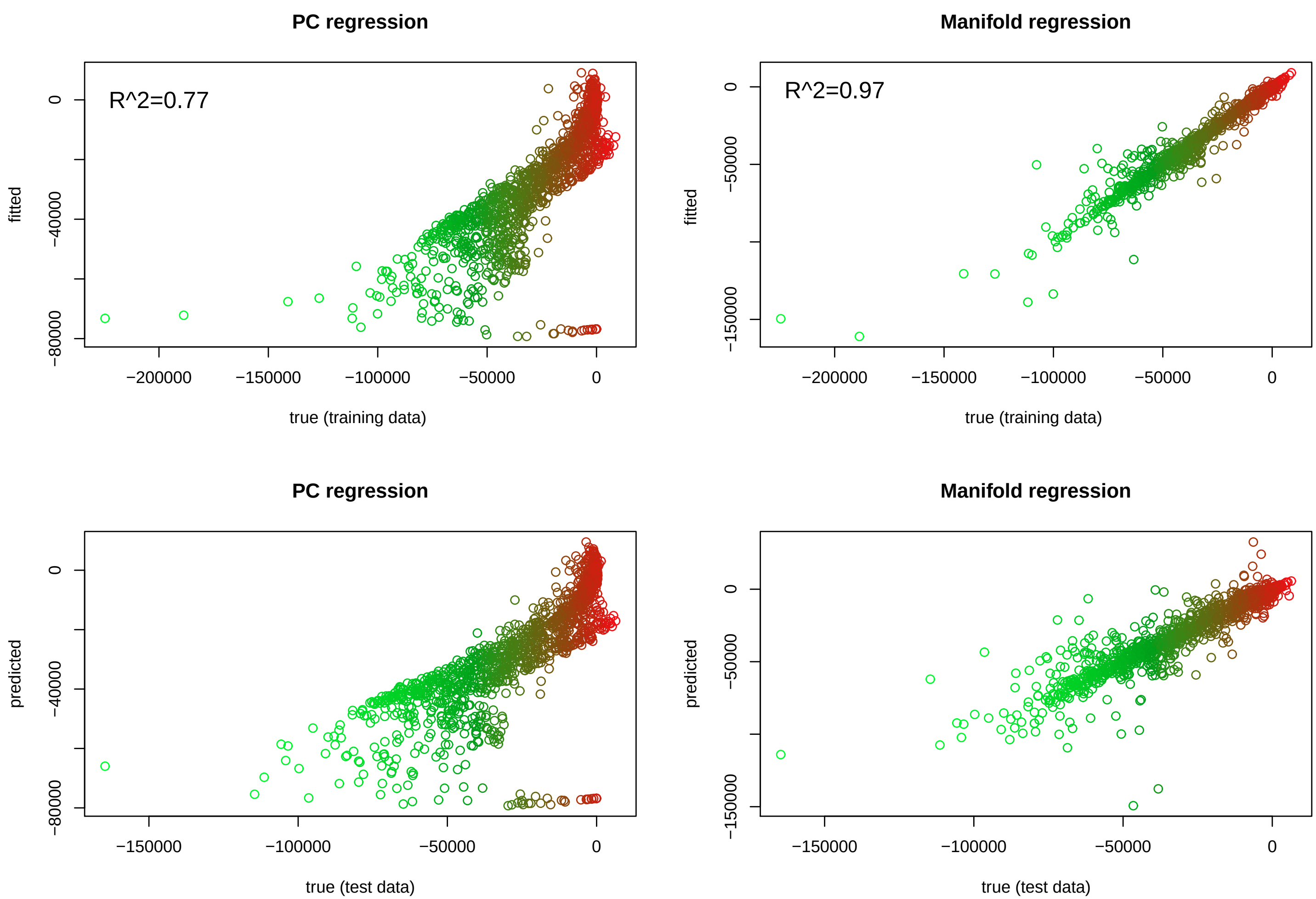
- Local principal surface, with data coloured by (true, tabulated) PC source terms s_i (left); after regression onto principal surface (right).





Model validation

- Fitted versus true response for 4000 training data (top) and 4000 test data (bottom), using PC regression (left) and surface regression (right):



Model comparison

For comparison, we consider a wider range of regression methods:

- Traditional methods:
  - Linear (principal component) regression:
$$s_i = \beta_0 + \beta_1 \mathsf{PC}_{1,i} + \beta_2 \mathsf{PC}_{2,i} + \beta_3 \mathsf{PC}_{3,i} + \epsilon_i$$
  - Additive models:
$$s_i = f_1(\mathsf{PC}_{1,i}) + f_2(\mathsf{PC}_{2,i}) + f_3(\mathsf{PC}_{3,i}) + \epsilon_i$$

Model comparison

For comparison, we consider a wider range of regression methods:

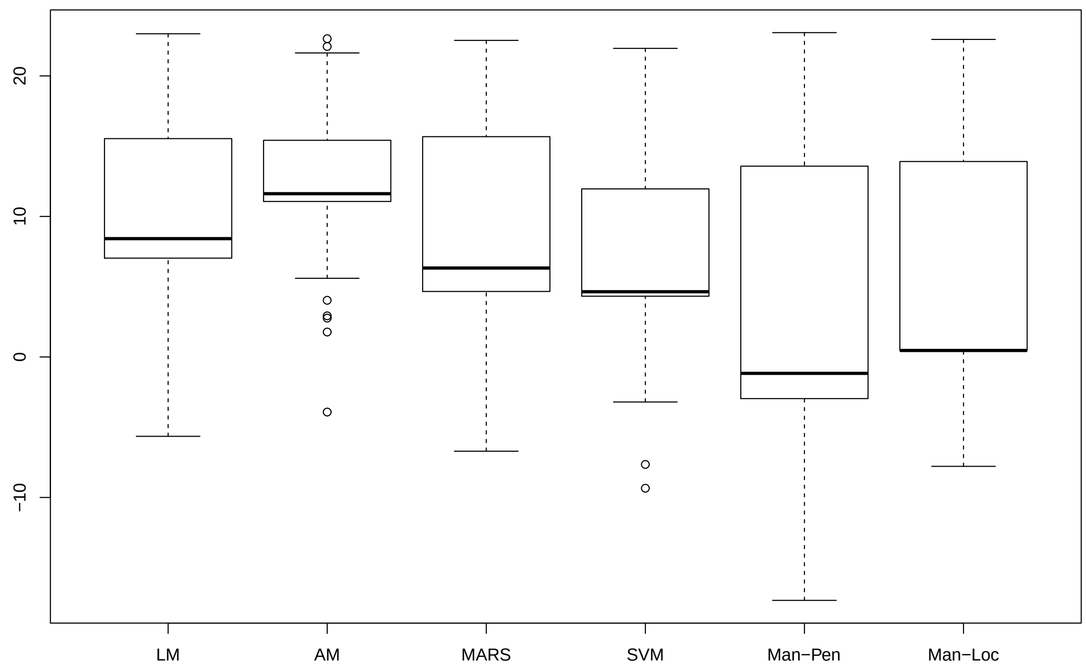
- Traditional methods:
  - Linear (principal component) regression:
  $$s_i = \beta_0 + \beta_1 \mathsf{PC}_{1,i} + \beta_2 \mathsf{PC}_{2,i} + \beta_3 \mathsf{PC}_{3,i} + \epsilon_i$$
  - Additive models:
  $$s_i = f_1(\mathsf{PC}_{1,i}) + f_2(\mathsf{PC}_{2,i}) + f_3(\mathsf{PC}_{3,i}) + \epsilon_i$$
- Modern "black–box" methods:
  - Multivariate adaptive regression splines (MARS);
  - Support vector machine (SVM);
  - Penalized principal–surface–based regression (as explained).
  - Localized principal–surface–based regression (Einbeck, Evers & Powell, 2010).

Model comparison (cont'd)

- Boxplots of test data residuals,

$$\log((s_i - \hat{s}_i)^2),$$

for all six regression techniques:

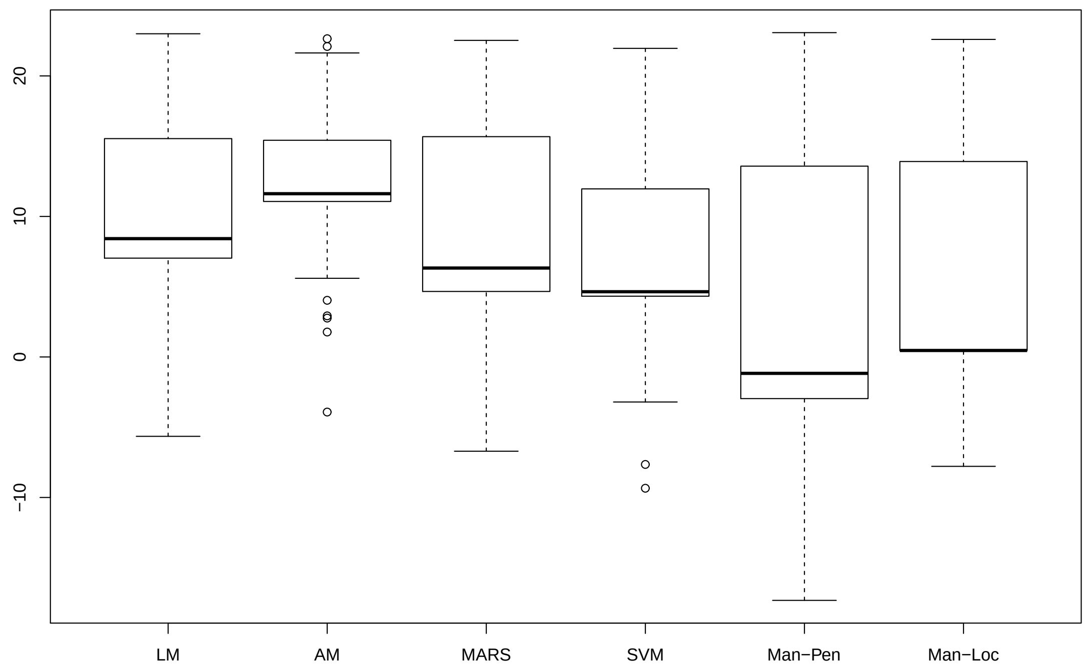


Model comparison (cont'd)

- Boxplots of test data residuals,

$$\log((s_i - \hat{s}_i)^2),$$

for all six regression techniques:



- Clear evidence in favour of the manifold.

Conclusion

- Principal curves and surfaces form a powerful tool for compressing high–dimensional non–linear data structures...
- ...which can be used as a building block for further statistical procedures (such as, nonparametric regression).
- Technique extends to manifolds of higher dimension by considering tetrahedrons ($d = 3$) or simplices ($d \geq 4$).
- Open problems:
  - We don't have yet a (reliable) tool to determine the 'right' intrinsic dimension.
  - In higher dimensions, it is hard to judge whether the fitted surface or manifold is 'good'.
  - Automated smoothing parameter selection only available for principal curves.
- Software: R package **LPCM** for principal curves (on CRAN); and **lpmforge** for principal manifolds (L. Evers, unpublished).

References

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

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

Einbeck & Dwyer (2011): Using principal curves to analyse traffic patterns on freeways. *Transportmetrica* **7**, 229–246.

Einbeck & Zayed (2013). Some asymptotics for localized principal components and curves. *Communications in Statistics – T&M*, doi:10.1080/03610926.2012.673676.

Evers & Einbeck (2013): Local principal manifolds. *Working paper, unpublished*.

Sutherland & Parente (2009): Combustion modeling using principal component analysis. *Proceedings of the Combustion Institute* **32**, 1563–1570.