Penalized regression on principal manifolds with application to combustion modelling

Jochen Einbeck

Department of Mathematical Sciences, Durham University

Prague, 18th of July 2012

Durham University
Durham Energy Institute

in collaboration with:

Ludger Evers (University of Glasgow)

Alessandro Parente (University Libré de Bruxelles)

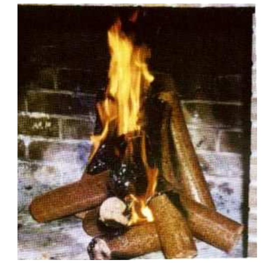
Ben Isaac (University of Utah)

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

$$2H_2 + O_2 \longrightarrow 2H_2 0$$

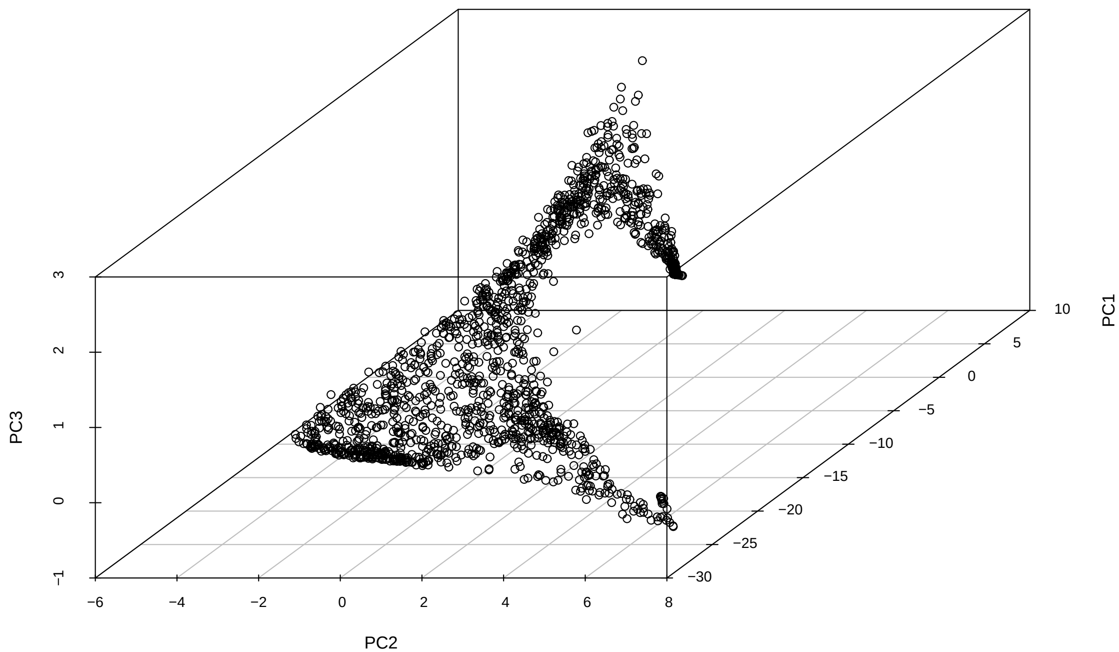
Combustion



- 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.

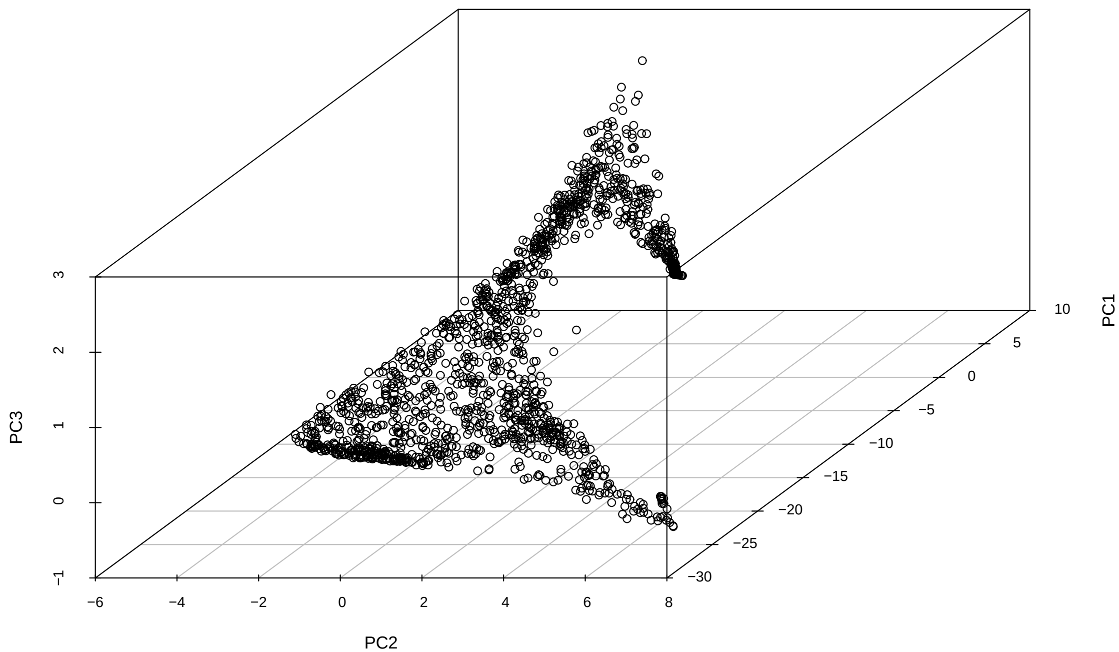
Combustion data

- 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 ($n = 4000$):



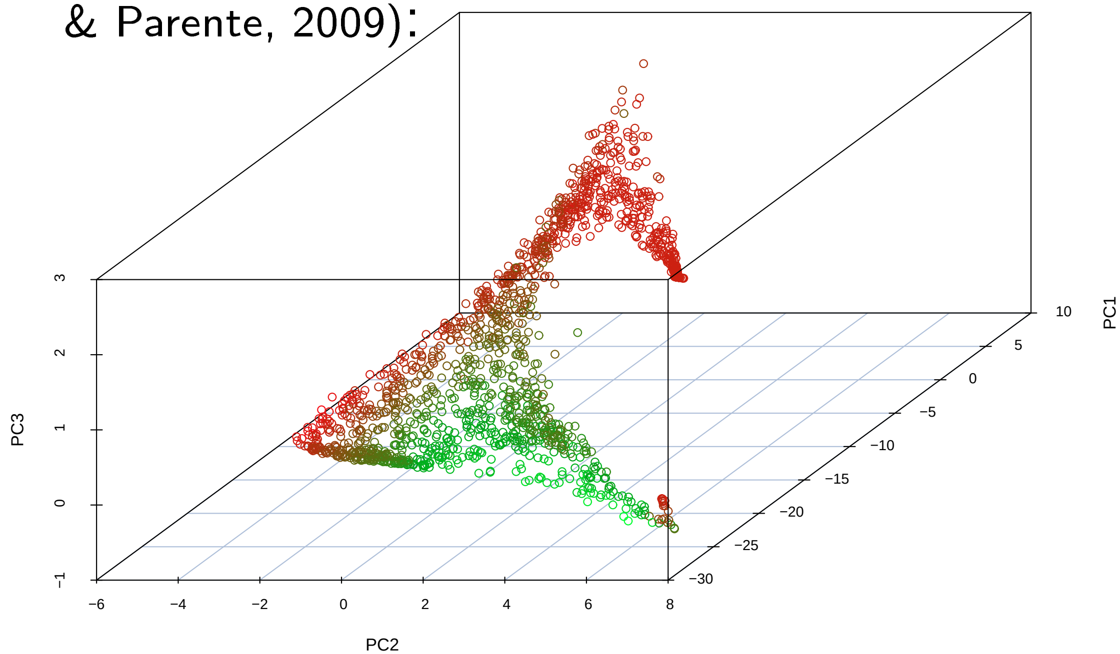
- It is well-known that the thermochemical state space of combustion systems resides on low–dimensional manifolds.
- This is convenient, as the transport equations based on the reduced system of, say, 3 principal components *are* tractable.

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

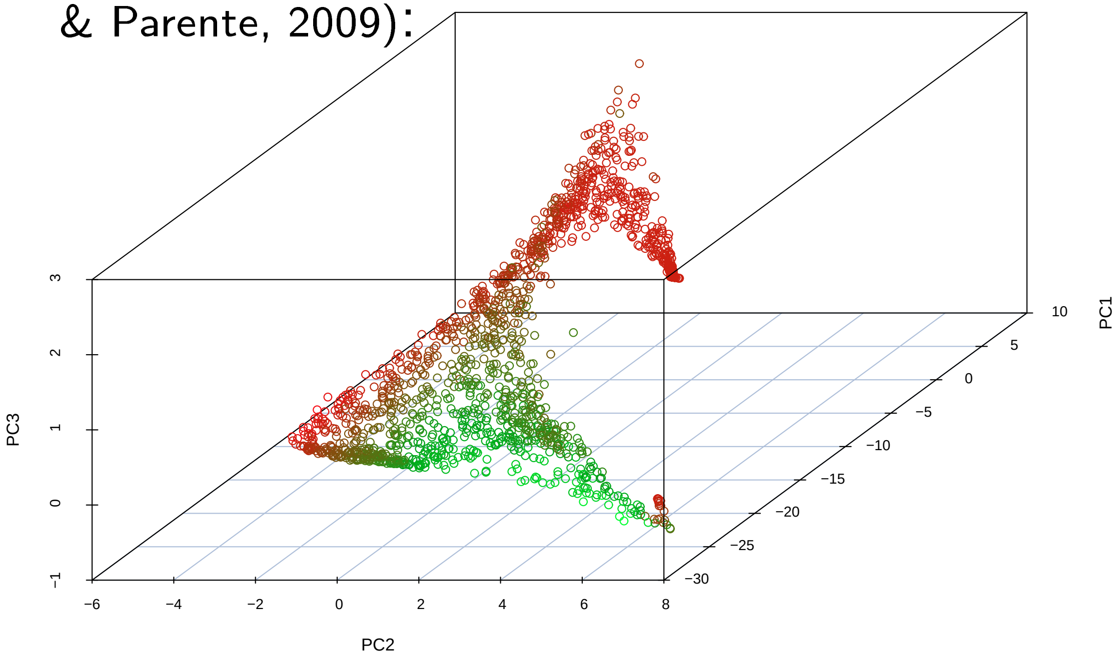
- Complication: The rates of production ('source terms') of the principal components are unknown.
- In practice, they have to be found by regression on the principal components.
- Requires 'high–fidelity' data with tabulated source terms (Sutherland & Parente, 2009):



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 manifold is informative for the source terms.

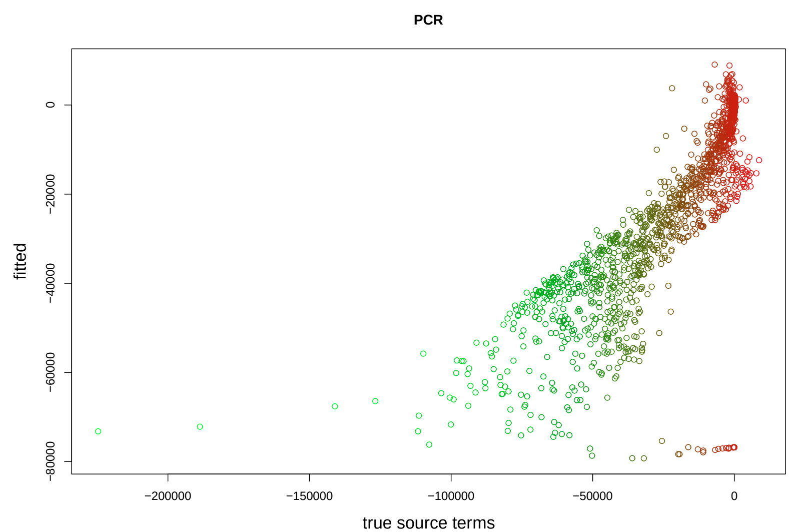
Principal component regression

- 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:

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

(Sutherland & Parente, 2009).

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



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

Principal manifolds

- Can we make use of the manifold structure more explicitly?
- Requires data approximation via principal manifolds (in 2D: principal surfaces).

Principal manifolds

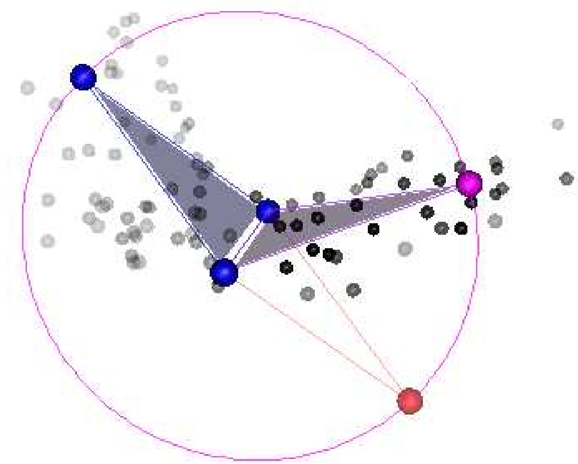
- Can we make use of the manifold structure more explicitly?
- Requires data approximation via principal manifolds (in 2D: principal surfaces).
- Local principal surfaces, using triangles as building blocks (Einbeck & Evers, 2010):

Starting from an initial triangle, iteratively . . .

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

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

. . . until all triangles have been considered.



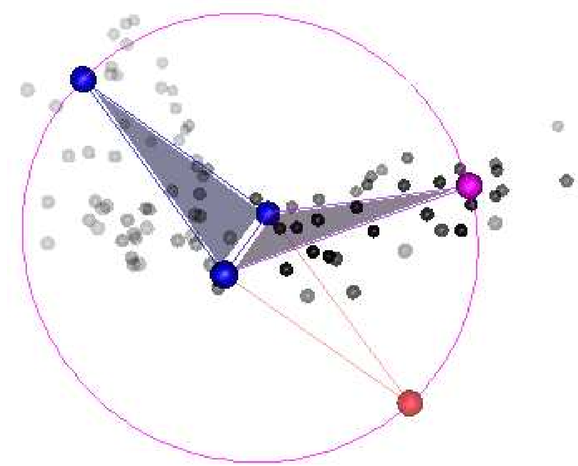
Principal manifolds

- Can we make use of the manifold structure more explicitly?
- Requires data approximation via principal manifolds (in 2D: principal surfaces).
- Local principal surfaces, using triangles as building blocks (Einbeck & Evers, 2010):

Starting from an initial triangle, iteratively . . .

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

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

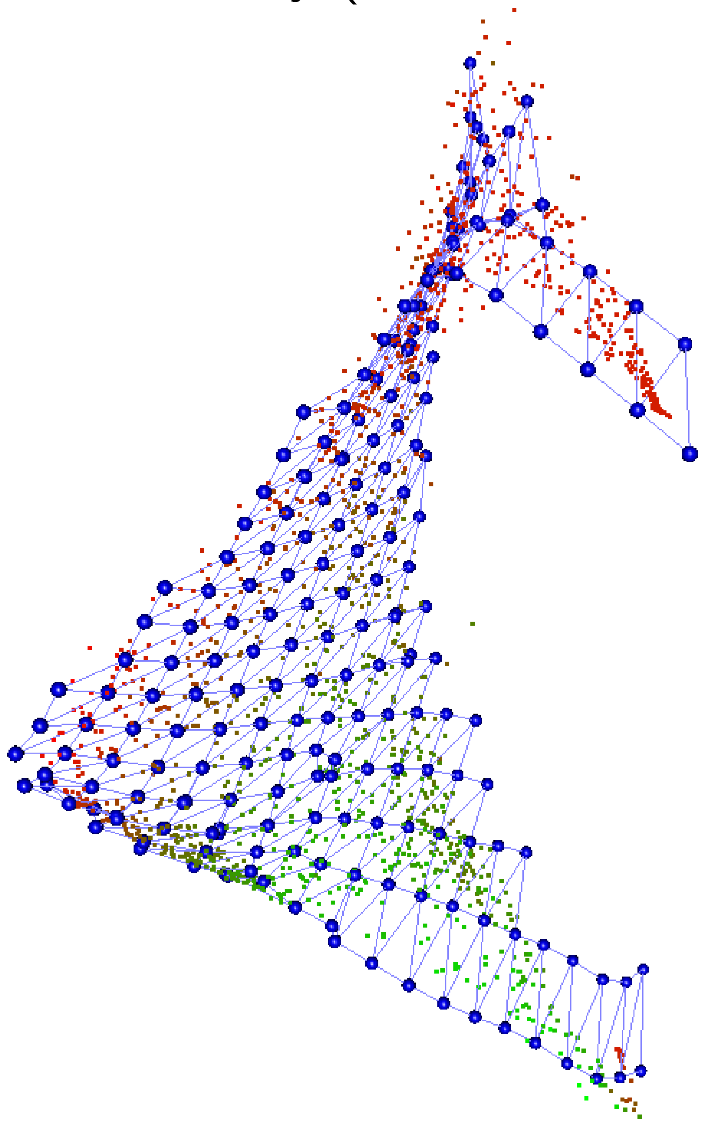


. . . until all triangles have been considered.

- Extends to principal manifolds of any dimension when replacing triangles (2D) by tetrahedrons (3D) or simplices (>3D).

Principal manifolds (cont'd)

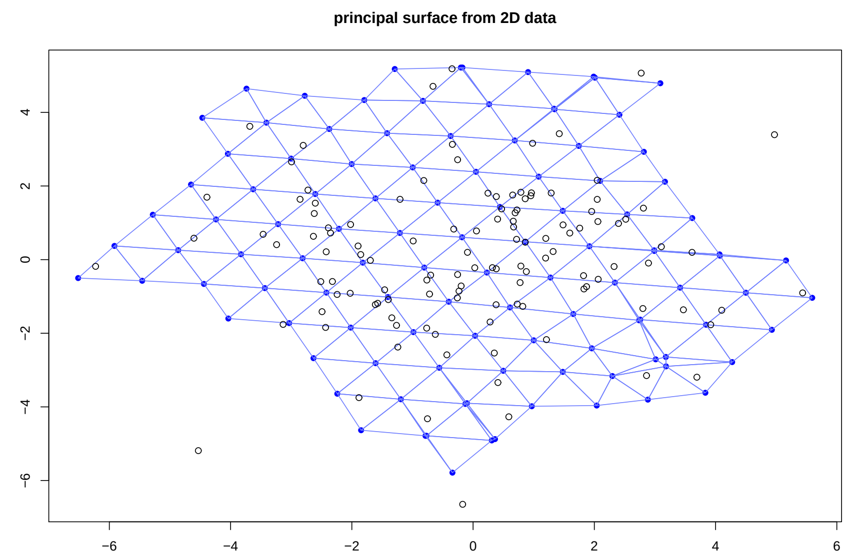
- Fitted local principal surface to combustion data, with data couloured by (true, tabulated) PC source terms:



- Neat . . .
- . . . but the actual challenge is to regress the source terms onto the surface.

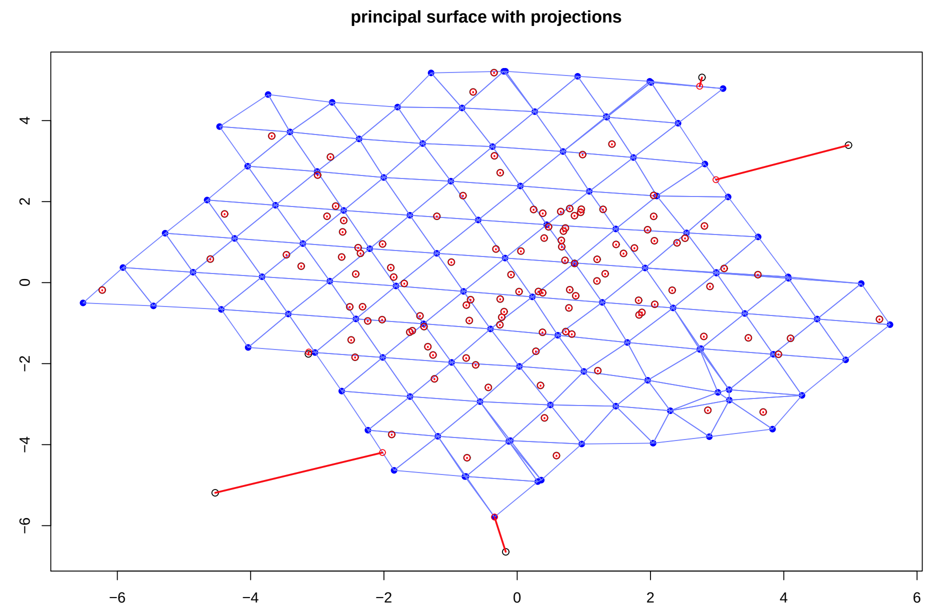
Regression on principal manifolds

- Toy example: A principal surface for bivariate data.



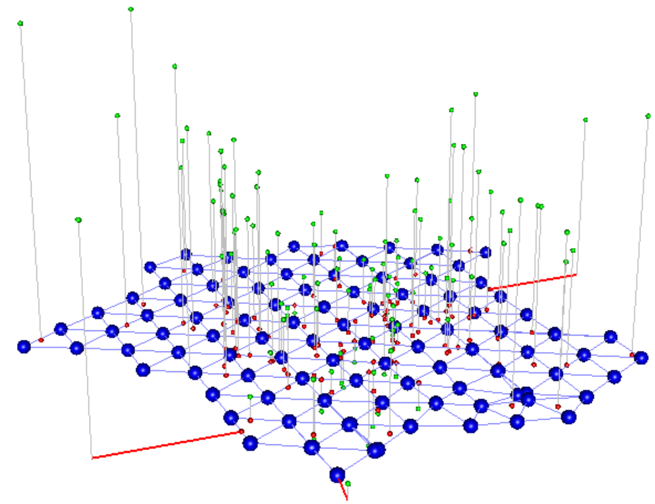
Regression on principal manifolds

- 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 manifolds

- 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.



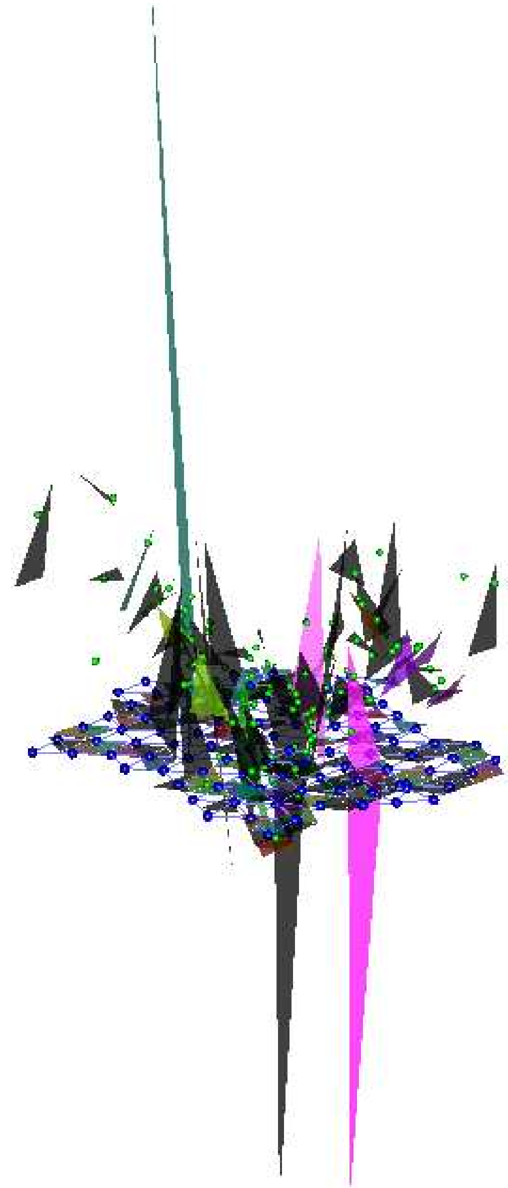
- Next, consider a response y_i.
- Assume separate regression models for each triangle j

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

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

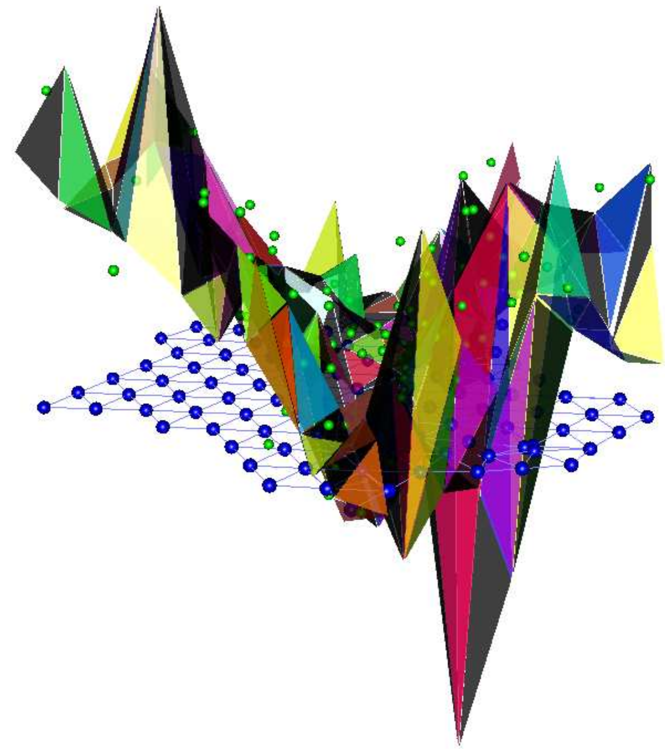
Penalized regression

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



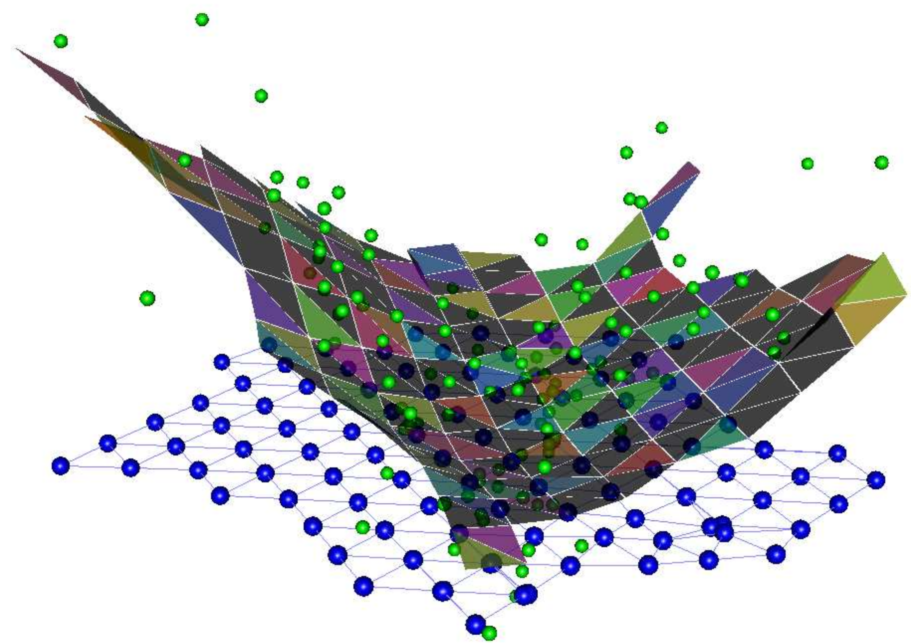
Penalized regression

- 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 $\boldsymbol{\beta}' = \left(\boldsymbol{\beta}'_{(1)}, \boldsymbol{\beta}'_{(2)}, \dots\right)$,
  - 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

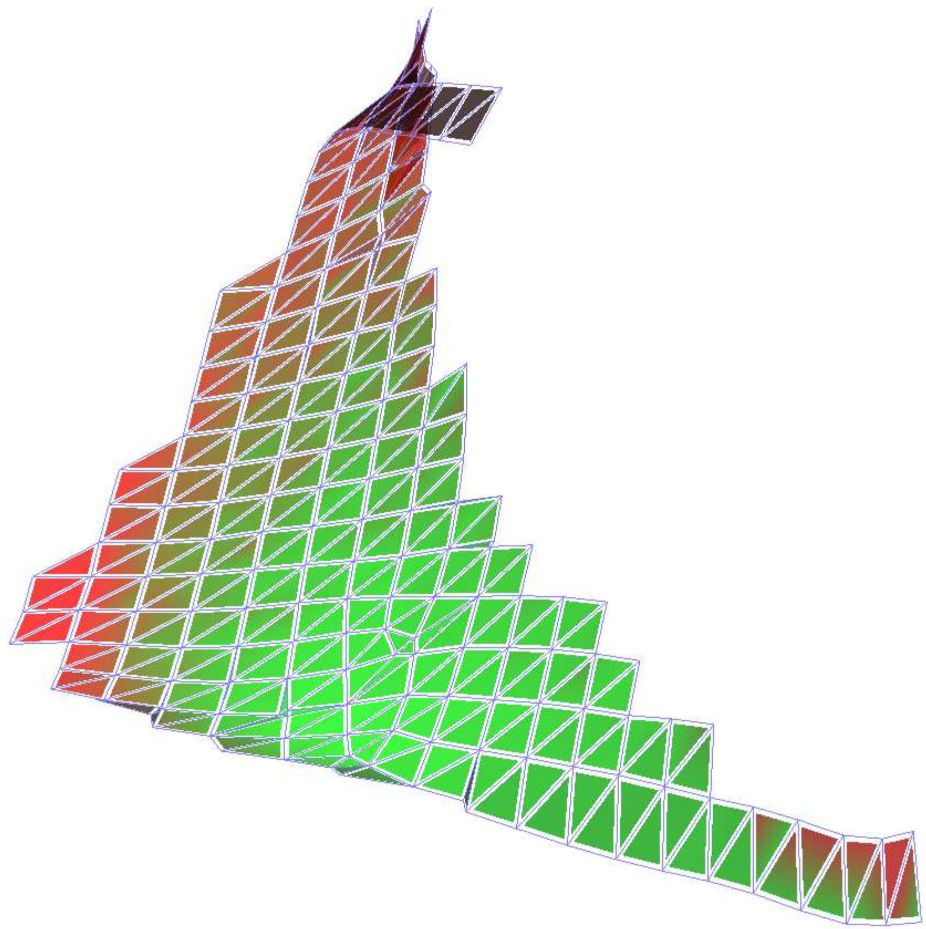
$$\|\mathbf{Z}\boldsymbol{\beta} - \mathbf{y}\|^2 + \lambda\|\mathbf{D}\boldsymbol{\beta}\|^2 + \mu\|\mathbf{E}\boldsymbol{\beta}\|^2. \tag{1}$$

- 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

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

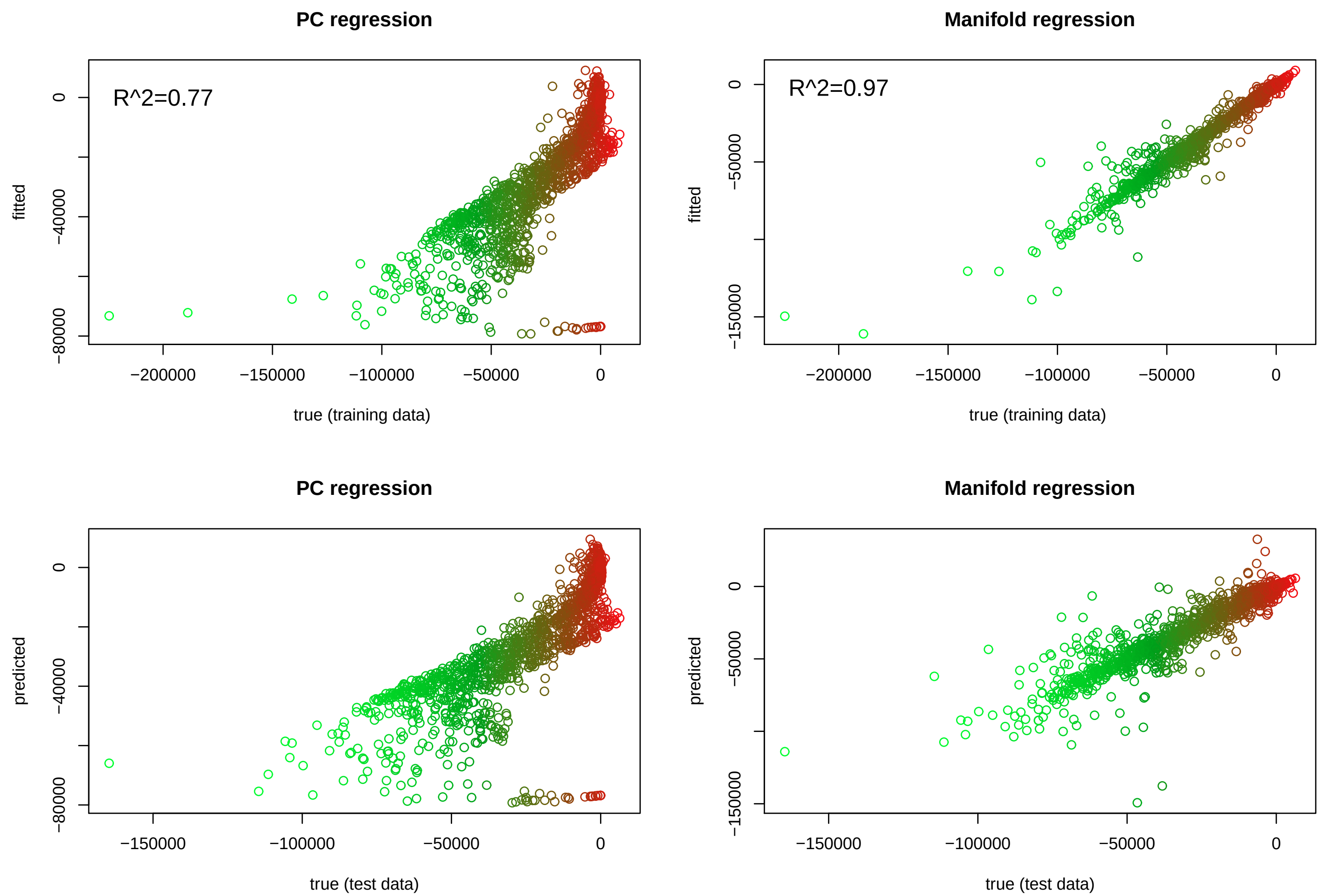
Back to combustion problem

- Using this technique, the source terms $s_i, i = 1, \ldots, n$ are regressed onto the principal surface.



Simulation study

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



Simulation study (cont'd)

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:
  $$s_i = f_1(\mathsf{PC}_{1,i}) + f_2(\mathsf{PC}_{2,i}) + f_3(\mathsf{PC}_{3,i}) + \epsilon_i$$

Simulation study (cont'd)

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

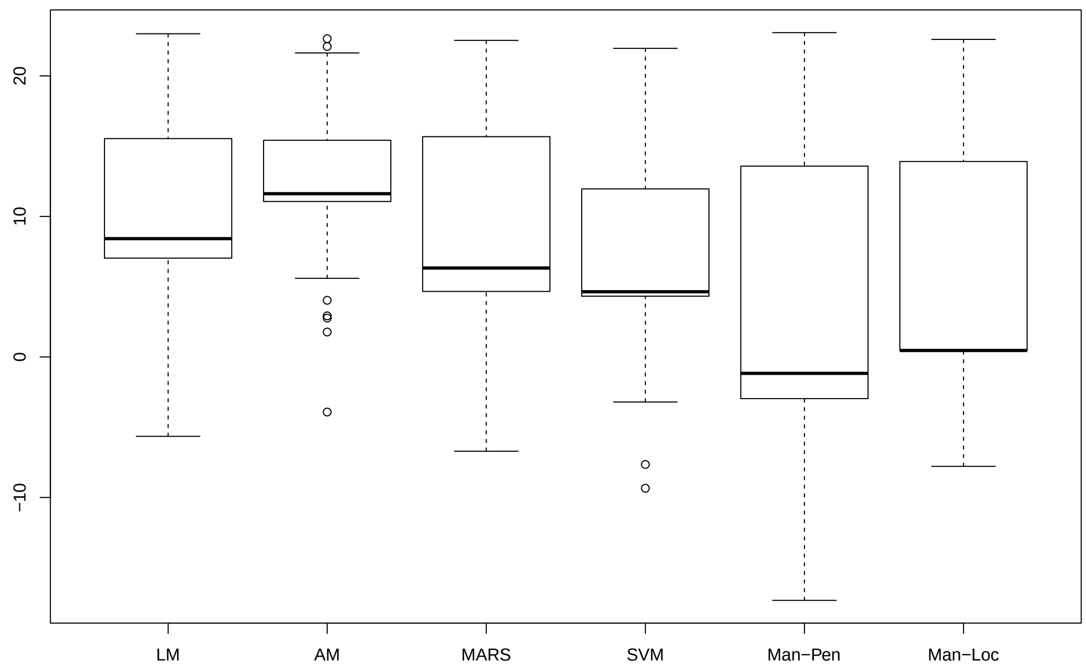
- Traditional methods:
  - Linear (principal component) regression:
    $$s_i = \beta_0 + \beta_1 \mathrm{PC}_{1,i} + \beta_2 \mathrm{PC}_{2,i} + \beta_3 \mathrm{PC}_{3,i} + \epsilon_i$$
  - Additive models:
    $$s_i = f_1(\mathrm{PC}_{1,i}) + f_2(\mathrm{PC}_{2,i}) + f_3(\mathrm{PC}_{3,i}) + \epsilon_i$$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

Simulation study (cont'd)

- Boxplots of test data residuals,

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

for all six regression techniques:

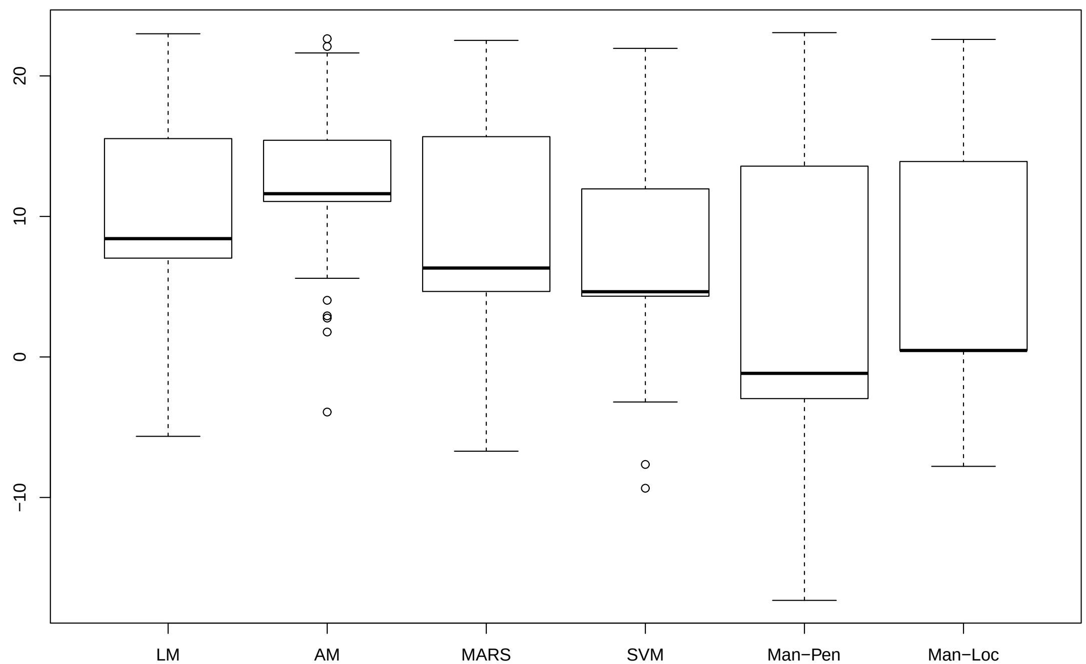


Simulation study (cont'd)

- 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

- For the combustion problem, the estimation of source terms is one of a series of steps towards the construction of a practical combustion model (for Direct Numerical Simulation, etc).
- The next step is the numerical solution of the reduced set of transport equations.
- Results depend on type of scaling before PCA (Isaac et al, 2012).
- Our predictions tend to give excellent results for most of the predictor space, but quite 'bad' results for a few small subregions (usually at manifold tails and boundaries). In our application, those 'bad' predictions could be traced back to the burn–in–process.
- Other applications of principal manifolds in: astrophysics, neuroimaging, particle physics, oceanography, ...
- Working paper (Evers & Einbeck, 2012) and R package (**lpmforge**) available on request.

References

Einbeck & Evers (2010): Localized regression on principal manifolds. *Proc' of the 25th International Workshop on Statistical Modelling*, University of Glasgow, pp 179–184.

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

Evers & Einbeck (2012): Local principal manifolds. *Working paper, unpublished*.

Isaac, Parente, Einbeck, Evers, Sutherland, Thornock & Smith (2012): Principal component conservation equations: source terms. *Working paper, unpublished*.

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