Dimension reduction for high dimensional regression problems based on local principal curves

Jochen Einbeck

jochen.einbeck@durham.ac.uk

ERCIM, Neuchâtel, 21th of June 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.
Simulated GAIA data

- Photon counts are simulated from APs
- AP Design space:
Simulated GA\|A data (cont.)

Photon counts simulated from APs through computer models:
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?

For instance, temperature as response:

```r
> gaia.lm <- lm(temperature ~ spec1 + spec2 + ... + spec16, data= gaia)
```
> 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    |
| spec10  |          |            |         |          |
| spec11  |          |            |         |          |
| spec12  |          |            |         |          |
| spec13  |          |            |         |          |
| spec14  |          |            |         |          |
| spec15  |          |            |         |          |
| spec16  | 13886697  | 21106076  | 0.658  | 0.511    |

Residual standard error: 1978 on 983 degrees of freedom

All variables are insignificant.
Dimension reduction

- Usual remedies:
  - Model/ variable selection procedures
  - Dimension reduction techniques
- The second one is obviously the more promising here.
- Look at scree plot:

<table>
<thead>
<tr>
<th>Comp. 1</th>
<th>Comp. 2</th>
<th>Comp. 3</th>
<th>Comp. 4</th>
<th>Comp. 5</th>
<th>Comp. 6</th>
<th>Comp. 7</th>
<th>Comp. 8</th>
<th>Comp. 9</th>
<th>Comp. 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0000</td>
<td>0.0002</td>
<td>0.0004</td>
<td>0.0006</td>
<td>0.0008</td>
<td>0.0010</td>
<td>0.0012</td>
<td>0.0014</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

Fit temperature against the first three PC scores:

```
gaiapclm <- 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 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.
Hence, the following is to do:

1. Estimate the smooth curve capturing the structure of the (three-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 $X = (X_1, \ldots, X_n)$, where $X_i = (X_{i1}, \ldots, X_{id})$.

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 $1^{st}$ local eigenvector $\gamma^x$ of $\Sigma^x = (\sigma^x_{jk})_{(1 \leq j, k \leq d)}$, where
   $$\sigma^x_{jk} = \sum_{i=1}^{n} w_i (X_{ij} - \mu^x_j)(X_{ik} - \mu^x_k).$$

4. Step from $\mu^x$ to $x := \mu^x + t_0 \gamma^x_1$.

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).
**Simpler example: Speed-flow data**

- Data recorded on the Californian Freeway FR57-N on 9th of July 2007:

![Scatter plot of speed vs. flow]

(data from: PemS)

- The red squares correspond to the points $\mu^x$ making up the LPC.
Example: Speed-flow data

- Technical question: How to connect the points?
- For descriptive purposes, a linear interpolation is sufficient.
- If one wants to compute projections onto it, a cubic spline can be laid through the curve:

![Speed-flow data graph](image-url)
Parametrization

Next, the origin is fixed to $t = 0$ at one of the two ends. We consider the LPC as a curve

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

and the parameterization along the curve is defined through the arc length 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 representant in a regression problem involving speed and flow as covariates.
Projections
Local principal curves seem to capture the structure of the data cloud adequately.

If this curve is to be used for regression, then:
- Fit the spline through the 3-dim LPC points
- Compute projection indices $t_i$ of all data points onto this spline
LPC Regression with GAIA data

- Assume we want to predict stellar temperature from spectral data.
- This is now a simple one-dimensional regression problem

\[ y_i = f(t_i) + \varepsilon_i \]
LPC Regression with GAIA data

- Assume we want to predict stellar temperature from spectral data.
- This is now a simple one-dimensional regression problem:

\[ y_i = f(t_i) + \varepsilon_i \]

![Graph showing smoothing spline regression](image.png)
Is there a shortcut?

LPC fitted *directly* through 16-dimensional space:
Direct data compression with LPCs

Zoom into 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 important patterns of the predictor space and its performance will depend on the choice of the starting point.
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 ($\times 10^3$) for 200 observations sampled from the training data set:

<table>
<thead>
<tr>
<th></th>
<th>LM</th>
<th>PC+Regr</th>
<th>LPC+Regr</th>
<th>PC+LPC+Regr</th>
</tr>
</thead>
<tbody>
<tr>
<td>average($\hat{\epsilon}_i^2$)</td>
<td>4,119</td>
<td>4,395</td>
<td>2,215</td>
<td>2,633</td>
</tr>
<tr>
<td>median($\hat{\epsilon}_i^2$)</td>
<td>1,035</td>
<td>1,300</td>
<td>66</td>
<td>51</td>
</tr>
</tbody>
</table>

where $\hat{\epsilon}_i$ is the difference between true and predicted temperature.
Take care with boundaries!

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

- Close to the endpoints of the LPC, all points in that cluster will be
Robustness issues

- LPCs are by construction resistant to outlying data patterns, if the starting point is fixed.

- However, resistance is not necessarily a desired feature when exploring the covariate space. Here, it is important that all regions of it are covered, even if outlying.

- If the data pattern is not connected but scattered in space, one needs multiple/branched/disconnected LPCs to describe it.
  - Technically possible (Einbeck, Tutz, Evers, 2005b), but yet to incorporate into regression framework.....

- An important issue is robustness to the choice of the starting point, which is the more critical the more sparse the data are.
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

