Some asymptotics for localized principal components and curves

Jochen Einbeck

Department of Mathematical Sciences, Durham University

jochen.einbeck@durham.ac.uk

joint work with Mohammad Zayed (Mansoura University, Egypt),

Budapest, 20 July 2013

Durham University

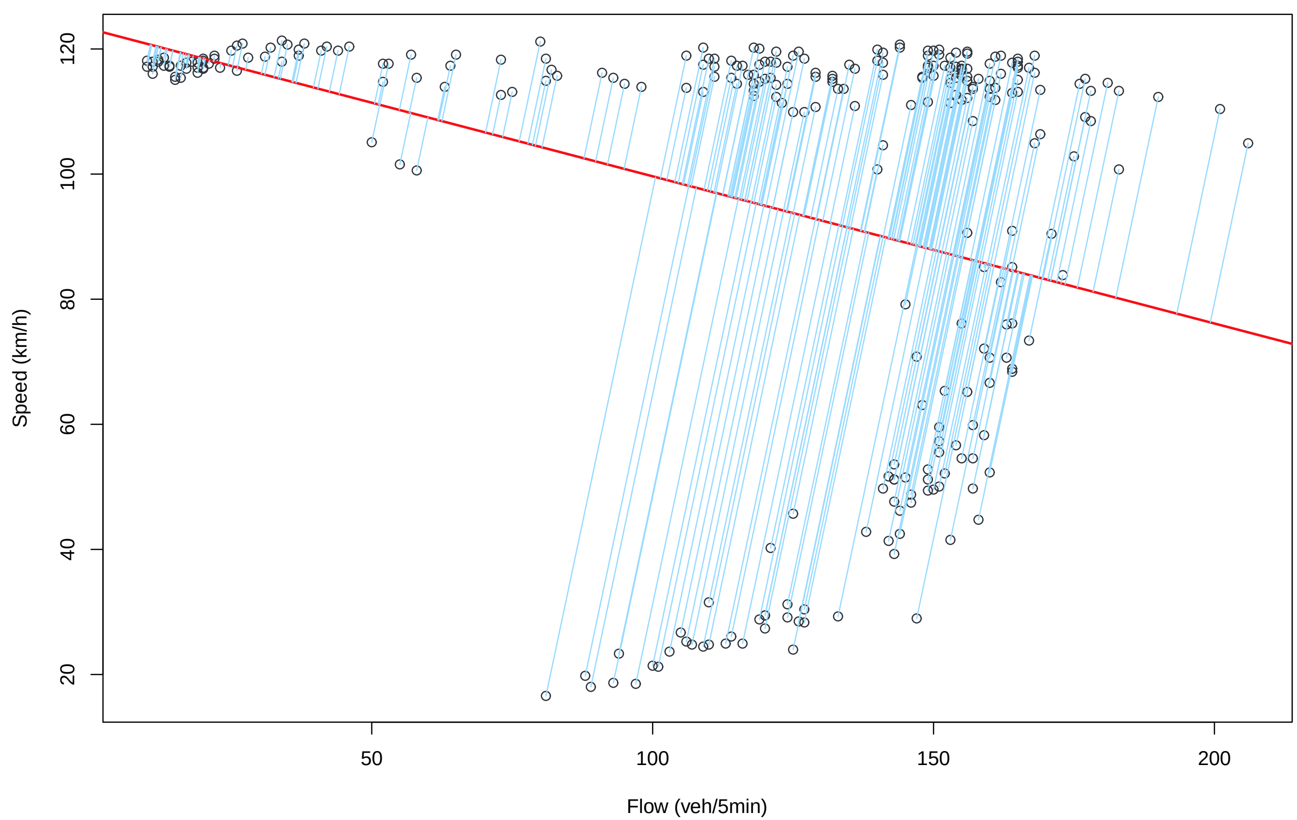
Principal components

The first principal component provides the best fit to a multivariate data cloud. It minimizes the sum of squared distances between the data points and their orthogonal projections onto the line.

Principal components

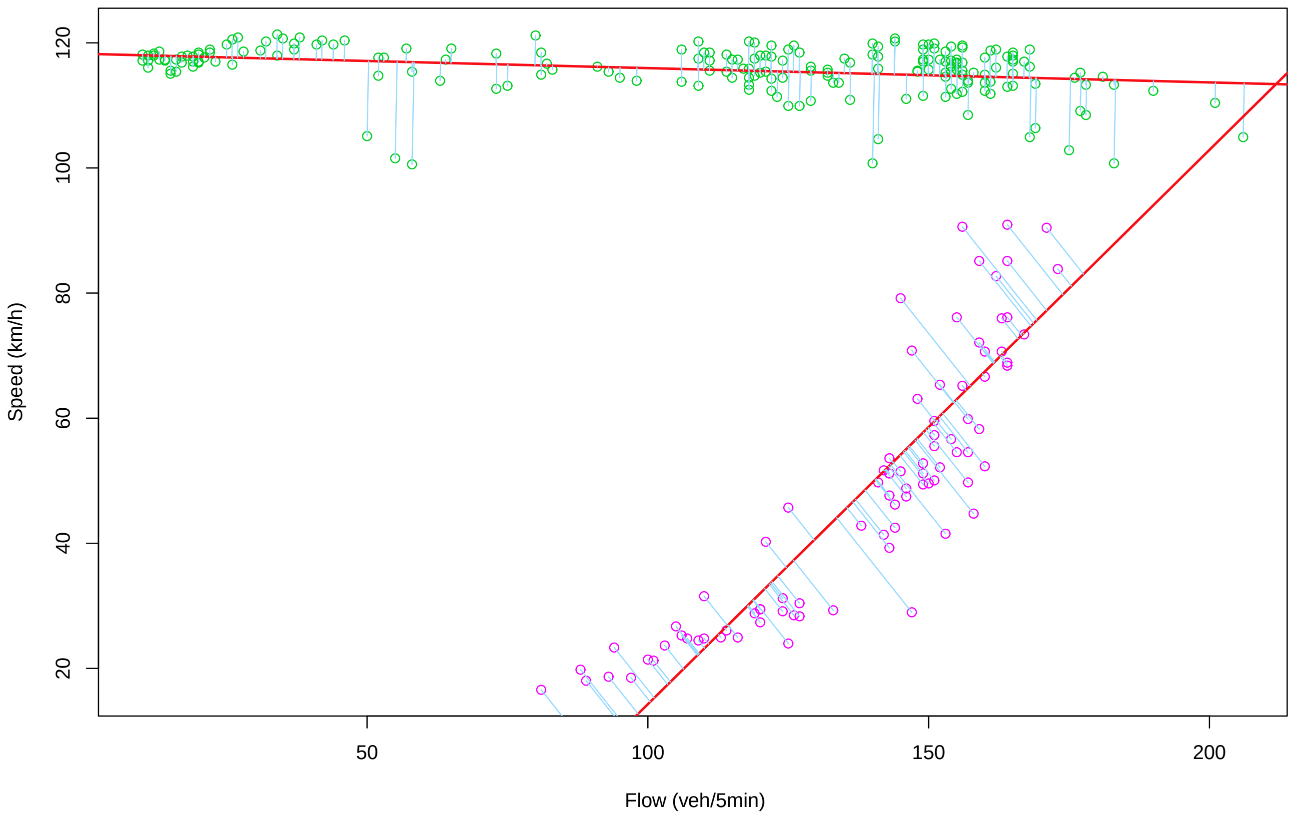
The first principal component provides the best fit to a multivariate data cloud. It minimizes the sum of squared distances between the data points and their orthogonal projections onto the line.

Example: Speed-Flow data from a Californian “freeway”.



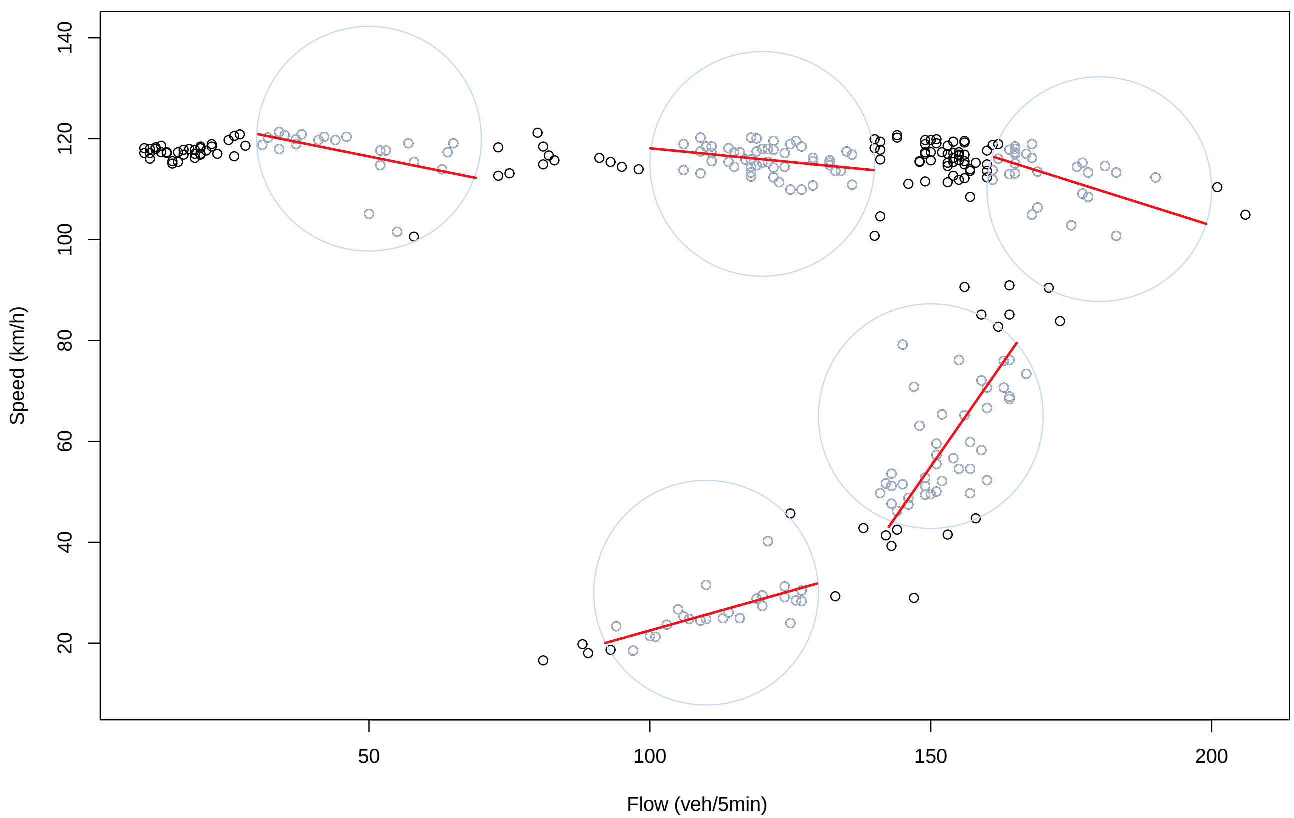
Localized principal components

Principal components can be computed over certain subregions ...



Localized principal components

Principal components can be computed over certain subregions or local neighborhoods.



Localized principal component analysis

- Given: Data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ sampled from a d–variate random vector $\boldsymbol{X}$ with density $f : \mathbb{R}^d \longrightarrow \mathbb{R}$.
- At a target point $\boldsymbol{x}$, we wish to find a line $g(s) = \boldsymbol{m} + s\boldsymbol{\gamma}$ which "locally gives the best fit".

Localized principal component analysis

- Given: Data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ sampled from a d–variate random vector $\boldsymbol{X}$ with density $f : \mathbb{R}^d \longrightarrow \mathbb{R}$.
- At a target point $\boldsymbol{x}$, we wish to find a line $g(s) = \boldsymbol{m} + s\boldsymbol{\gamma}$ which "locally gives the best fit".
- This leads to the concept of *locally weighted PCA*, i.e., we wish to minimize

$$Q(\boldsymbol{m}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} w^{\boldsymbol{x}}(\boldsymbol{x}_i)||\boldsymbol{x}_i - \boldsymbol{x}_i'||^2 - \lambda(\boldsymbol{\gamma}^T\boldsymbol{\gamma} - 1)$$

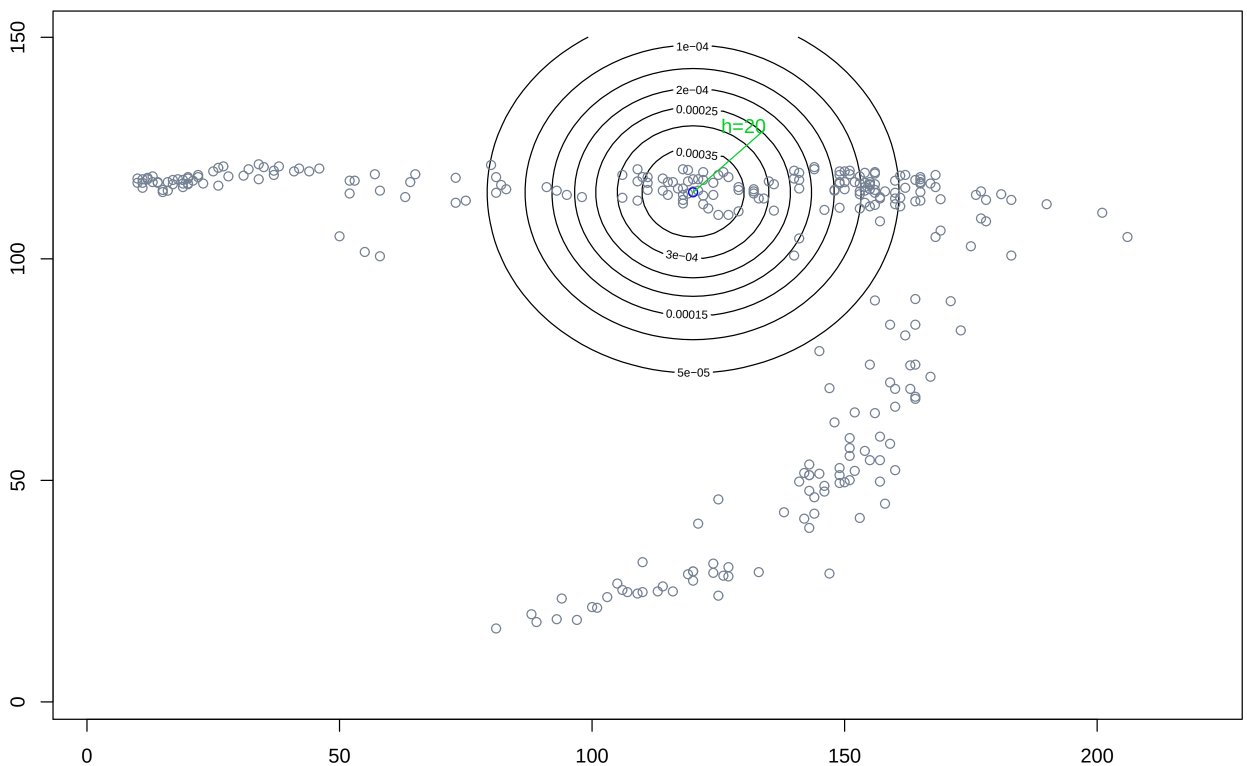
w.r.t. $\boldsymbol{m}$ and $\boldsymbol{\gamma}$, where $\boldsymbol{x}_i'$ is the projection of $\boldsymbol{x}_i$ onto the line through $\boldsymbol{m}$ with direction $\boldsymbol{\gamma}$, and with kernel weights

$$w^{\boldsymbol{x}}(\boldsymbol{x}_i) = K_{\boldsymbol{H}}\left(\boldsymbol{x}_i - \boldsymbol{x}\right)$$

centered at the target point $\boldsymbol{x}$.

Kernel weights

- We use $K_{\boldsymbol{H}}(\boldsymbol{x}) \equiv K_{\boldsymbol{H}}(x_1, \dots x_d) = \prod_{j=1}^{d} N(x_j, h_j^2)$ which implies $\boldsymbol{H} = \text{diag}(h_1^2, \dots h_d^2)$.
- For instance, for estimation at $\boldsymbol{x} = (120, 115)^T$, with $h \equiv h_1 = h_2 = 20$, one has the weight diagram



Solving localized PCA

- Writing the projections as

$$\boldsymbol{x}_i' = \boldsymbol{m} + \boldsymbol{\gamma}\boldsymbol{\gamma}^T(\boldsymbol{x}_i - \boldsymbol{m}) = (\boldsymbol{I} - \boldsymbol{\gamma}\boldsymbol{\gamma}^T)\boldsymbol{m} + \boldsymbol{\gamma}\boldsymbol{\gamma}^T\boldsymbol{x}_i,$$

the minimization problem takes the shape

$$Q(\boldsymbol{m},\boldsymbol{\gamma}) = \sum_{i=1}^{n} w^{\boldsymbol{x}}(\boldsymbol{x}_i)(\boldsymbol{x}_i - \boldsymbol{m})^T(\boldsymbol{I} - \boldsymbol{\gamma}\boldsymbol{\gamma}^T)(\boldsymbol{x}_i - \boldsymbol{m}) - \lambda(\boldsymbol{\gamma}^T\boldsymbol{\gamma} - 1)$$

- By taking partial derivatives of $Q(\boldsymbol{m},\boldsymbol{\gamma})$, one finds

$$\boldsymbol{m} = \sum_{i=1}^{n} w^{\boldsymbol{x}}(\boldsymbol{x}_i)\boldsymbol{x}_i / \sum_{i=1}^{n} w^{\boldsymbol{x}}(\boldsymbol{x}_i).$$

and

$$\boldsymbol{\Sigma}^{\boldsymbol{x}}\boldsymbol{\gamma} = -\lambda\boldsymbol{\gamma},$$

where $\boldsymbol{\Sigma}^{\boldsymbol{x}} = \sum_{i=1}^{n} w^{\boldsymbol{x}}(\boldsymbol{x}_i)(\boldsymbol{x}_i - \boldsymbol{m})(\boldsymbol{x}_i - \boldsymbol{m})^T / \sum_{i=1}^{n} w^{\boldsymbol{x}}(\boldsymbol{x}_i)$.

Asymptotics for localized PCA

- We are interested in the behavior of localized PCA for large n and small $\boldsymbol{H} = \text{diag}(h_1^2, \ldots, h_d^2)$.
- Theory developed in another context (Ruppert and Wand, 1994) gives us directly

$$\boldsymbol{m} = \boldsymbol{x} + \mu_2(K)\boldsymbol{H}\boldsymbol{\nabla} f(\boldsymbol{x})/f(\boldsymbol{x}) + o_p(\boldsymbol{H}\boldsymbol{1})$$

- Find $\boldsymbol{\gamma}$ minimizing the expected minimization problem

$$\mathbb{E}_f Q(\boldsymbol{m}, \boldsymbol{\gamma}) = nf(\boldsymbol{x})(\boldsymbol{x} - \boldsymbol{m})^T(\boldsymbol{I} - \boldsymbol{\gamma}\boldsymbol{\gamma}^T)(\boldsymbol{x} - \boldsymbol{m}) - \lambda(\boldsymbol{\gamma}^T\boldsymbol{\gamma} - 1),$$

yielding

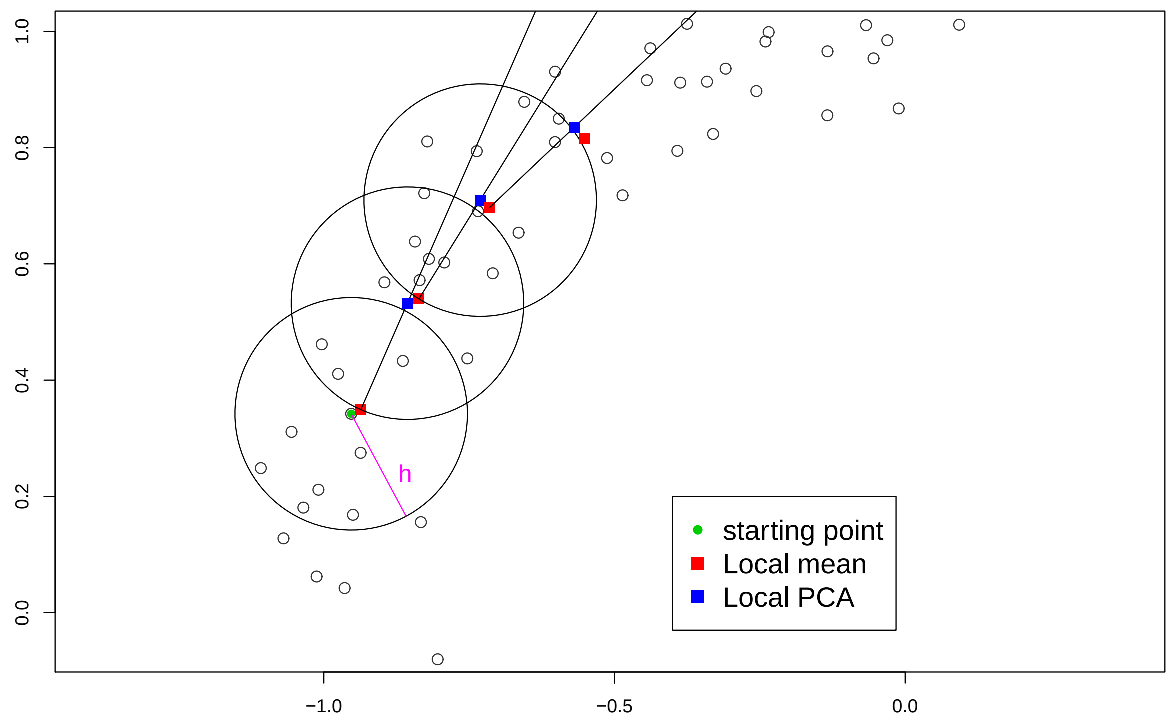
$$\boldsymbol{\gamma} \stackrel{a}{=} -\frac{\boldsymbol{H}\boldsymbol{\nabla} f(\boldsymbol{x})}{||\boldsymbol{H}\boldsymbol{\nabla} f(\boldsymbol{x})||}.$$

Local principal curves (LPC)

- Goal: Find a smooth curve through the “middle of the data” (like a “nonparametric principal component”)

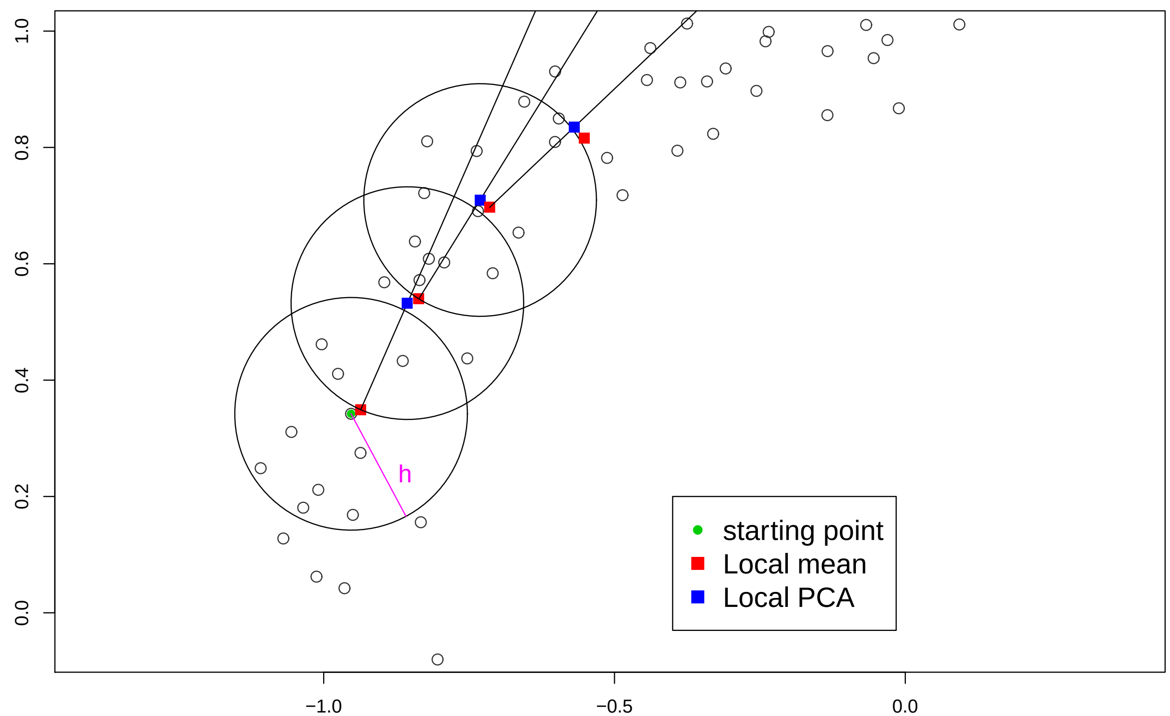
Local principal curves (LPC)

- Goal: Find a smooth curve through the "middle of the data" (like a "nonparametric principal component")
- Idea: Calculate alternately a local mean and a first local principal component, each within a certain bandwidth h.



Local principal curves (LPC)

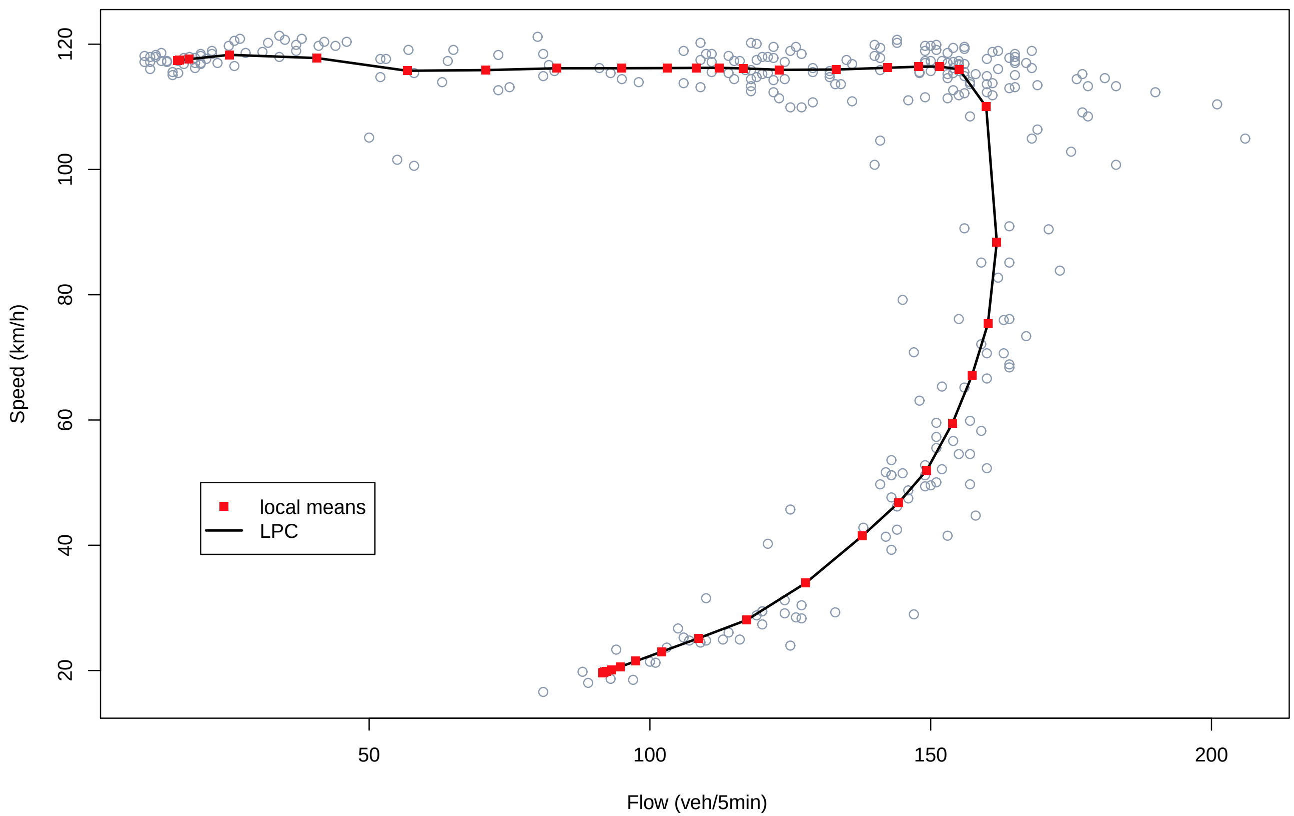
- Goal: Find a smooth curve through the "middle of the data" (like a "nonparametric principal component")
- Idea: Calculate alternately a local mean and a first local principal component, each within a certain bandwidth h.



- At j–th iteration, $\boldsymbol{x}_{(j+1)} = \boldsymbol{m}_{(j)} \pm t\boldsymbol{\gamma}_{(j)}$
- The LPC is the series of local means, $(\boldsymbol{m}_{(j)})_{j \geq 0}$

Local principal curves (cont.)

- Local principal curve ($h = t = 12$) through speed-flow data.



Asymptotics for LPCs

- The distance between two local centers of mass, say $\boldsymbol{m}_{(j)}$ and $\boldsymbol{m}_{(j+1)}$, is given by

$$\begin{aligned}\boldsymbol{m}_{(j+1)} - \boldsymbol{m}_{(j)} &= (\boldsymbol{m}_{(j+1)} - \boldsymbol{x}_{(j+1)}) - (\boldsymbol{x}_{(j+1)} - \boldsymbol{m}_{(j)}) \\ &\overset{a}{=} \mu_2(K)\boldsymbol{H}\frac{\boldsymbol{\nabla} f(\boldsymbol{x}_{(j+1)})}{f(\boldsymbol{x}_{(j+1)})} \pm t\frac{\boldsymbol{H}\boldsymbol{\nabla} f(\boldsymbol{x})}{||\boldsymbol{H}\boldsymbol{\nabla} f(\boldsymbol{x})||}\end{aligned}$$

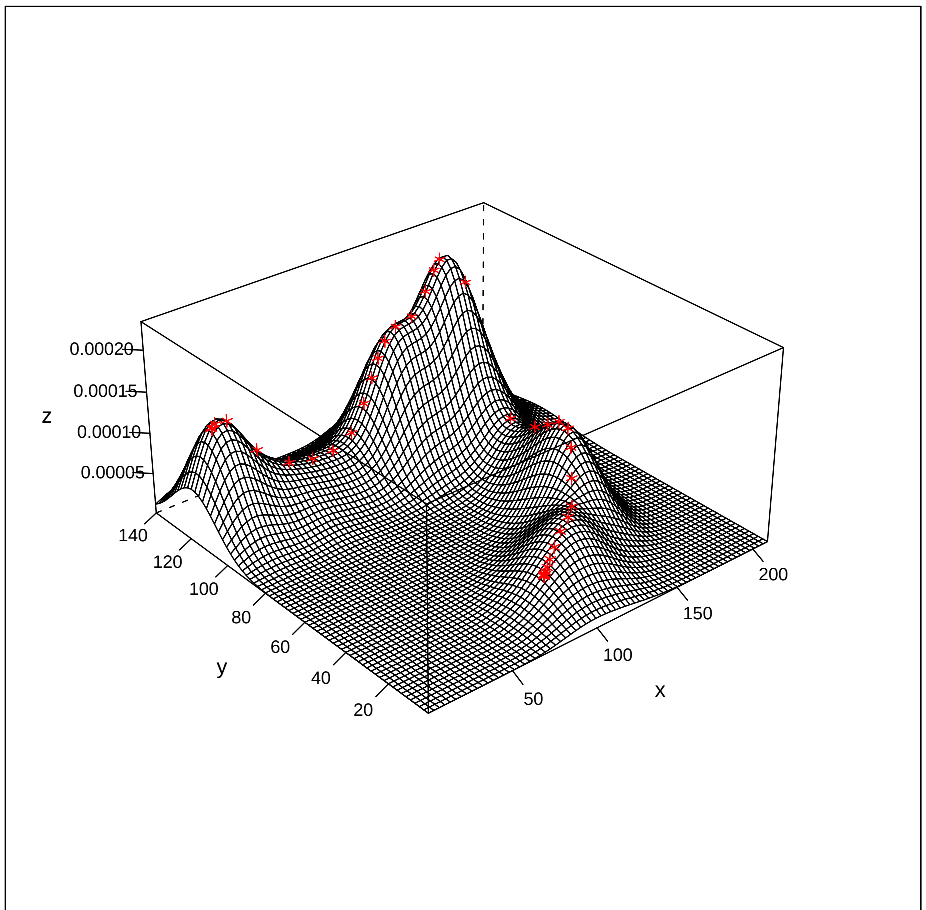
- Using $\boldsymbol{H} = \text{diag}(h^2)$,

$$\boldsymbol{m}_{(j+1)} - \boldsymbol{m}_{(j)} \overset{a}{=} \left[\frac{1}{f(\boldsymbol{x}_{(j)})}h^2 \pm \frac{1}{||\nabla f(\boldsymbol{x}_{(j)})||}t\right]\boldsymbol{\nabla} f(\boldsymbol{x}_{(j)})$$

- Hence, the LPC always turns in direction of the gradient.
- The first term is the mean shift; the second term local PCA.

LPC as ridge estimator

- In practice the LPC will follow the density ridge.
- Both terms have the same sign “uphill” and opposite sign “downhill”.



Kernel density estimate:

$$\hat{f}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\boldsymbol{H}}\left(\boldsymbol{x}_i - \boldsymbol{x}\right)$$

Convergence behavior of LPCs

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

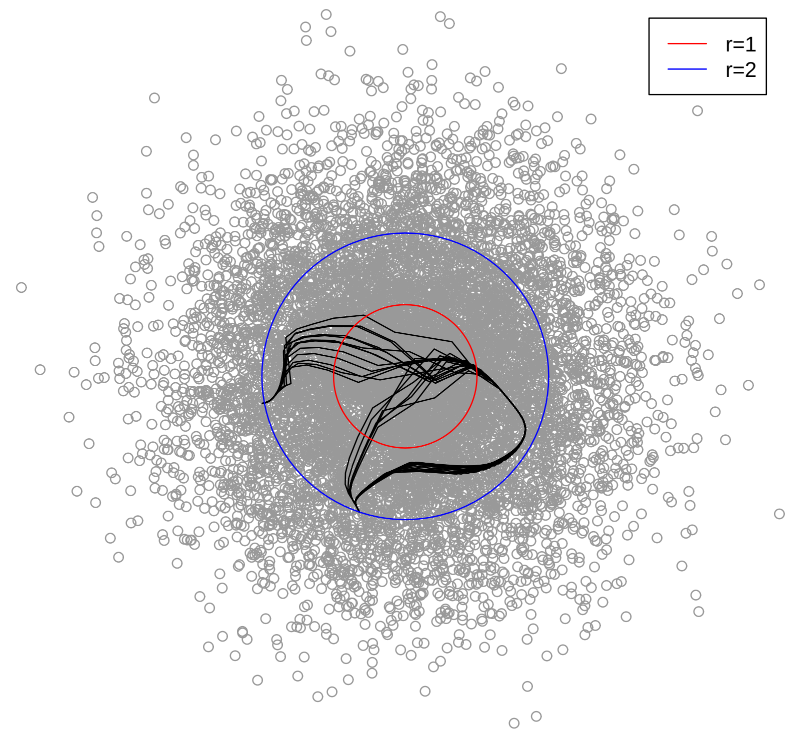
Convergence behavior of LPCs

- Secondly, we observe that the LPC stops when

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

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

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



Conclusion

- Using the ‘expected (localized and penalized) sum of squared orthogonal distances’, one can derive the asymptotic shape of the first localized PC, which turns out to be a unit vector turning into the direction $\boldsymbol{H}\boldsymbol{\nabla} f(\boldsymbol{x})$.
- That is, asymptotically, localized PCA is entirely determined by the bandwidth matrix and the local topology of the data (in terms of the density).
- Localized principal components lead to powerful data approximation algorithms in conjunction with the mean shift, known as ‘local principal curves’ (LPC).
- Using the asymptotic results, the convergence behavior of LPCs can be established, and verified by simulation.
- Can be exploited for boundary extension of principal curves (Zayed, 2011).

References (cont.)

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

Einbeck, J. & Zayed, M. (201x) Some asymptotics for localized principal components and curves; *Communications in Statistics – Theory and Methods*, to appear.

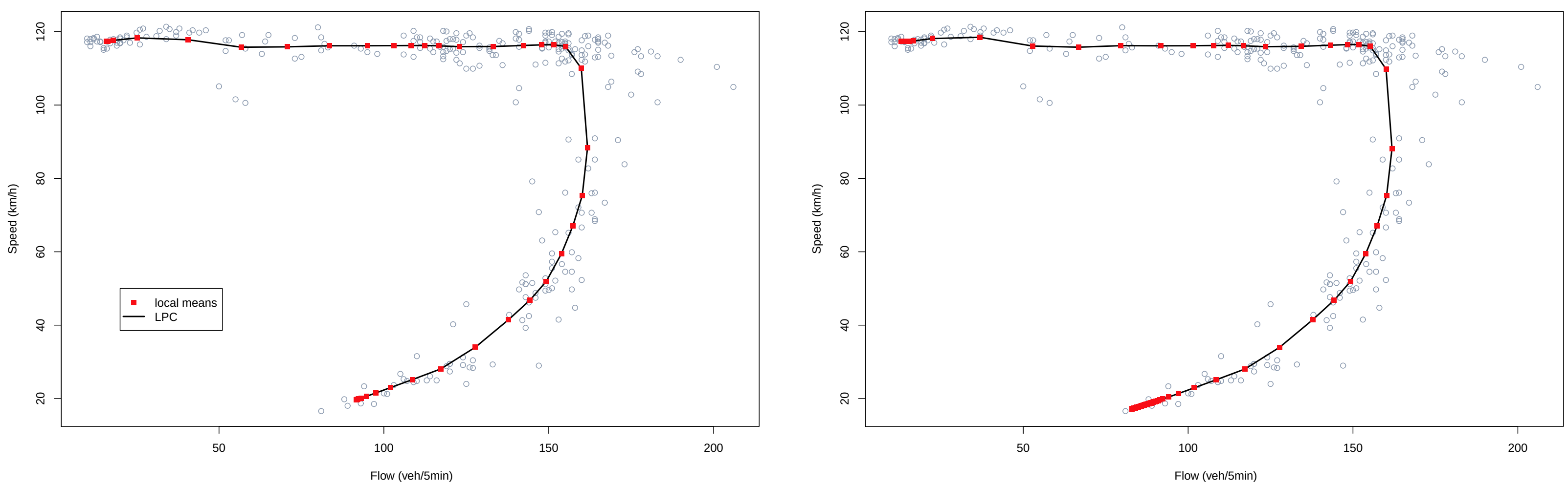
Ruppert, D., & Wand, M.P. (1994) Multivariate locally weighted least squares regression. *Annals of Statistics*, **22**, 1346–1370.

Zayed, M. (2011) *Curve Estimation Based on Localised Principal Components - Theory and Applications*. PhD thesis, Durham University.

Boundary extension

- Reduce h adaptively (compared to t) as principal curve approaches boundary.

without bounday correction　　　　　with boundary correction



- Especially useful for time series data!