Principal... what?

Stanislav Kolenikov
skolenik@unc.edu
117 New West, Cameron Ave,
University of North Carolina,
Chapel Hill, NC 27599-3260, US

March 29, 2001

"Principal" things in statistics

Slide 1

4211 items at researchindex.com, 1381 at Current Index to Statistics, including:

- Principal components
  - linear
  - categorical
  - nonlinear
  - functional
- Principal curves and surfaces
- Principal Hessian directions

Generalization of PC

Elaborate on the main aspects of principal components:

Slide 2

1. Maximum variance of the PC

$$PC_1 = \arg \max_{\mathbf{a}_1 : \|\mathbf{a}_1\| = 1} \mathrm{Var}(\mathbf{X}'\mathbf{a}_1) \tag{1}$$

$$PC_2 = \arg \max_{\mathbf{a}_2 : \|\mathbf{a}_2\| = 1, \mathbf{a}_2 \perp \mathbf{a}_1} \mathrm{Var}(\mathbf{X}'\mathbf{a}_2), \quad \text{etc.} \tag{2}$$

2. Minimum variance of the residuals

$$PC_1 = \arg \min_{\mathbf{a} : \|\mathbf{a}\| = 1} \mathsf{E}\, \mathrm{distance}(\mathbf{a}t + \mathbf{b}, \mathbf{X}), \quad \text{etc.} \tag{3}$$

3. Multivariate normal: self-consistency w.r.t. projection

Nonlinear PCA I

Slide 3

Gifi (1990): the book is interesting *per se* as an approach aiming at better understanding the structure of your data by taking different views of it. The main accent is made on categorical data.

In Gifi's notation, the non-linear PCA can be motivated as the generalization of (3):

$$\sum_j SSQ(\mathbf{x} - \phi_j(\mathbf{h}_j)) \to \min, \tag{4}$$

where $\mathbf{x}$ are scores ($\mathbf{x}'\mathbf{x} = 1$ for normalization) and $\mathbf{h}_j$ are the entries of the data matrix.

All nonlinear transformations allowed $\Rightarrow \min = 0$?

Nonlinear PCA II

To get nontrivial solutions, we need to impose some restrictions. *Smoothness* is the general condition I would think of, but Gifi proposes some other things.

Slide 4

1. *Monotonicity*
2. *Basis expansion*: e.g. polynomials of low order
3. *Categorization* (Gifi's favorite): discretize into a small number of categories, and...

we are back to the convenient setting:

$$\sum_j SSQ(\mathbf{X} - \mathbf{G}_j \mathbf{Y}_j) \to \min \tag{5}$$

Principal curves

Slide 5

The basic reference is Hastie and Stuetzle (1989): definitions, algorithm, applications, discussion.

Principal curve is defined informally as a smooth curve that passes through the middle of the data and is self-consistent under the projections to it.

	Linear	Smooth
Selected dependent variable	Linear regression	Non-parametric regression
All variables treated symmetrically	PCA	Principal curves

Projections

If the curve is parametrized with the parameter λ, then the *projection index* of a data point is the arg of the point on the curve to which the data point is projected:

Slide 6

$$\lambda_{\mathbf{f}}(\mathbf{x}) = \sup_{\lambda} \left\{ \lambda : \|\mathbf{x} - \mathbf{f}(\lambda)\| = \inf_{\mu} \|\mathbf{x} - \mathbf{f}(\mu)\| \right\}, \tag{6}$$

i.e. the value of λ for which $\mathbf{f}(\lambda)$ is closest to $\mathbf{x}$.

With this projection index, self-consistent / principal curves are the curves such that

$$E(\mathbf{X}|\lambda_{\mathbf{f}}(\mathbf{X}) = \lambda) = \mathbf{f}(\lambda) \tag{7}$$

Algorithm

Hastie and Stuetzle (1989):

Slide 7

1. Start from the first principal component:

$$\mathbf{f}^{(0)} = \bar{\mathbf{x}} + \mathbf{a}\lambda, \lambda^{(0)}(\mathbf{x}) = \lambda_{\mathbf{f}^{(0)}}(\mathbf{x}). \tag{8}$$

2. Set

$$\mathbf{f}^{(j)}(\cdot) = E(\mathbf{X}|\lambda_{\mathbf{f}^{(j-1)}}(\mathbf{X}) = \cdot) \tag{9}$$

   Done by scatterplot smoother: perform local fitting for each dimension.

Algorithm (continued)

Slide 8

3. Define

$$\lambda^{(j)}(\mathbf{x}) = \lambda_{\mathbf{f}^{(j)}}(\mathbf{x}) \tag{10}$$

   and transform to unit speed paramterization.

4. Evaluate

$$\Delta^{(j)} = E\left[\|\mathbf{X} - \mathbf{f}(\lambda^{(j)}(\mathbf{X}))\|^2\right] \tag{11}$$

5. Stop if $\Delta^{(j)}$ is small enough, otherwise $j \leftarrow j + 1$ and go to step 2.

Does it all make sense?

Slide 9

Are principal curves really so nice objects to work with?

1. Uniqueness?
2. Bias in the parts of high curvature.
3. Generalizations to 2D, 3D, ... : possible, but cumbersome.
4. Algorithmic issues: choice of the bandwidth? convergence?

Alternatives?

Tibshirani (1992):

1. generate a variable S according to some distribution $g_S(s)$;
2. generate $\mathbf{Y} \in \mathbb{R}^p$ from the conditional distribution $g_{\mathbf{Y}|s}$.

Then a principal curve is a triple $< g_S, g_{\mathbf{Y}|s}, \mathbf{f} >$ satisfying

Slide 10

I. $g_{\mathbf{Y}}(y) = \int g_{\mathbf{Y}|s} g_S(s) ds$;

II. $Y_1, \ldots, Y_p$ conditionally independent given s;

III. $\mathbf{f} : \Gamma \to \mathbb{R}^p$, Γ is a closed interval in $\mathbb{R}$, and $E\mathbf{Y}|s = \mathbf{f}(s)$.

This definition does not suffer from bias, but it only coincides with HS if the support of the conditional distribution $g_{\mathbf{Y}|s}$ is orthogonal to the curve $\mathbf{f}(\cdot)$ at s. The algorithm is a version of the EM-algorithm for a finite mixture of normal distributions with a support on at most n points which are essentially the projections of the data.

My own experience

Slide 11

- Coarser implementation (only projections of the data points) $\Rightarrow$ convergence in the sense of Step 5 never achieved: SS did not decrease monotonically. It makes sense to trace convergence by some graphical means.
- Some strange figures appeared on the way... T. Hastie commented that he had some of them, too.
- Tricks to avoid the previous problem — chop pieces where no points project to; increase bandwidth.
- Starting point was crucial for convergence, or at least for the speed of convergence.
- Performed reasonably well with high dimensional example: dealt with non-normalities, non-linearities, many dimensions. The data were rather simple, though.

Functional data used

Slide 12

Simulation model:

$$x_{ik} = 1 + t - (\frac{1}{2} + \frac{3}{2}\beta_k \exp\left[-(9 - 6\alpha_k)t^2\right] - 2\gamma_k I_{t>0} + \delta_k + \varepsilon_{ik}, \qquad (12)$$
$$t = (i - 21)/20, k = 1, \ldots, 84,$$

where i and t correspond to the measurement points / dimensions, and k enumerates observations / curves, $\delta_k \sim N(0, 0.03^2)$, $\varepsilon_{ik} \sim N(0, 0.02^2)$ are errors independent of each other and of anything else. The parameters $\alpha_k, \beta_k, \gamma_k$ are distributed independently inside a tube that goes along the edges of the unit cube.

30 points: $\alpha_k \sim U[0, 1], \beta_k$ and $\gamma_k \sim U[0, 0.1]$

27 points: $\alpha_k \sim U[0.9, 1], \beta_k \sim U[0, 1], \gamma_k \sim U[0, 0.1]$

27 points: α_k and $\beta_k \sim U[0.9, 1], \gamma_k \sim U[0, 1]$

References

Slide 13

Hastie, T., and Stuetzle, W. Principal Curves. *JASA*, **84**, 502–516 (1989).

Tibshirani, R. Principal Curves Revisited. http://www-stat.stanford.edu/˜tibs/ftp/princcurve.ps

Ramsey, J.O., and Silverman, B.W. Functional Data Analysis (Springer Series in Statistics). Springer (1997).

Gifi, A. Nonlinear multivariate analysis. Wiley (1990).