Constructing ψ Satisfying (A)

$$\int \boldsymbol{x}\psi(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} = 0 \quad \text{(A)}$$

- $h(\boldsymbol{x})$: Smooth non-linear function from $\mathbb{R}^d$ to $\mathbb{R}$
- Then following $\psi(\boldsymbol{x})$ satisfies (A):

$$\psi(\boldsymbol{x}) = \boldsymbol{x}^\top \int \boldsymbol{x}h(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} - h(\boldsymbol{x})$$

Constructing $\boldsymbol{\beta}$ from h

- Then $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\beta}(h) = \int g(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}$$

$$g(\boldsymbol{x}) = \boldsymbol{x} h(\boldsymbol{x}) - \nabla h(\boldsymbol{x})$$

- Empirical approximation:

$$\widehat{\boldsymbol{\beta}}(h) = \frac{1}{n} \sum_{i=1}^{n} g(\boldsymbol{x}_i)$$

- Note: $\widehat{\boldsymbol{\beta}}(h)$ only approximately belongs to the non-Gaussian space.

Identifying Non-Gaussian Subspace

- Each function $h(\boldsymbol{x})$ yields a vector $\widehat{\boldsymbol{\beta}}(h)$
- Prepare a set of different non-linear functions: $\{h_i(\boldsymbol{x})\}_{i=1}^{p}$
- Calculate corresponding vectors: $\{\widehat{\boldsymbol{\beta}}_i\}_{i=1}^{p}$

$$\widehat{\boldsymbol{\beta}}_i = \frac{1}{n}\sum_{j=1}^{n} g_i(\boldsymbol{x}_j)$$

$$g_i(\boldsymbol{x}) = \boldsymbol{x}h_i(\boldsymbol{x}) - \nabla h_i(\boldsymbol{x})$$

Identifying Non-Gaussian Subspace (cont.)

- All $\{\widehat{\boldsymbol{\beta}}_i\}_{i=1}^p$ approximately belong to the non-Gaussian subspace.
- The non-Gaussian subspace may be estimated by "principal subspace" of $\{\widehat{\boldsymbol{\beta}}_i\}_{i=1}^p$.
- We apply PCA to $\{\widehat{\boldsymbol{\beta}}_i\}_{i=1}^p$ and extract leading m directions.

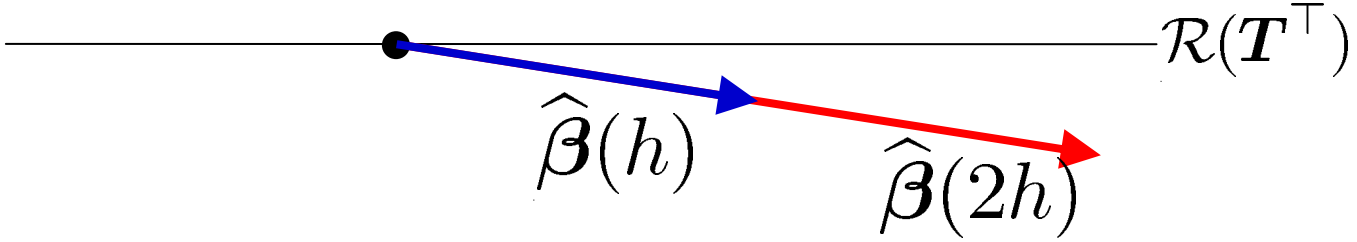
Examples of Non-Linear Functions

- For a random vector w ,
  - $h(\boldsymbol{x}) = \sin(\langle \boldsymbol{w}, \boldsymbol{x} \rangle)$
  - $h(\boldsymbol{x}) = \cos(\langle \boldsymbol{w}, \boldsymbol{x} \rangle)$
  - $h(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle^3 \exp(-\langle \boldsymbol{w}, \boldsymbol{x} \rangle^2)$
  - $h(\boldsymbol{x}) = \tanh(\langle \boldsymbol{w}, \boldsymbol{x} \rangle)$
- Corresponding derivatives are
  - $\nabla h(\boldsymbol{x}) = \boldsymbol{w} \cos(\langle \boldsymbol{w}, \boldsymbol{x} \rangle)$
  - $\nabla h(\boldsymbol{x}) = -\boldsymbol{w} \sin(\langle \boldsymbol{w}, \boldsymbol{x} \rangle)$
  - $\nabla h(\boldsymbol{x}) = 3\boldsymbol{w}\langle \boldsymbol{w}, \boldsymbol{x} \rangle^2 \exp(-\langle \boldsymbol{w}, \boldsymbol{x} \rangle^2)$
    $-2\boldsymbol{w}\langle \boldsymbol{w}, \boldsymbol{x} \rangle^4 \exp(-\langle \boldsymbol{w}, \boldsymbol{x} \rangle^2)$
  - $\nabla h(\boldsymbol{x}) = \boldsymbol{w}(1 - \tanh(\langle \boldsymbol{w}, \boldsymbol{x} \rangle))$

Norm of $\widehat{\boldsymbol{\beta}}$

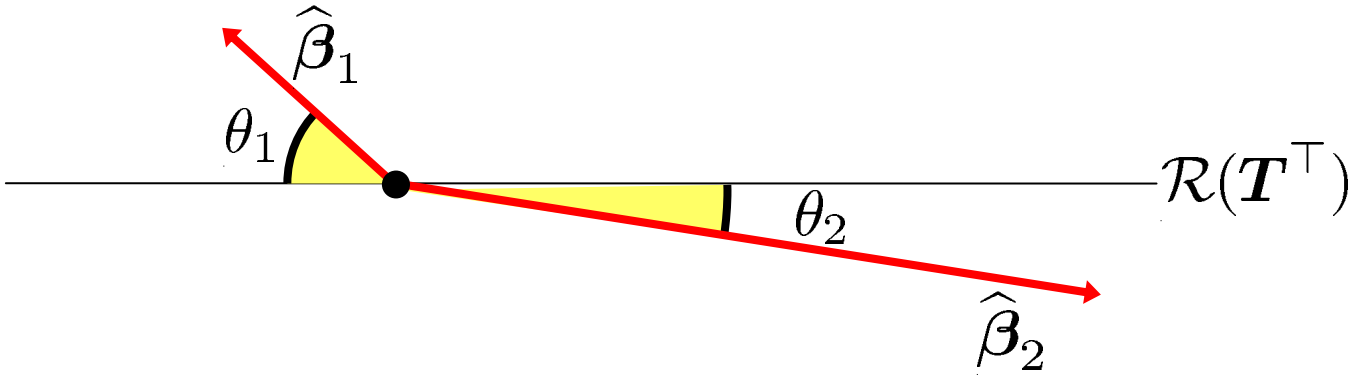
$$\widehat{\boldsymbol{\beta}}_i = \frac{1}{n}\sum_{j=1}^{n} g_i(\boldsymbol{x}_j) \qquad g_i(\boldsymbol{x}) = \boldsymbol{x}h_i(\boldsymbol{x}) - \nabla h_i(\boldsymbol{x})$$

- Derivative is a linear operation.
- Mapping $h \longrightarrow \widehat{\boldsymbol{\beta}}$ is therefore linear.
- Norm of $\widehat{\boldsymbol{\beta}}$ can be arbitrary by rescaling h .
- For example, $h(\boldsymbol{x})$ and $2h(\boldsymbol{x})$ give the same direction but length is different.



Normalization

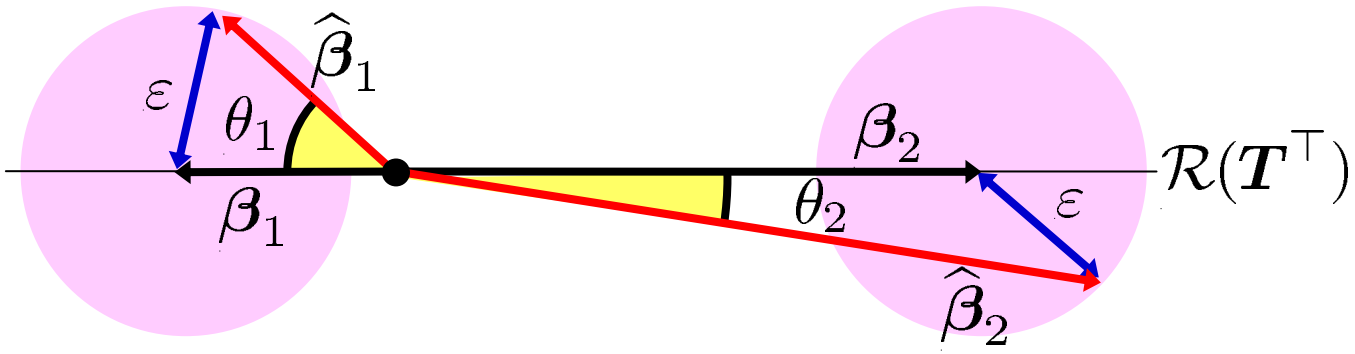
- In PCA, long vectors are more “powerful” than short ones.
- In order to have better estimate of the non-Gaussian subspace by PCA, $\{\widehat{\boldsymbol{\beta}}_i\}_{i=1}^{p}$ should be reasonably normalized.
- Normalization should be carried out such that accurate $\widehat{\boldsymbol{\beta}}_i$ has large norm.



Normalization (cont.)

- This may be achieved by normalizing $\{\widehat{\boldsymbol{\beta}}_i\}_{i=1}^p$ such that the standard deviation $\sqrt{\varepsilon_i}$ is equivalent for all $\{\widehat{\boldsymbol{\beta}}_i\}_{i=1}^p$.

$$\varepsilon_i = \mathbb{E}_{\{\boldsymbol{x}_i\}_{i=1}^n} \|\widehat{\boldsymbol{\beta}}_i - \mathbb{E}_{\{\boldsymbol{x}_i\}_{i=1}^n} \widehat{\boldsymbol{\beta}}_i\|$$



- However, ε_i is inaccessible.

Empirical Approximation

- Variance is expressed as

$$\varepsilon_i = \frac{1}{n}\mathbb{E}_{\boldsymbol{x}}\|g_i(\boldsymbol{x})\|^2 - \frac{1}{n}\|\mathbb{E}_{\boldsymbol{x}}g_i(\boldsymbol{x})\|^2$$

- Empirical approximation:

$$\widehat{\varepsilon}_i = \frac{1}{n^2}\sum_{j=1}^{n}\|g_i(\boldsymbol{x}_j)\|^2 - \frac{1}{n}\|\frac{1}{n}\sum_{j=1}^{n}g_i(\boldsymbol{x}_j)\|^2$$

Algorithm of Non-Gaussian Component Analysis

- Prepare p different non-linear functions:

$$\{h_i(\boldsymbol{x})\}_{i=1}^{p}$$

- Calculate $\widehat{\boldsymbol{\beta}}_i = \frac{1}{n}\sum_{j=1}^{n} g_i(\boldsymbol{x}_j)$

$$g_i(\boldsymbol{x}) = \boldsymbol{x}h_i(\boldsymbol{x}) - \nabla h_i(\boldsymbol{x})$$

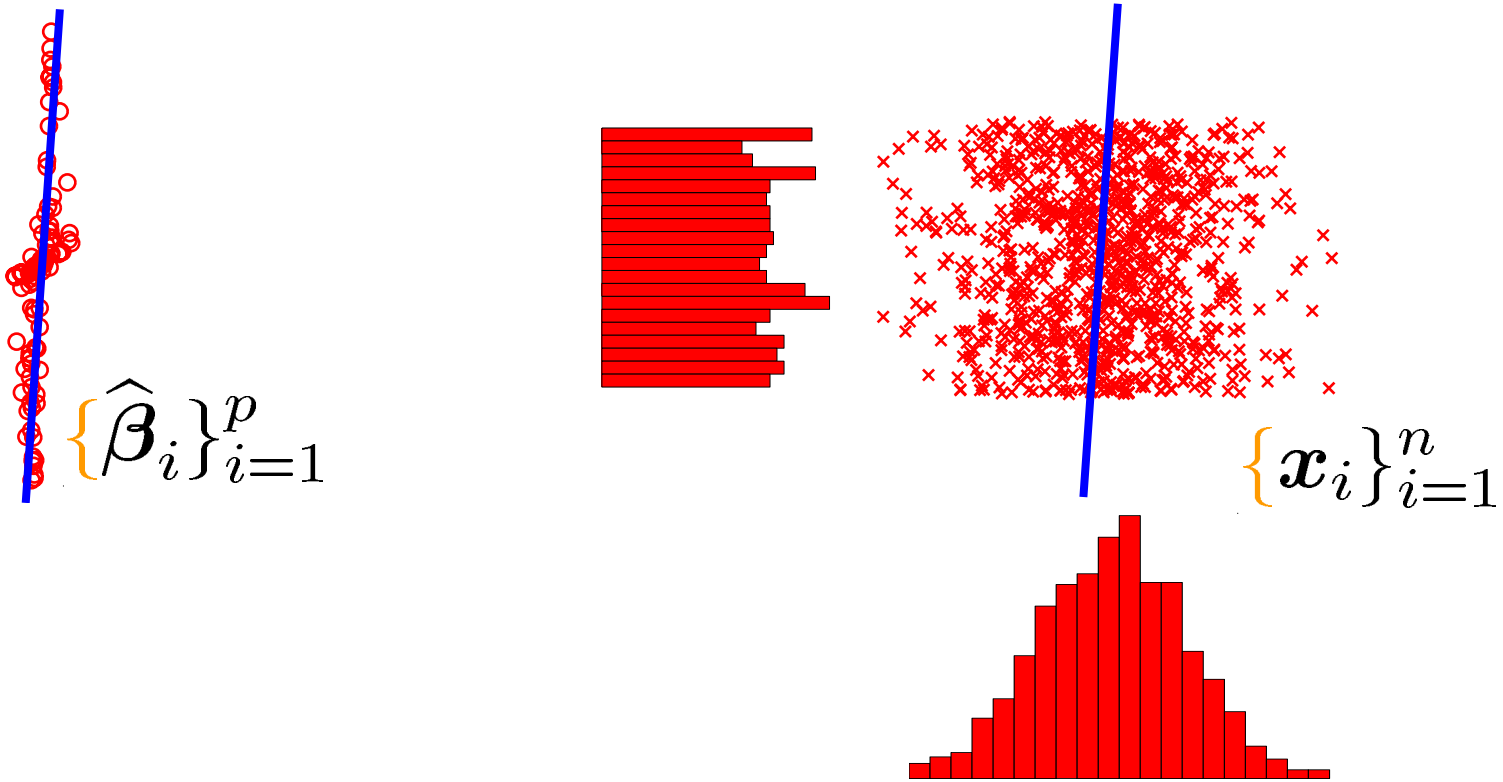
- Normalize $\{\widehat{\boldsymbol{\beta}}_i\}_{i=1}^{p}$: $\widehat{\boldsymbol{\beta}}_i \longleftarrow \widehat{\boldsymbol{\beta}}_i / \sqrt{\widehat{\varepsilon}_i}$

$$\widehat{\varepsilon}_i = \frac{1}{n^2}\sum_{j=1}^{n} \|g_i(\boldsymbol{x}_j)\|^2 - \frac{1}{n}\|\frac{1}{n}\sum_{j=1}^{n} g_i(\boldsymbol{x}_j)\|^2$$

- Apply PCA to $\{\widehat{\boldsymbol{\beta}}_i\}_{i=1}^{p}$ and extract leading m components.

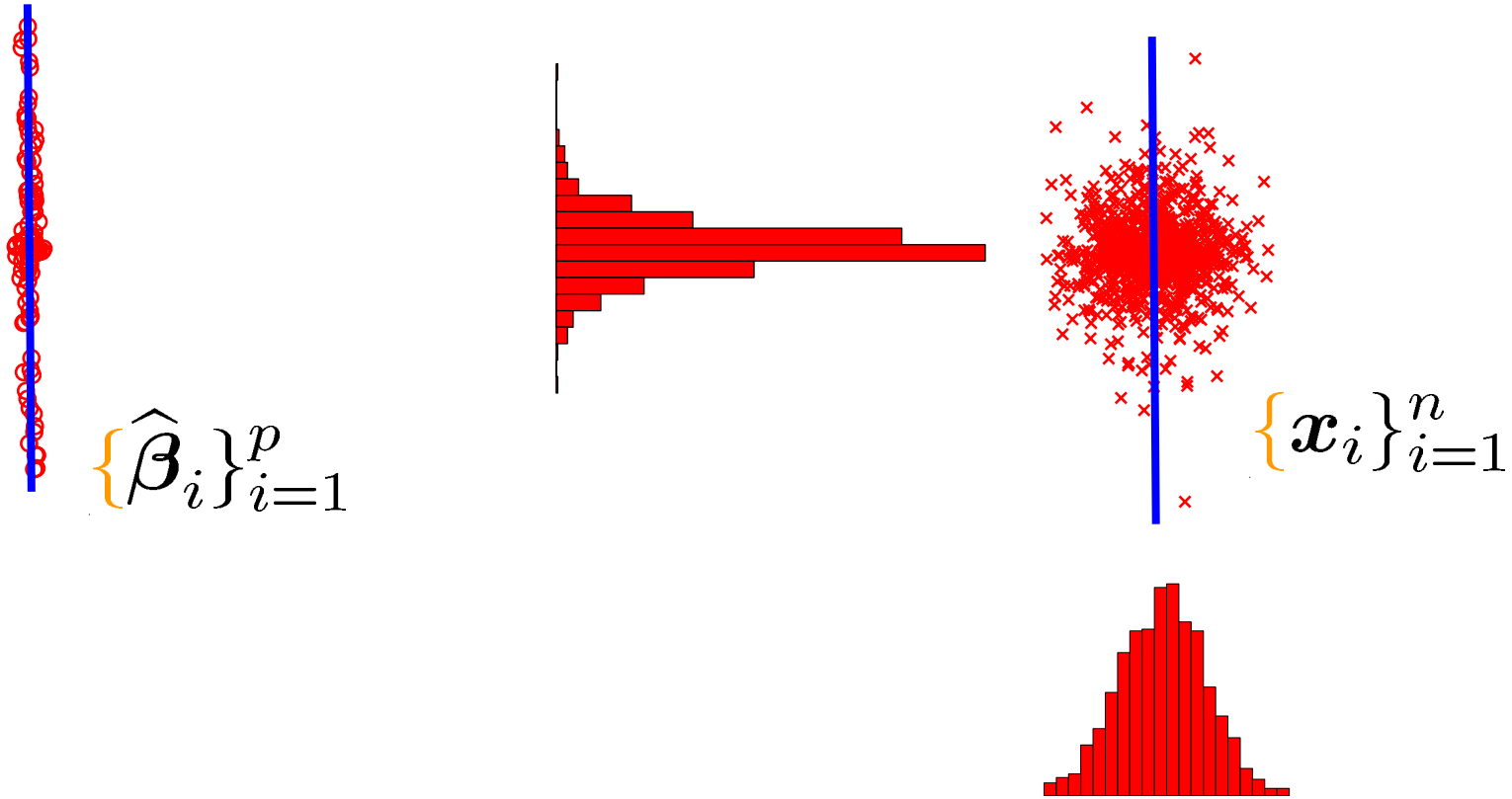
Examples

- $p = 100$
- $h_i(\boldsymbol{x}) = \sin(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle)$



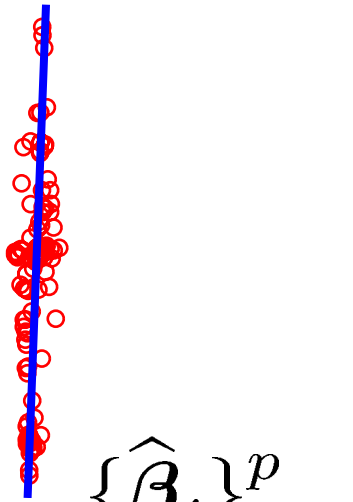
Examples (cont.)

- $p = 100$
- $h_i(\boldsymbol{x}) = \sin(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle)$

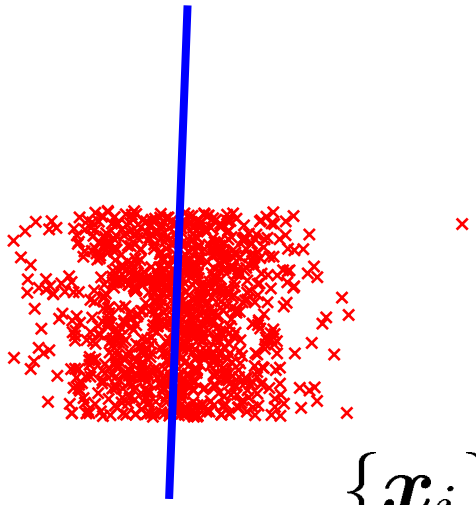


Examples (cont.)

- With outliers, NGCA with sin works well.



$\{\widehat{\boldsymbol{\beta}}_i\}_{i=1}^p$



$\{\boldsymbol{x}_i\}_{i=1}^n$

Homework

- $\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{\epsilon} \qquad \boldsymbol{s} \sim q(\boldsymbol{s}) \qquad \boldsymbol{\epsilon} \sim \phi_{\mathbf{0},\boldsymbol{\Gamma}}(\boldsymbol{\epsilon})$
- $\boldsymbol{P}$: Projection onto $\boldsymbol{\Gamma}^{-1}\mathcal{S}$
- $\boldsymbol{T}$: $m \times d$ matrix such that $\boldsymbol{P} = \boldsymbol{T}^\top \boldsymbol{T}$

$$m = \dim(\boldsymbol{\Gamma}^{-1}\mathcal{S})$$

- Prove $p(\boldsymbol{x}) = f(\boldsymbol{T}\boldsymbol{x})\phi_{\mathbf{0},\boldsymbol{\Gamma}}(\boldsymbol{x})$

$$f(\boldsymbol{z}) = g(\boldsymbol{T}^\top \boldsymbol{z})$$

$$g(\boldsymbol{x}) = \int q(\boldsymbol{s}) e^{-\frac{1}{2}\langle \boldsymbol{\Gamma}^{-1}\boldsymbol{s}, \boldsymbol{s}\rangle} e^{\langle \boldsymbol{\Gamma}^{-1}\boldsymbol{s}, \boldsymbol{x}\rangle} d\boldsymbol{s}$$

Homework (cont.)

- Prove $\varepsilon_i = \mathbb{E}_{\{\boldsymbol{x}_i\}_{i=1}^n} \|\widehat{\boldsymbol{\beta}}_i - \mathbb{E}_{\{\boldsymbol{x}_i\}_{i=1}^n} \widehat{\boldsymbol{\beta}}_i\|^2$ is expressed as

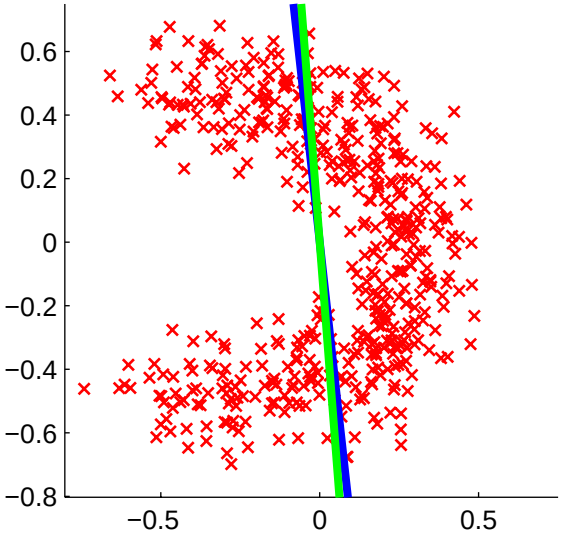
$$\varepsilon_i = \frac{1}{n}\mathbb{E}_{\boldsymbol{x}} \|g_i(\boldsymbol{x})\|^2 - \frac{1}{n}\|\mathbb{E}_{\boldsymbol{x}} g_i(\boldsymbol{x})\|^2$$

$$\widehat{\boldsymbol{\beta}}_i = \frac{1}{n}\sum_{j=1}^{n} g(\boldsymbol{x}_j)$$

Linear Dimensionality Reduction Methods: Summary

Method	Advantages	Disadvantages
PCA	Good data description Analytic solution available No tuning parameter	Structure can be missed
LPP	Local structure preservation Analytic solution available	Tuning parameters included
PP	Interesting structure discovery	No analytic solution NG measure prefixed
NGCA	Interesting structure discovery	No analytic solution

Data with Curved Structures



- If the data cloud is bent, any linear methods fail to find the curved structure.

→ Limitation of linear method!

Suggestion

- Read the following article for the next class:
- B. Schölkopf, A. Smola and K.-R. Müller: Nonlinear Component Analysis as a Kernel Eigenvalue Problem, *Neural Computation*, 10(5), 1299-1319, 1998.