Advanced Data Analysis: Projection Pursuit

Masashi Sugiyama (Computer Science)

W8E-505, sugi@cs.titech.ac.jp

http://sugiyama-www.cs.titech.ac.jp/~sugi

I.i.d. Samples

- Independent and identically distributed (i.i.d.) samples

$$\boldsymbol{x}_i \overset{i.i.d.}{\sim} P(\boldsymbol{x})$$

  - Independent: joint probability is a product of each probability

$$P(\boldsymbol{x}_i, \boldsymbol{x}_j) = P(\boldsymbol{x}_i)P(\boldsymbol{x}_j)$$

  - Identically distributed: each variable follow the identical distribution:

$$\boldsymbol{x}_i \sim P(\boldsymbol{x})$$

Gaussian Distribution

- Gaussian distribution: Probability density function is given by

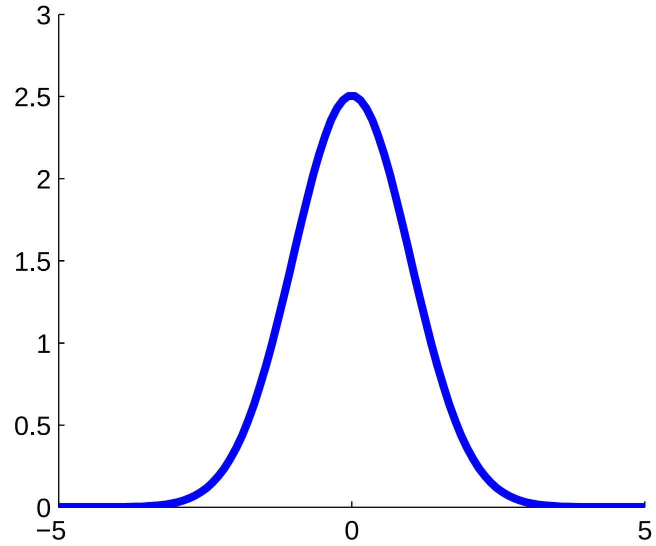
$$\phi_{\boldsymbol{\theta},\boldsymbol{\Gamma}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Gamma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\theta})^\top \boldsymbol{\Gamma}^{-1}(\boldsymbol{x}-\boldsymbol{\theta})\right)$$

- $\boldsymbol{\theta}, \boldsymbol{\Gamma}$:Mean, covariance

$$\mathbb{E}[\boldsymbol{x}] = \boldsymbol{\theta}$$

$$\mathbb{E}[(\boldsymbol{x}-\theta)(\boldsymbol{x}-\theta)^\top] = \boldsymbol{\Gamma}$$

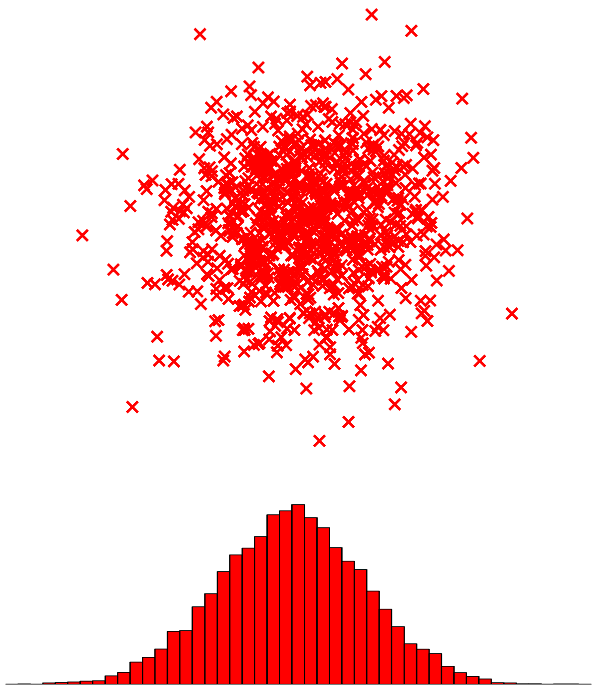
- When one-dimensional,



$$\phi_{\theta,\gamma}(x) = \frac{1}{\sqrt{2\pi\gamma^2}} \exp\left(-\frac{(x-\theta)^2}{2\gamma^2}\right)$$

Interesting Directions for Data Visualization

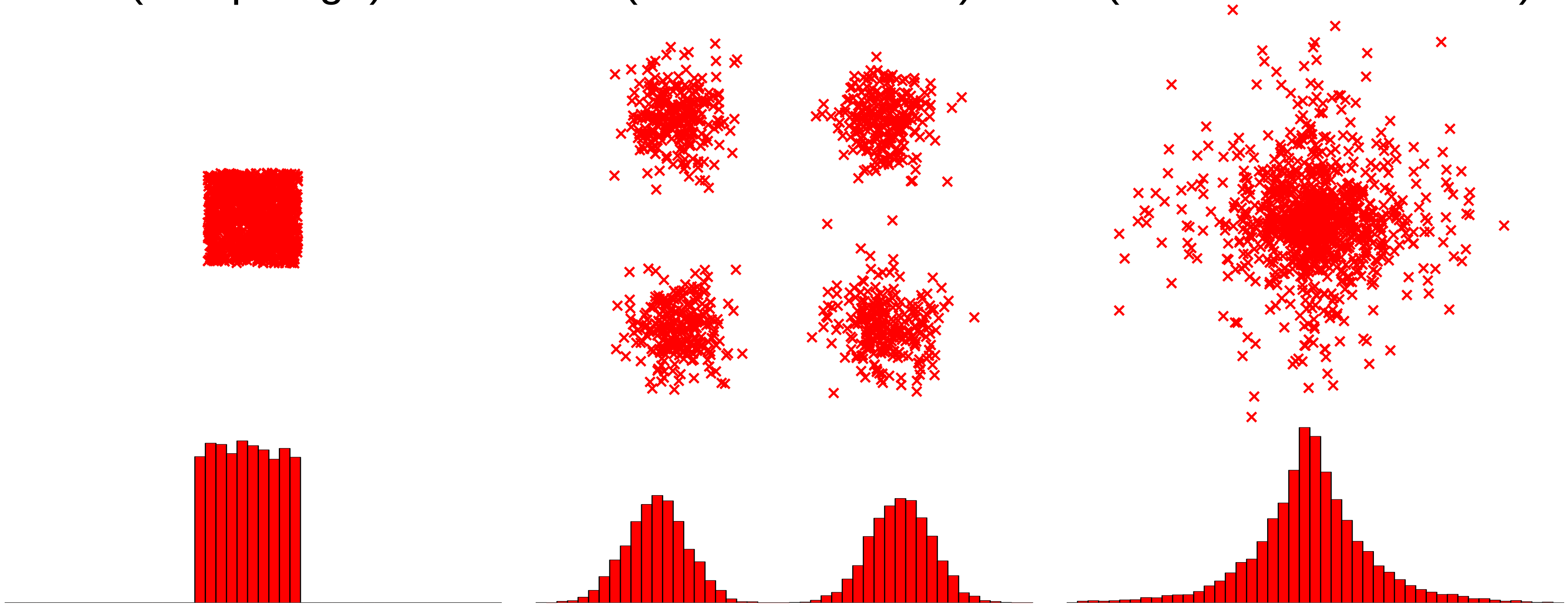
- Which distribution is interesting to visualize?
- If data follows the Gaussian distribution, samples are spherically distributed.
- Visualizing spherically distributed samples is not so interesting.
- What about "non-Gaussian" data?



Non-Gaussian Distributed Data

- Non-Gaussian data look more interesting than Gaussian:

Uniform (sharp edge) | Gaussian mixture (cluster structure) | Laplacian (existence of outliers)



Projection Pursuit

- Idea: Find the most non-Gaussian direction in the data
- For this purpose, we need a criterion to measure non-Gaussianity of data as a function of the direction.

Kurtosis

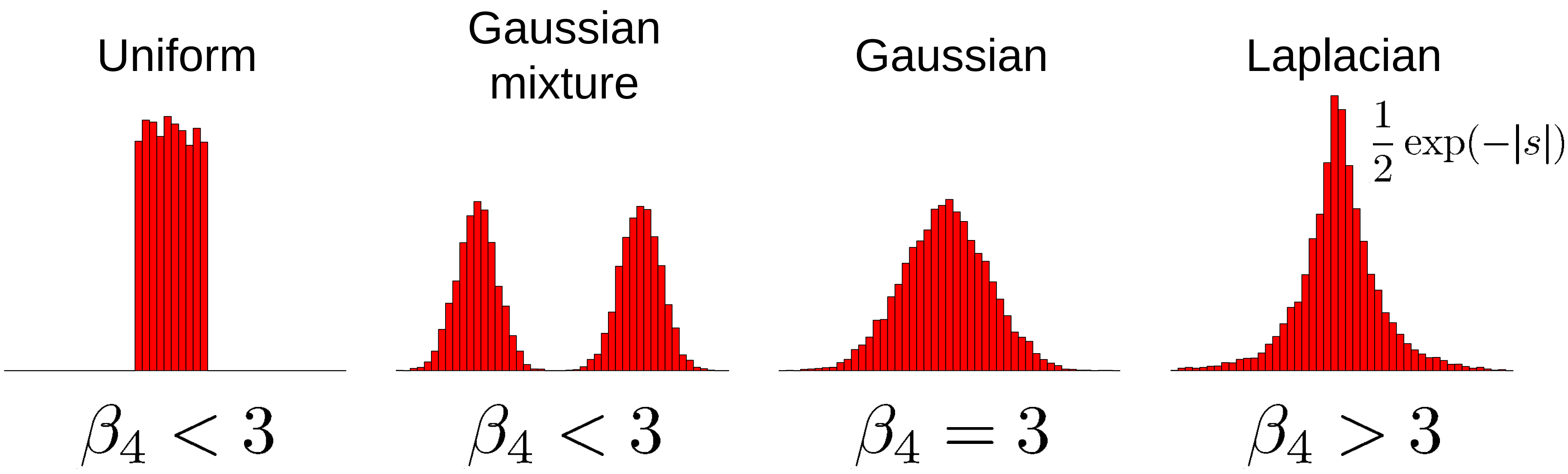
- Kurtosis for a one-dimensional random variable s:

$$\beta_4 = \frac{\mathbb{E}[(s-\mathbb{E}[s])^4]}{(\mathbb{E}[(s-\mathbb{E}[s])^2])^2} \quad (>0)$$

- Kurtosis measures the "sharpness" of the distributions.
- If tail of distribution is
  - Heavy ➡ β_4 is large
  - Light ➡ β_4 is small

Kurtosis (cont.)

- $\beta_4 = 3$: Gaussian distribution
- $\beta_4 < 3$: Sub-Gaussian distribution
- $\beta_4 > 3$: Super-Gaussian distribution



Kurtosis-Based Non-Gaussianity Measure

$$\beta_4 = \frac{\mathbb{E}[(s - \mathbb{E}[s])^4]}{(\mathbb{E}[(s - \mathbb{E}[s])^2])^2}$$

- Non-Gaussianity is strong if $(\beta_4 - 3)^2$ is large.
- Non-Gaussianity of the data for a direction $\boldsymbol{b}$ can be measured by letting $s = \langle \boldsymbol{b}, \boldsymbol{x} \rangle$ and $\|\boldsymbol{b}\| = 1$.

PP Criterion

- In practice, we use empirical approximation:

$$J_{PP}(\boldsymbol{b}) = \left(\frac{\frac{1}{n}\sum_{i=1}^{n}(s_i - \overline{s})^4}{(\frac{1}{n}\sum_{i=1}^{n}(s_i - \overline{s})^2)^2} - 3 \right)^2$$

$$s_i = \langle \boldsymbol{b}, \boldsymbol{x}_i \rangle$$

$$\overline{s} = \frac{1}{n}\sum_{i=1}^{n} s_i$$

- PP criterion:

$$\psi = \underset{\boldsymbol{b}\in\mathbb{R}^d}{\operatorname{argmax}}\, J_{PP}(\boldsymbol{b})$$

$$\text{subject to } \|\boldsymbol{b}\| = 1$$

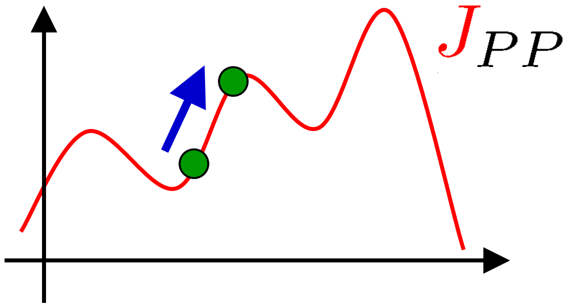
- There is no known method for analytically solving this optimization problem.
- We resort to numerical methods.

Gradient Ascent Approach

- Repeat until convergence:
  - Update $\boldsymbol{b}$ to increase J_{PP} :

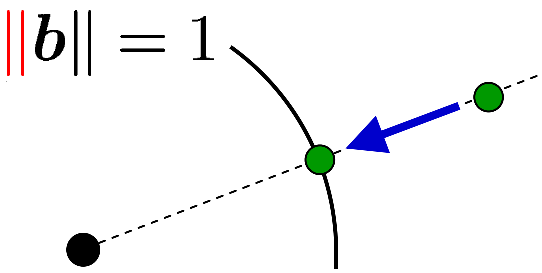
$$\boldsymbol{b} \longleftarrow \boldsymbol{b} + \varepsilon \frac{\partial J_{PP}}{\partial \boldsymbol{b}}$$

$$(\varepsilon > 0)$$



  - Modify $\boldsymbol{b}$ to satisfy $\|\boldsymbol{b}\| = 1$:

$$\boldsymbol{b} \longleftarrow \boldsymbol{b} / \|\boldsymbol{b}\|$$



Data Centering and Sphering

- We center and sphere for easy calculation.
- Centering:

$$\overline{\boldsymbol{x}}_i = \boldsymbol{x}_i - \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{x}_j$$

- Sphering (or pre-whitening):

$$\widetilde{\boldsymbol{x}}_i = \left(\frac{1}{n}\sum_{i=1}^{n}\overline{\boldsymbol{x}}_i\overline{\boldsymbol{x}}_i^\top\right)^{-\frac{1}{2}}\overline{\boldsymbol{x}}_i$$

- In matrix,

$$\widetilde{\boldsymbol{X}} = (\frac{1}{n}\boldsymbol{X}\boldsymbol{H}^2\boldsymbol{X}^\top)^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{H}$$

$$\widetilde{\boldsymbol{X}} = (\widetilde{\boldsymbol{x}}_1|\widetilde{\boldsymbol{x}}_2|\cdots|\widetilde{\boldsymbol{x}}_n)$$

$$\boldsymbol{X} = (\boldsymbol{x}_1|\boldsymbol{x}_2|\cdots|\boldsymbol{x}_n)$$

$$\boldsymbol{H} = \boldsymbol{I}_n - \frac{1}{n}\boldsymbol{1}_{n\times n}$$

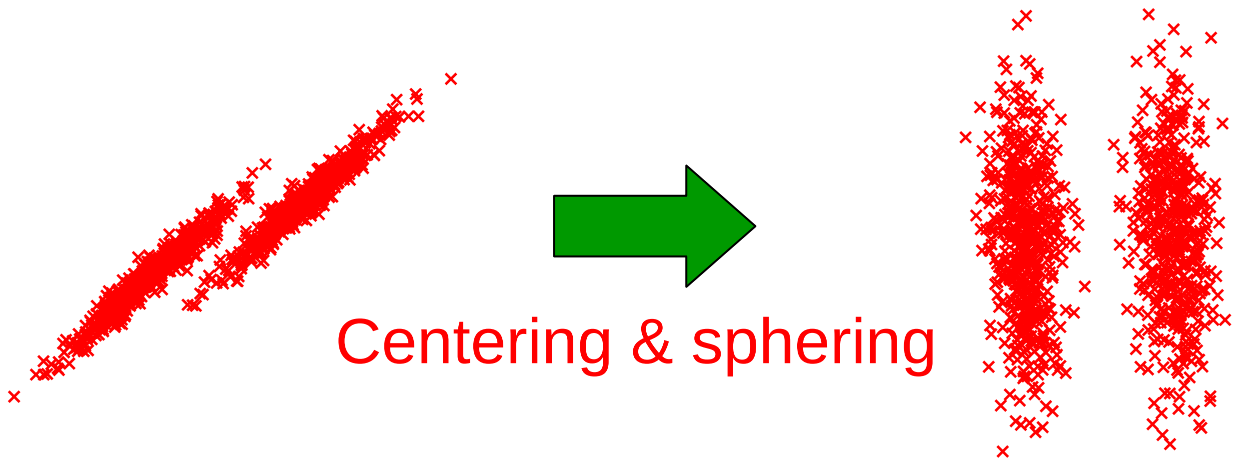
$\boldsymbol{I}_n$: n-dimensional identity matrix

$\boldsymbol{1}_{n\times n}$: $n \times n$ matrix with all ones

Data Centering and Sphering

- By centering and sphering, covariance matrix becomes identity:

$$\frac{1}{n}\sum_{i=1}^{n}\widetilde{\boldsymbol{x}}_i\widetilde{\boldsymbol{x}}_i^\top = \boldsymbol{I}_d$$



Simplification for Sphered Data

- For centered and sphered samples $\{\widetilde{\boldsymbol{x}}_i\}_{i=1}^n$,

$$J_{PP}(\boldsymbol{b}) = \left(\frac{1}{n}\sum_{i=1}^{n}\langle \boldsymbol{b}, \widetilde{\boldsymbol{x}}_i\rangle^4 - 3\right)^2$$

$$\frac{\partial J_{PP}}{\partial \boldsymbol{b}} = 2\left(\frac{1}{n}\sum_{i=1}^{n}\langle \boldsymbol{b}, \widetilde{\boldsymbol{x}}_i\rangle^4 - 3\right)\left(\frac{4}{n}\sum_{i=1}^{n}\widetilde{\boldsymbol{x}}_i\langle \boldsymbol{b}, \widetilde{\boldsymbol{x}}_i\rangle^3\right)$$

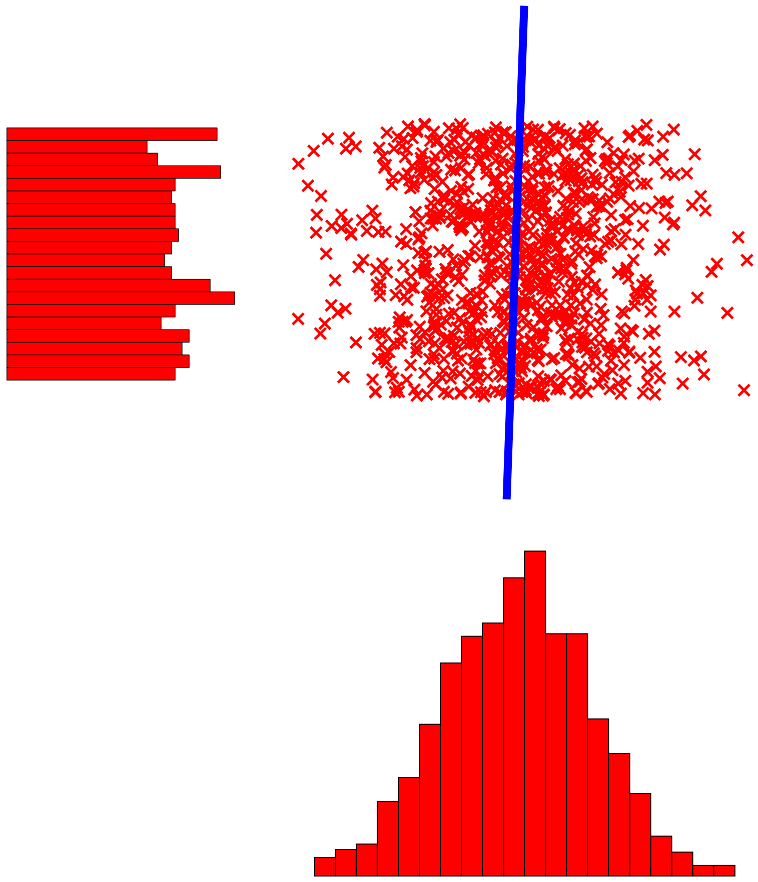
- Gradient update rule is

$$\boldsymbol{b} \longleftarrow \boldsymbol{b} + \varepsilon\left(\frac{1}{n}\sum_{i=1}^{n}\langle \boldsymbol{b}, \widetilde{\boldsymbol{x}}_i\rangle^4 - 3\right)\frac{1}{n}\sum_{i=1}^{n}\widetilde{\boldsymbol{x}}_i\langle \boldsymbol{b}, \widetilde{\boldsymbol{x}}_i\rangle^3$$

- Don’t forget normalization: $\boldsymbol{b} \longleftarrow \boldsymbol{b}/\|\boldsymbol{b}\|$
- Homework: Prove them!

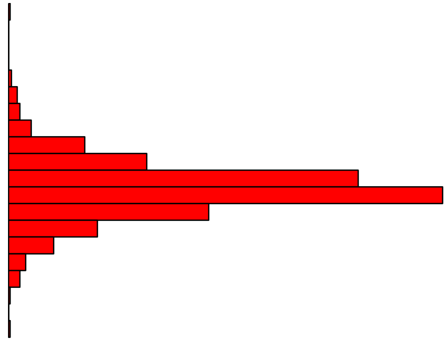
Examples

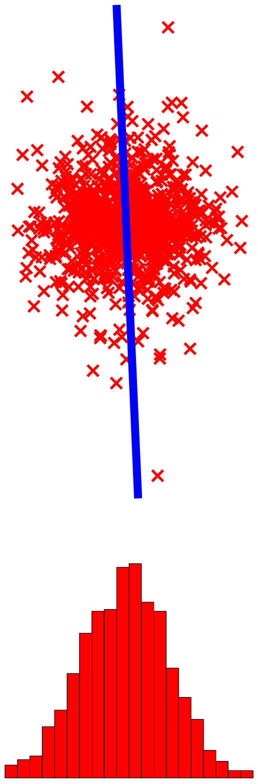
- $d = 2, \quad m = 1, \quad n = 1000$
- $x = \begin{pmatrix} s \\ t \end{pmatrix}$
- $s \sim N(0, 1)$
- $t \sim U(-\sqrt{3}, \sqrt{3})$



Examples (cont.)

- $d = 2, \quad m = 1, \quad n = 1000$
- $x = \begin{pmatrix} s \\ t \end{pmatrix}$
- $s \sim N(0, 1)$
- $t \sim Lap(0, 1)$





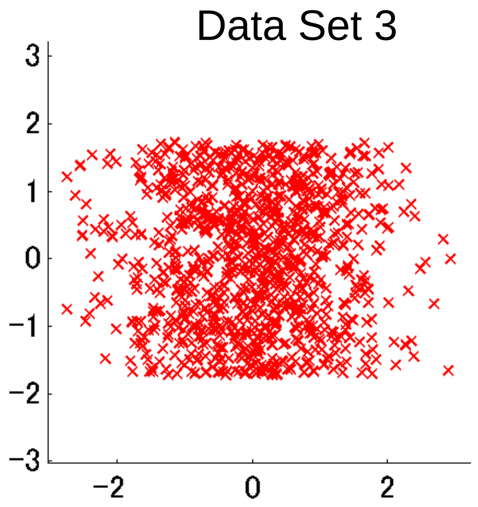
Important Notice

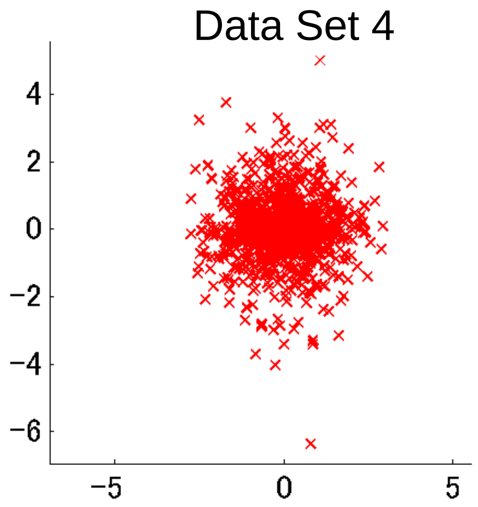
- There will be no class next week (Jun. 26th)

Homework

1. Implement PP and reproduce the 2-dimensional examples shown in the class.

http://sugiyama-www.cs.titech.ac.jp/~sugi/data/DataAnalysis





You may create similar (or more interesting) data sets by yourself.

Homework (cont.)

2. Prove the following for centered and sphered samples $\{\widetilde{\boldsymbol{x}}_i\}_{i=1}^n$:

   A) Covariance matrix is given by

$$\frac{1}{n}\sum_{i=1}^{n}\widetilde{\boldsymbol{x}}_i\widetilde{\boldsymbol{x}}_i^\top = \boldsymbol{I}_d$$

   B) J_{PP} under $\|\boldsymbol{b}\| = 1$ is given by

$$J_{PP}(\boldsymbol{b}) = \left(\frac{1}{n}\sum_{i=1}^{n}\langle\boldsymbol{b}, \widetilde{\boldsymbol{x}}_i\rangle^4 - 3\right)^2$$

   C) Gradient $\partial J_{PP}/\partial\boldsymbol{b}$ is given by

$$\frac{\partial J_{PP}}{\partial\boldsymbol{b}} = 2\left(\frac{1}{n}\sum_{i=1}^{n}\langle\boldsymbol{b}, \widetilde{\boldsymbol{x}}_i\rangle^4 - 3\right)\left(\frac{4}{n}\sum_{i=1}^{n}\widetilde{\boldsymbol{x}}_i\langle\boldsymbol{b}, \widetilde{\boldsymbol{x}}_i\rangle^3\right)$$