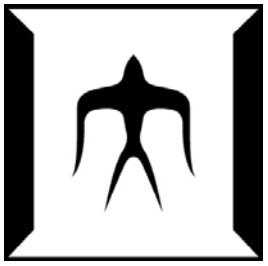
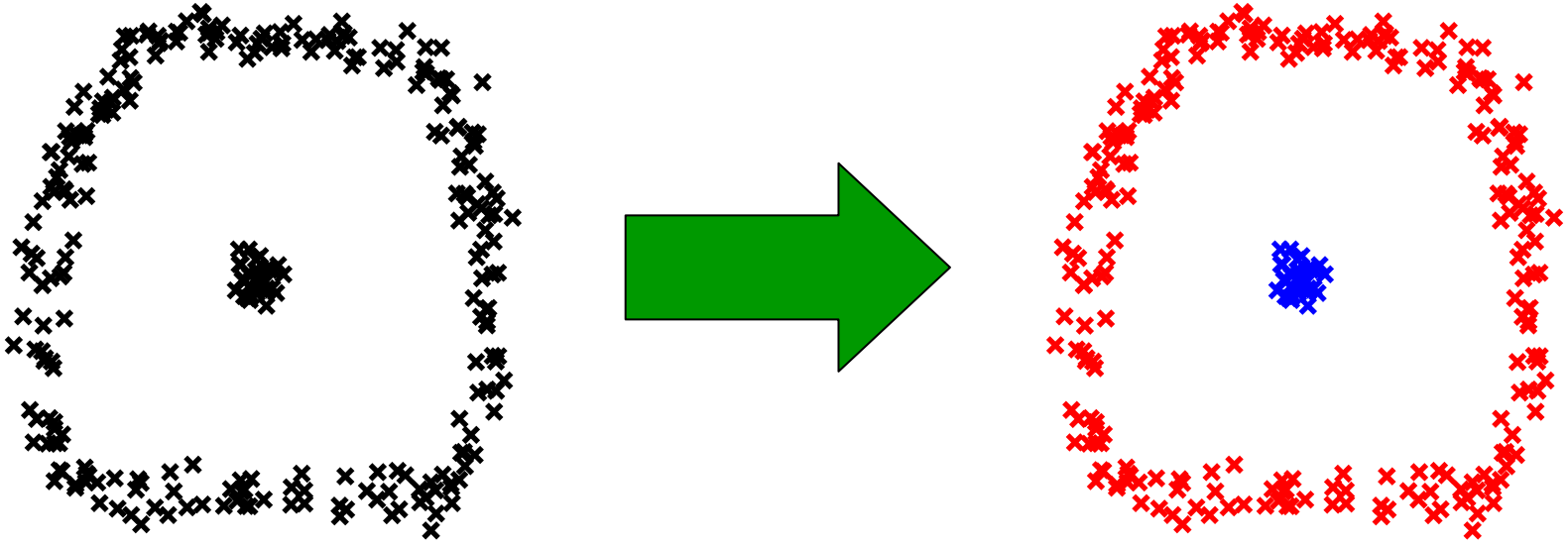
On Information-Maximization Clustering: Tuning Parameter Selection and Analytic Solution



Masashi Sugiyama, Makoto Yamada,
Manabu Kimura, and Hirotaka Hachiya
Department of Computer Science
Tokyo Institute of Technology

Goal of Clustering

- Given unlabeled samples $\{\boldsymbol{x}_i \in \mathbb{R}^d\}_{i=1}^n$,
  assign cluster labels $\{y_i \in \{1, \ldots, c\}\}_{i=1}^n$ so that
  - Samples in the same cluster are similar.
  - Samples in other clusters are dissimilar.
- Throughout this talk, we assume c is known.



Contents

1. Problem formulation
2. Review of existing approaches
3. Proposed method
   A) Clustering
   B) Tuning parameter optimization
4. Experiments

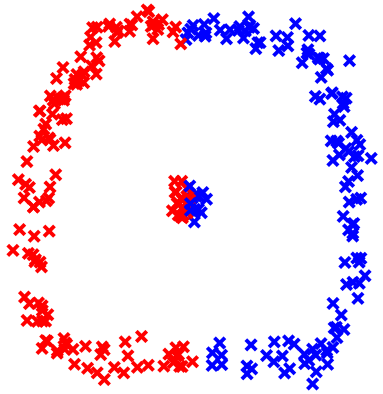
Model-based Clustering

- Learn a mixture model by maximum-likelihood or Bayes estimation:

$$p(\boldsymbol{x}) = \sum_{y=1}^{c} p(\boldsymbol{x}|y)p(y)$$

  - K-means (MacQueen, 1967)
  - Dirichlet process mixture (Ferguson, 1973)

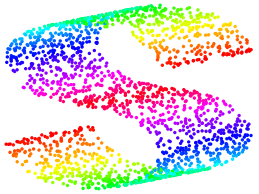
- Pros and cons:
  - ☺ No tuning parameters.
  - ☹ Cluster shape depends on pre-defined cluster models ($\approx$ Gaussian).
  - ☹ Initialization is difficult.



Model-free Clustering

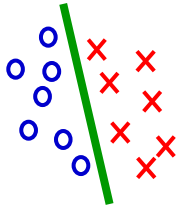
- No parametric assumption on clusters:
  - Spectral clustering: K-means after non-linear manifold embedding

    (Shi & Malik, 2000; Ng *et al*., 2002)



  - Discriminative clustering: Learn a classifier and cluster labels simultaneously

    (Xu *et al*., 2005; Bach & Harchaoui, 2008)



  - Dependence maximization: Determine labels so that dependence on samples is maximized

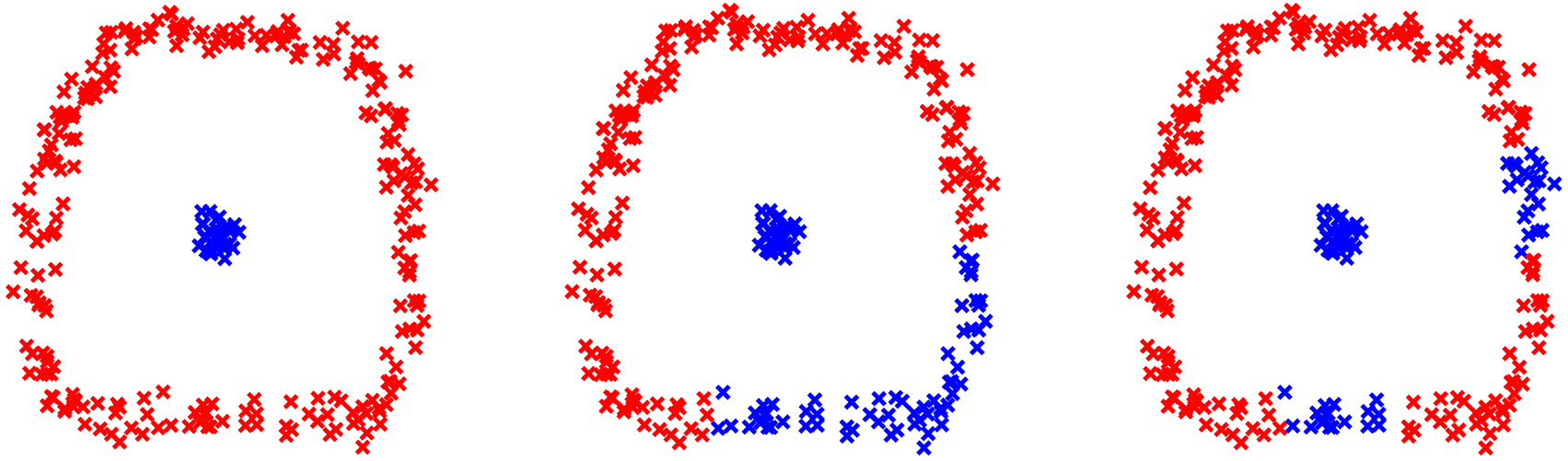
    (Song *et al*., 2007; Faivishevsky & Goldberger, 2010)

  - Information maximization: Learn a classifier so that some information measure is maximized

    (Agakov & Barberu, 2006; Gomes *et al*., 2010)

Model-free Clustering (cont.)

- Pros and cons:
  - ☺ Cluster shape is flexible.
  - ☹ Kernel/similarity parameter choice is difficult.
  - ☹ Initialization is difficult.



Contents

1. Problem formulation
2. Review of existing approaches
3. Proposed method
   A) Clustering
   B) Tuning parameter optimization
4. Experiments

Goal of Our Research

- We propose a new information-maximization clustering method:
  - ☺ Global analytic solution is available.
  - ☺ Objective tuning-parameter choice is possible.

- In the proposed method:
  - A non-parametric kernel classifier is learned so that an information measure is maximized.
  - Tuning parameters are chosen so that an information measure is maximized.

Squared-loss Mutual Information (SMI)

- As an information measure, we use SMI:

$$\mathrm{SMI} = \frac{1}{2} \int \sum_{y=1}^{c} p(\boldsymbol{x}) p(y) \left(\frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x}) p(y)} - 1 \right)^2 \mathrm{d}\boldsymbol{x}$$

  - Ordinary MI is the KL divergence.
  - SMI is the Pearson (PE) divergence.
  - Both KL and PE are f-divergences (thus they have similar properties).
  - Indeed, as ordinary MI, SMI satisfies

$$\mathrm{SMI} = 0 \iff \boldsymbol{x} \perp\!\!\!\perp y$$

Contents

1. Problem formulation
2. Review of existing approaches
3. Proposed method
   A) Clustering
   B) Tuning parameter optimization
4. Experiments

Kernel Probabilistic Classifier

- Kernel probabilistic classifier:

$$p(y|\boldsymbol{x};\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_{y,i} K(\boldsymbol{x}, \boldsymbol{x}_i)$$

- Learn the classifier so that SMI is maximized.

$$\mathrm{SMI} = \frac{1}{2} \int \sum_{y=1}^{c} p(\boldsymbol{x}) p(y) \left(\frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x}) p(y)} - 1 \right)^2 \mathrm{d}\boldsymbol{x}$$

- Challenge: only $\{\boldsymbol{x}_i\}_{i=1}^{n}$ is available for training

SMI Approximation

- Approximate cluster-posterior by kernel model:

$$p(y|\boldsymbol{x}) \approx \sum_{i=1}^{n} \alpha_{y,i} K(\boldsymbol{x}, \boldsymbol{x}_i)$$

- Approximate expectation by sample average:

$$\int f(\boldsymbol{x}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \approx \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{x}_i)$$

$$\{\boldsymbol{x}_i\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x})$$

- Assume cluster-prior is uniform:

$$p(y) = 1/c$$

c : # clusters

- Then we obtain the following SMI approximator:

$$\widehat{\mathrm{SMI}} = \frac{c}{2n} \sum_{y=1}^{c} \boldsymbol{\alpha}_y^\top \boldsymbol{K}^2 \boldsymbol{\alpha}_y - \frac{1}{2}$$

$$\boldsymbol{\alpha}_y = (\alpha_{y,1}, \ldots, \alpha_{y,n})^\top$$

$$K_{i,j} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

Maximizing SMI Approximator

$$\max_{\{\boldsymbol{\alpha}_y\}_{y=1}^c} \widehat{\mathrm{SMI}}$$

$$\widehat{\mathrm{SMI}} = \frac{c}{2n} \sum_{y=1}^{c} \boldsymbol{\alpha}_y^\top \boldsymbol{K}^2 \boldsymbol{\alpha}_y - \frac{1}{2}$$

- Under mutual orthonormality of $\{\boldsymbol{\alpha}_y\}_{y=1}^c$, a solution is given by principal components of kernel matrix $\boldsymbol{K}$.
  - Similar to Ding & He (ICML2004)

SMI-based Clustering (SMIC)

- Post-processing:
  - Adjusting sign of principal components $\{\boldsymbol{\phi}_y\}_{y=1}^{c}$:

$$\widetilde{\boldsymbol{\phi}}_y = \boldsymbol{\phi}_y \times \mathrm{sign}(\boldsymbol{\phi}_y^\top \mathbf{1})$$

$\mathbf{1}$:Vector with all ones

  - Normalization according to $p(y) = 1/c$.
  - Rounding-up negative probability estimates to 0.
- Final solution (analytically computable):

$$y_i = \operatorname*{argmax}_y \frac{[\max(\mathbf{0}, \widetilde{\boldsymbol{\phi}}_y)]_i}{\max(\mathbf{0}, \widetilde{\boldsymbol{\phi}}_y)^\top \mathbf{1}}$$

$[\cdot]_i$:i-th element of a vector

$\mathbf{0}$:Vector with all zeros

Contents

1. Problem formulation
2. Review of existing approaches
3. Proposed method
   A) Clustering
   B) Tuning parameter optimization
4. Experiments

Tuning Parameter Choice

- Solution of SMIC depends on kernel functions.
- We determine kernels so that SMI is maximized.
- We may use the same $\widehat{\mathrm{SMI}}$ for this purpose.

$$\widehat{\mathrm{SMI}} = \frac{c}{2n} \sum_{y=1}^{c} \boldsymbol{\alpha}_y^\top \boldsymbol{K}^2 \boldsymbol{\alpha}_y - \frac{1}{2}$$

- However, $\widehat{\mathrm{SMI}}$ is not accurate enough since it is an unsupervised estimator of SMI.
- In the phase of tuning parameter choice, estimated labels are available!

Supervised SMI Estimator

$$\mathrm{SMI} = \frac{1}{2} \int \sum_{y=1}^{c} p(\boldsymbol{x}) p(y) \left(\frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x}) p(y)} - 1 \right)^2 \mathrm{d}\boldsymbol{x}$$

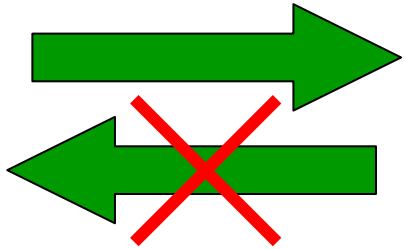
- Least-squares mutual information (LSMI):
  - Directly estimate the density ratio

    Suzuki & Sugiyama (AISTATS2010)

$$r(\boldsymbol{x}, y) = \frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x}) p(y)}$$

    without going through density estimation.
  - Density-ratio estimation is substantially easier than density estimation (*à la* Vapnik).

Knowing $p(\boldsymbol{x}, y), p(\boldsymbol{x}), p(y)$



Knowing $r(\boldsymbol{x}, y) = \dfrac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x}) p(y)}$

Density-Ratio Estimation

- Kernel density-ratio model:

$$\widehat{r}(\boldsymbol{x}, y) = \sum_{\ell : y_\ell = y} \theta_\ell L(\boldsymbol{x}, \boldsymbol{x}_\ell)$$

$L(\boldsymbol{x}, \boldsymbol{x}')$: Kernel function

(We use Gaussian kernel)

- Least-squares fitting:

$$\frac{1}{2} \int \sum_{y=1}^{c} \left(\widehat{r}(\boldsymbol{x}, y) - r(\boldsymbol{x}, y) \right)^2 p(\boldsymbol{x}) p(y) d\boldsymbol{x}$$

$$r(\boldsymbol{x}, y) = \frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x}) p(y)}$$

Density-Ratio Estimation (cont.)

- Empirical and regularized training criterion:

$$\min_{\boldsymbol{\theta}_y} \left[\frac{1}{2} \boldsymbol{\theta}_y^\top \widehat{\boldsymbol{H}}^{(y)} \boldsymbol{\theta}_y - \boldsymbol{\theta}_y^\top \widehat{\boldsymbol{h}}^{(y)} + \frac{\delta}{2} \boldsymbol{\theta}_y^\top \boldsymbol{\theta}_y \right]$$

$$\widehat{H}_{\ell,\ell'}^{(y)} = \frac{n_y}{n^2} \sum_{i=1}^{n} L(\boldsymbol{x}_i, \boldsymbol{x}_\ell^{(y)}) L(\boldsymbol{x}_i, \boldsymbol{x}_{\ell'}^{(y)}) \qquad \widehat{h}_\ell^{(y)} = \frac{1}{n} \sum_{i:y_i=y} L(\boldsymbol{x}_i, \boldsymbol{x}_\ell^{(y)})$$

- Global solution can be obtained analytically:

$$\widehat{\boldsymbol{\theta}}^{(y)} = (\widehat{\boldsymbol{H}}^{(y)} + \delta \boldsymbol{I})^{-1} \widehat{\boldsymbol{h}}^{(y)}$$

$$\widehat{r}(\boldsymbol{x}, y) = \sum_{\ell=1}^{n_y} \widehat{\theta}_\ell^{(y)} L(\boldsymbol{x}, \boldsymbol{x}_\ell^{(y)})$$

- Kernel and regularization parameter can be determined by cross-validation.

Least-Squares Mutual Information (LSMI)

- SMI approximator is given analytically as

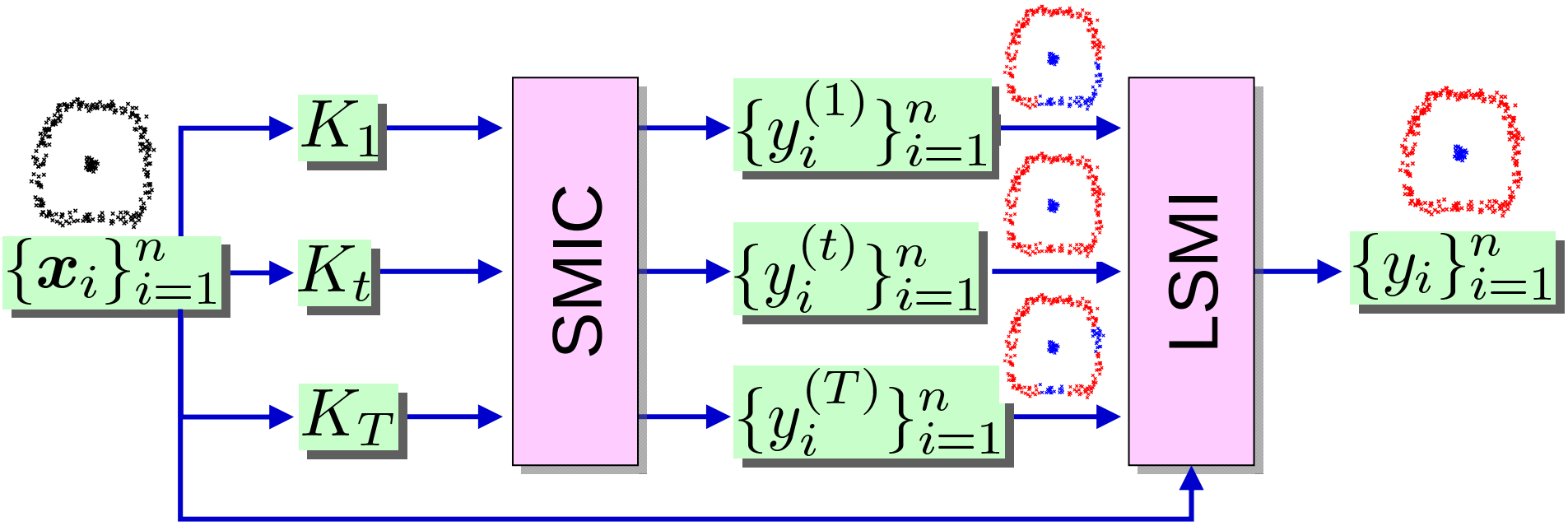
$$\mathrm{LSMI} = -\frac{1}{2n^2}\sum_{i,j=1}^{n}\widehat{r}(\boldsymbol{x}_i, y_j)^2 + \frac{1}{n}\sum_{i=1}^{n}\widehat{r}(\boldsymbol{x}_i, y_i) - \frac{1}{2}$$

- LSMI achieves a fast non-parametric convergence rate! Suzuki & Sugiyama (AISTATS2010)

- We determine the kernel function in SMIC so that LSMI is maximized.

Summary of Proposed Method

- SMI Clustering with LSMI:
  - Input: Unlabeled samples $\{\boldsymbol{x}_i\}_{i=1}^n$
    Kernel candidates $\{K_t(\boldsymbol{x}, \boldsymbol{x}')\}_{t=1}^T$
  - Output: Cluster labels $\{y_i\}_{i=1}^n$



Contents

1. Problem formulation
2. Review of existing approaches
3. Proposed method
   A) Clustering
   B) Tuning parameter optimization
4. Experiments

Experimental Setup

- For SMIC, we use a sparse variant of the local scaling kernel: (Zelnik-Manor & Perona, NIPS2004)

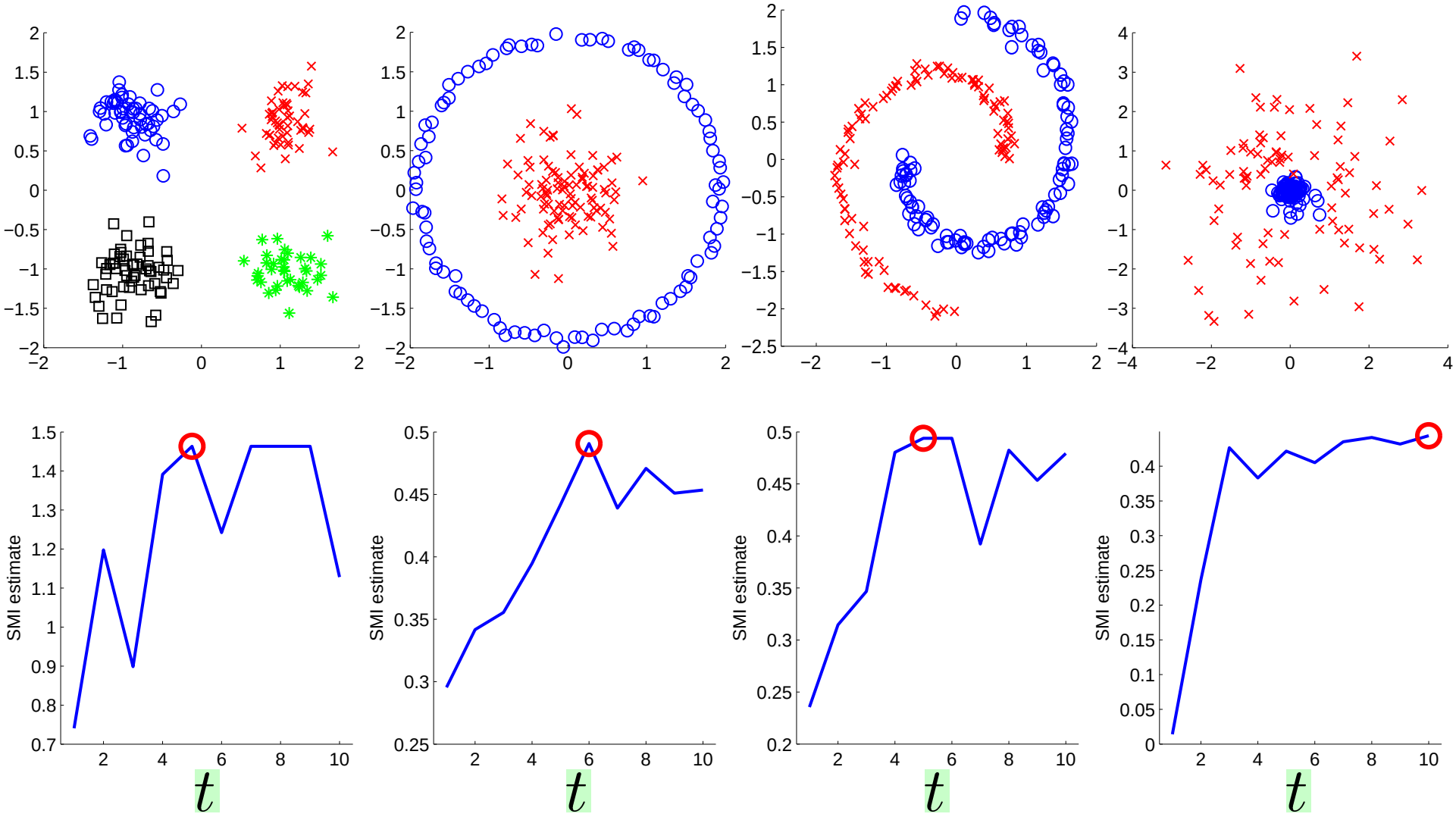
$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \begin{cases} \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma_i \sigma_j}\right) & \text{if } \boldsymbol{x}_i, \boldsymbol{x}_j \text{ are } t \text{ neighbors} \\ 0 & \text{otherwise} \end{cases}$$

$\sigma_i = \|\boldsymbol{x}_i - \boldsymbol{x}_i^{(t)}\|$

$\boldsymbol{x}_i^{(t)}$: t-th neighbor of $\boldsymbol{x}_i$

- Tuning parameter t is determined by LSMI maximization.

Illustration of SMIC



- SMIC with model selection by LSMI works well!

Performance Comparison

- KM: K-means clustering (MacQueen, 1967)
- SC: Self-tuning spectral clustering (Zelnik-Manor & Perona, NIPS2004)
- MNN: Dependence-maximization clustering based on mean nearest neighbor approximation (Faivishevsky & Goldberger, ICML2010)
- MIC: Information-maximization clustering for kernel logistic models with model selection by maximum-likelihood mutual information (Gomes, Krause & Perona, NIPS2010) (Suzuki, Sugiyama, Sese & Kanamori, FSDM2008)

Experimental Results

Digit ($d = 256, n = 5000,$ and $c = 10$)

	KM	SC	MNN	MIC	SMIC
ARI	0.42(0.01)	0.24(0.02)	0.44(0.03)	**0.63(0.08)**	**0.63(0.05)**
Time	835.9	973.3	318.5	84.4[3631.7]	14.4[359.5]

Face ($d = 4096, n = 100,$ and $c = 10$)

	KM	SC	MNN	MIC	SMIC
ARI	0.60(0.11)	**0.62(0.11)**	0.47(0.10)	**0.64(0.12)**	**0.65(0.11)**
Time	93.3	2.1	1.0	1.4[30.8]	0.0[19.3]

20News Group

Document ($d = 50, n = 700,$ and $c = 7$)

	KM	SC	MNN	MIC	SMIC
ARI	0.00(0.00)	0.09(0.02)	0.09(0.02)	0.01(0.02)	**0.19(0.03)**
Time	77.8	9.7	6.4	3.4[530.5]	0.3[115.3]

Sens eval2

Word ($d = 50, n = 300,$ and $c = 3$)

	KM	SC	MNN	MIC	SMIC
ARI	0.04(0.05)	0.02(0.01)	0.02(0.02)	0.04(0.04)	**0.08(0.05)**
Time	6.5	5.9	2.2	1.0[369.6]	0.2[203.9]

Accelerometry ($d = 5, n = 300,$ and $c = 3$)

	KM	SC	MNN	MIC	SMIC
ARI	0.49(0.04)	0.58(0.14)	**0.71(0.05)**	0.57(0.23)	**0.68(0.12)**
Time	0.4	3.3	1.9	0.8[410.6]	0.2[92.6]

Speech ($d = 50, n = 400,$ and $c = 2$)

	KM	SC	MNN	MIC	SMIC
ARI	0.00(0.00)	0.00(0.00)	0.04(0.15)	**0.18(0.16)**	**0.21(0.25)**
Time	0.9	4.2	1.8	0.7[413.4]	0.3[179.7]

- Adjusted Rand index (ARI): larger is better
- Red: Best or comparable by 1%t-test
- SMIC works well and computationally efficient!

Conclusions

- Weaknesses of existing clustering methods:
  - Cluster initialization is difficult.
  - Tuning parameter choice is difficult.
- SMIC: A new information-maximization clustering method based on squared-loss mutual information (SMI):
  - Analytic global solution is available.
  - Objective tuning parameter choice is possible.
- MATLAB code is available from
  http://sugiyama-www.cs.titech.ac.jp/~sugi/software/SMIC/

Other Usage of SMI

$$\mathrm{SMI} = \frac{1}{2} \int \sum_{y=1}^{c} p(\boldsymbol{x}) p(y) \left(\frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x}) p(y)} - 1 \right)^2 \mathrm{d}\boldsymbol{x}$$

- Feature selection

  Suzuki, Sugiyama, Sese & Kanamori (BMC Bioinfo. 2009)

- Dimensionality reduction

  Suzuki & Sugiyama (AISTATS2010)

  Yamada, Niu, Takagi & Sugiyama (ArXiv2011)

- Independent component analysis

  Suzuki & Sugiyama (Neural Comp. 2011)

- Independence test

  Sugiyama & Suzuki (IEICE-ED2011)

- Causal inference

  Yamada & Sugiyama (AAAI2010)