IEICE Transactions on Information and Systems,
vol.E97-D, no.10, pp.2806–2809, 2014.

Unsupervised Dimension Reduction via Least-Squares Quadratic Mutual Information

Janya Sainui
Tokyo Institute of Technology
nguyen@sg.cs.titech.ac.jp

Masashi Sugiyama
Tokyo Institute of Technology
sugi@cs.titech.ac.jp
http://sugiyama-www.cs.titech.ac.jp/~sugi

Abstract

The goal of dimension reduction is to represent high-dimensional data in a lower-dimensional subspace, while intrinsic properties of the original data are kept as much as possible. An important challenge in unsupervised dimension reduction is the choice of tuning parameters, because no supervised information is available and thus parameter selection tends to be subjective and heuristic. In this paper, we propose an information-theoretic approach to unsupervised dimension reduction that allows objective tuning parameter selection. We employ *quadratic mutual information* (QMI) as our information measure, which is known to be less sensitive to outliers than ordinary mutual information, and QMI is estimated analytically by a least-squares method in a computationally efficient way. Then, we provide an eigenvector-based efficient implementation for performing unsupervised dimension reduction based on the QMI estimator. The usefulness of the proposed method is demonstrated through experiments.

Keywords

unsupervised dimension reduction, quadratic mutual information, least-squares density difference, Epanechnikov kernel, hyperparameter tuning.

1 Introduction

Dimension reduction is aimed at reducing the dimensionality of data, while preserving the "information" contained in the original data as much as possible. In this paper, we consider the problem of unsupervised dimension reduction where no label information is available. Unsupervised dimension reduction may be used for various purposes such as visualization and clustering, as well as a pre-processing step for supervised learning.

Below, we focus on linear dimension reduction where the dimension of the original data $\boldsymbol{x} \in \mathbb{R}^d$ is reduced by using a linear mapping $\boldsymbol{W} \in \mathbb{R}^{r \times d}$ as

$$\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x} \in \mathbb{R}^r,$$

where $1 \leq r \leq d$.

Principal component analysis (PCA) [3] finds the low-dimensional subspace retaining the maximum variance of the data. PCA is a classical linear unsupervised dimension reduction method, but it is still one of the most commonly used methods. However, due to its global preserving nature, local properties of the data such as clusters tend to be lost by PCA. *Locality preserving projection* (LPP) [1] seeks a linear transformation that well preserves the cluster structure of the original data. However, LPP contains tuning parameters for defining the local structure of the original data, and no objective method is available for tuning parameter selection. Consequently, the result obtained by LPP tends to be ad-hoc and subjective. Lack of objective model selection is actually a common drawback in many unsupervised dimension reduction methods [4].

In this paper, we address this issue by proposing an information-theoretic approach. More specifically, we adopt *quadratic mutual information* (QMI) [7] as our information measure, which is known to be more robust against outliers than ordinary mutual information:

$$\mathrm{QMI} := \iint \Big(p(\boldsymbol{z}, \boldsymbol{x}) - p(\boldsymbol{z})p(\boldsymbol{x})\Big)^2 \mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{x},$$

where $p(\boldsymbol{z}, \boldsymbol{x})$ is the joint density of $\boldsymbol{z}$ and $\boldsymbol{x}$, and $p(\boldsymbol{z})$ and $p(\boldsymbol{x})$ are the marginal densities. We find $\boldsymbol{W}$ so that QMI is maximized. Since $p(\boldsymbol{z}, \boldsymbol{x})$, $p(\boldsymbol{z})$, and $p(\boldsymbol{x})$ contained in QMI are unknown in practice, we utilize a least-squares QMI estimator called LSQMI [2] for developing a dimension reduction method. An advantage of LSQMI is that all tuning parameters can be objectively chosen based on cross-validation. Furthermore, by borrowing the idea from [8], we develop a computationally efficient algorithm for dimension reduction. Through experiments, we demonstrate the usefulness of our proposed method over competitive approaches.

2 LSQMI Estimation

In this section, we review a QMI estimator called *least-squares QMI* (LSQMI) [2].

Suppose that we are given a set of paired samples $\{(\boldsymbol{z}_i, \boldsymbol{x}_i)\}_{i=1}^n$ independently drawn from a joint probability distribution with density $p(\boldsymbol{z}, \boldsymbol{x})$. The key idea in LSQMI is to directly approximate the following density-difference function without density estimation of $p(\boldsymbol{z}, \boldsymbol{x})$, $p(\boldsymbol{z})$, and $p(\boldsymbol{x})$ [5]:

$$f(\boldsymbol{z}, \boldsymbol{x}) := p(\boldsymbol{z}, \boldsymbol{x}) - p(\boldsymbol{z})p(\boldsymbol{x}).$$

Let us model the density difference $f(\boldsymbol{z}, \boldsymbol{x})$ by

$$g(\boldsymbol{z}, \boldsymbol{x}) = \sum_{\ell=1}^{n} \boldsymbol{\theta}_\ell K(\boldsymbol{z}, \boldsymbol{z}_\ell) L(\boldsymbol{x}, \boldsymbol{x}_\ell),$$

where $K(\boldsymbol{z}, \boldsymbol{z}')$ and $L(\boldsymbol{x}, \boldsymbol{x}')$ are kernel functions for $\boldsymbol{z}$ and $\boldsymbol{x}$, respectively. Then $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)^\top$ is learned by least-squares as

$$\min_{\boldsymbol{\theta}} \iint \Big(g(\boldsymbol{z}, \boldsymbol{x}) - f(\boldsymbol{z}, \boldsymbol{x}) \Big)^2 \mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{x}. \tag{1}$$

An empirical and regularized version of the above optimization problem is given as

$$\widehat{\boldsymbol{\theta}} := \operatorname*{argmin}_{\boldsymbol{\theta}} \left[\boldsymbol{\theta}^\top \boldsymbol{H} \boldsymbol{\theta} - 2\boldsymbol{\theta}^\top \widehat{\boldsymbol{h}} + \lambda \boldsymbol{\theta}^\top \boldsymbol{\theta} \right],$$

where $\lambda \geq 0$ is the regularization parameter and $\boldsymbol{H}$ and $\widehat{\boldsymbol{h}}$ are defined as

$$H_{\ell,\ell'} := \int K(\boldsymbol{z}, \boldsymbol{z}_\ell) K(\boldsymbol{z}, \boldsymbol{z}_{\ell'}) \mathrm{d}\boldsymbol{z} \int L(\boldsymbol{x}, \boldsymbol{x}_\ell) L(\boldsymbol{x}, \boldsymbol{x}_{\ell'}) \mathrm{d}\boldsymbol{x},$$

$$\widehat{h}_\ell := \frac{1}{n} \sum_{i=1}^{n} K(\boldsymbol{z}_i, \boldsymbol{z}_\ell) L(\boldsymbol{x}_i, \boldsymbol{x}_\ell) - \frac{1}{n^2} \sum_{i,j=1}^{n} K(\boldsymbol{z}_i, \boldsymbol{z}_\ell) L(\boldsymbol{x}_j, \boldsymbol{x}_\ell).$$

The solution $\widehat{\boldsymbol{\theta}}$ can be obtained analytically as

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

where $\boldsymbol{I}$ denotes the identity matrix. Finally, following [2], a QMI estimator is given by

$$\widehat{\mathrm{QMI}} := \widehat{\boldsymbol{\theta}}^\top \widehat{\boldsymbol{h}}.$$

The performance of LSQMI depends on the choice of the regularization parameter λ and kernel parameters included $K(\boldsymbol{z}, \boldsymbol{z}')$ and $L(\boldsymbol{x}, \boldsymbol{x}')$. These tuning parameters can be systematically optimized based on cross-validation with respect to the objective function (1) as follows: First, the sample set $\mathcal{S} = \{(\boldsymbol{z}_i, \boldsymbol{x}_i)\}_{i=1}^n$ is divided into disjoint subsets $\{\mathcal{S}_m\}_{m=1}^M$ of (approximately) the same size. Then an estimator $\widehat{f}_m$ is obtained from $\mathcal{S} \backslash \mathcal{S}_m$ (i.e., all samples without $\mathcal{S}_m$), and its objective value is evaluated using the hold-out samples $\mathcal{S}_m$ as

$$\iint \widehat{f}_m(\boldsymbol{z}, \boldsymbol{x})^2 \mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{x} - \frac{2}{|\mathcal{S}_m|} \sum_{(\boldsymbol{z},\boldsymbol{x}) \in \mathcal{S}_m} \widehat{f}_m(\boldsymbol{z}, \boldsymbol{x}) + \frac{2}{|\mathcal{S}_m|^2} \sum_{\boldsymbol{z},\boldsymbol{x} \in \mathcal{S}_m} \widehat{f}_m(\boldsymbol{z}, \boldsymbol{x}),$$

where $|\mathcal{S}_m|$ denotes the number of elements in the set $\mathcal{S}_m$, $\sum_{(\boldsymbol{z},\boldsymbol{x}) \in \mathcal{S}_m}$ indicates the summation over every paired sample $(\boldsymbol{z}, \boldsymbol{x})$ in $\mathcal{S}_m$ (i.e., summation over $|\mathcal{S}_m|$ elements), and $\sum_{\boldsymbol{z},\boldsymbol{x} \in \mathcal{S}_m}$ indicates the summation over every unpaired samples $\boldsymbol{z}$ and $\boldsymbol{x}$ in $\mathcal{S}_m$ (i.e., summation over $|\mathcal{S}_m|^2$ combinations). This procedure is repeated for $m = 1, \ldots, M$, and the model that minimizes the average of the above hold-out error over all m is chosen as the best one.

3 Unsupervised Dimension Reduction with LSQMI

In this section, we propose a dimension reduction method based on LSQMI.

We focus on linear dimension reduction where original data samples $\{\boldsymbol{x}_i | \boldsymbol{x}_i \in \mathbb{R}^d\}_{i=1}^n$ are transformed by using a linear mapping $\boldsymbol{W} \in \mathbb{R}^{r \times d}$ as

$$\boldsymbol{z}_i = \boldsymbol{W}\boldsymbol{x}_i \in \mathbb{R}^r,$$

where $1 \leq r \leq d$. We assume that r is fixed in advance and $\boldsymbol{W}$ is an orthogonal matrix, i.e.,

$$\boldsymbol{W}\boldsymbol{W}^\top = \boldsymbol{I}.$$

We try to find $\boldsymbol{W}$ that maximizes $\widehat{\mathrm{QMI}}$:

$$\max_{\boldsymbol{W} \in \mathbb{R}^{r \times d}} \widehat{\mathrm{QMI}} \text{ s.t. } \boldsymbol{W}\boldsymbol{W}^\top = \boldsymbol{I}.$$

A local maximizer may be obtained by a gradient-projection method or a natural gradient method [6], but we consider a computationally more efficient approach based on [8], which is described below.

We use the Gaussian kernel for $\boldsymbol{x}$ and the Epanechnikov kernel for $\boldsymbol{z}$:

$$L(\boldsymbol{x}, \boldsymbol{x}') := \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\sigma_x^2}\right),$$
$$K(\boldsymbol{z}, \boldsymbol{z}') := \max\left(0, 1 - \frac{\|\boldsymbol{z} - \boldsymbol{z}'\|^2}{2\sigma_z^2}\right).$$

Here, we approximate the integral in $\boldsymbol{H}$ for the Epanechnikov kernel by using the analytic form of the Gaussian kernel. Then $\boldsymbol{H}$ can be computed as

$$H_{\ell,\ell'} \approx (\pi\sigma_z^2)^{d_z/2} \exp\left(-\frac{\|\boldsymbol{z}_\ell - \boldsymbol{z}_{\ell'}\|^2}{4\sigma_z^2}\right) (\pi\sigma_x^2)^{d_x/2} \exp\left(-\frac{\|\boldsymbol{x}_\ell - \boldsymbol{x}_{\ell'}\|^2}{4\sigma_x^2}\right).$$

Let $I(c)$ be the indicator function, i.e., $I(c) = 1$ if c is true and zero otherwise. Then, $\widehat{\mathrm{QMI}}$ can be expressed as

$$\widehat{\mathrm{QMI}} = \mathrm{tr}(\boldsymbol{W}\boldsymbol{D}\boldsymbol{W}^\top),$$

where $\mathrm{tr}(\cdot)$ is the trace of a matrix and

$$\begin{aligned}\boldsymbol{D} = \frac{1}{n}\sum_{i=1}^n \sum_{\ell=1}^n \widehat{\theta}_\ell(\boldsymbol{W}) I\left(\frac{\|\boldsymbol{W}\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{x}_\ell\|^2}{2\sigma_z^2} < 1\right) \\ \times L(\boldsymbol{x}_i, \boldsymbol{x}_\ell)\left[\frac{1}{r}\boldsymbol{I} - \frac{1}{2\sigma_z^2}(\boldsymbol{x}_i - \boldsymbol{x}_\ell)(\boldsymbol{x}_i - \boldsymbol{x}_\ell)^\top\right] \\ - \frac{1}{n^2}\sum_{i,j=1}^n \sum_{\ell=1}^n \widehat{\theta}_\ell(\boldsymbol{W}) I\left(\frac{\|\boldsymbol{W}\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{x}_\ell\|^2}{2\sigma_z^2} < 1\right) \\ \times L(\boldsymbol{x}_j, \boldsymbol{x}_\ell)\left[\frac{1}{r}\boldsymbol{I} - \frac{1}{2\sigma_z^2}(\boldsymbol{x}_i - \boldsymbol{x}_\ell)(\boldsymbol{x}_i - \boldsymbol{x}_\ell)^\top\right].\end{aligned}$$

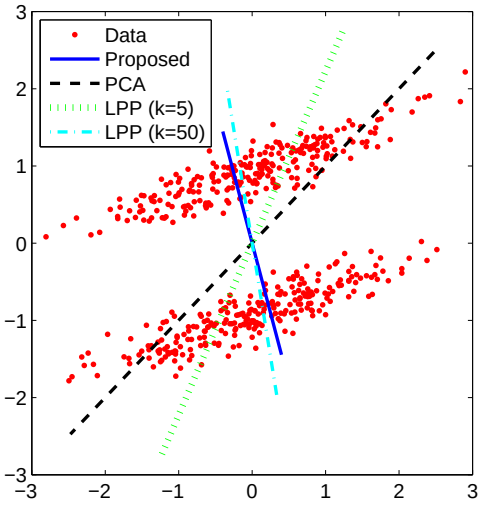


Figure 1: Dimension reduction for toy dataset.

Here, by $\widehat{\theta}_\ell(\boldsymbol{W})$, we explicitly indicated the fact that $\widehat{\theta}_\ell$ depends on $\boldsymbol{W}$.

Let us replace $\boldsymbol{D}$ in $\widehat{\mathrm{QMI}}$ by $\boldsymbol{D}'$, which is $\boldsymbol{D}$ with $\boldsymbol{W}$ replaced by the one obtained in the previous iteration:

$$\mathrm{tr}\left(\boldsymbol{W}\boldsymbol{D}'\boldsymbol{W}^\top\right).$$

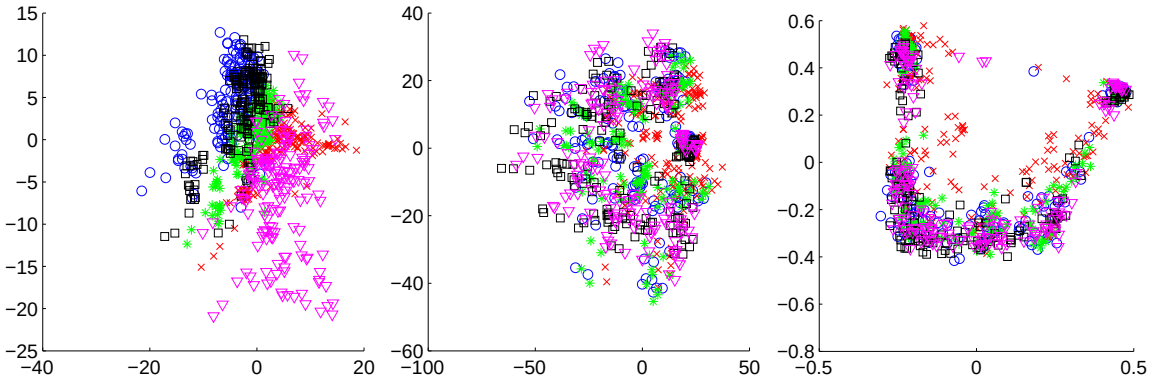
Its maximizer can then be analytically obtained as $(\boldsymbol{w}_1|...|\boldsymbol{w}_r)^\top$, where $\{\boldsymbol{w}_i\}_{i=1}^r$ are the r principal components of $\boldsymbol{D}'$.

We initialize $\boldsymbol{W}$ by the r principal components of $\boldsymbol{D}^{(0)}$ as $(\boldsymbol{w}_1^{(0)}|...|\boldsymbol{w}_r^{(0)})^\top$, where $\boldsymbol{D}^{(0)}$ is $\boldsymbol{D}$ with $\boldsymbol{z}$ replaced by $\boldsymbol{x}$.

4 Experiments

In this section, we compare the practical performance of the proposed method with PCA and LPP. In the proposed method, we choose the Gaussian width σ and the regularization parameter λ based on 5-fold cross-validation. In LPP, we use the Gaussian kernel for building the similarity matrix and the Gaussian width is set at the median of all pairwise sample distances. In LPP, we use the k-nearest neighbor similarity.

First, we illustrate the behavior of the dimension reduction methods using a 2-dimensional toy dataset in Figure 1. PCA only takes into account the global structure of the data and thus the cluster structure is lost. LPP tries to preserve the cluster structure, and LPP with $k = 50$ works relatively well for preserving clusters. However, LPP with $k = 5$ cannot separate the clusters well. Note that there is no objective method to choose k for LPP. On the other hand, tuning parameters in the proposed method can be objectively chosen by cross-validation and its performance is illustrated to be more reliable than other methods.



(a) Proposed (b) PCA (c) LPP $k = 50$

Figure 2: 2-dimensional embedding results for PIE face data.

Next, we use the *PIE face dataset*[1], which consists of 41,368 images of 68 people under 13 different poses, 43 different illumination conditions, and with 4 different expressions. We choose the images for 5 people and perform dimension reduction. The obtained 2-dimensional embedding results are exhibited in Figure 2, showing that the proposed method tends to preserve cluster structures corresponding to the true classes more clearly than PCA and LPP with $k = 50$.

Finally, we evaluate the clustering performance after dimension reduction using the *UCI benchmark datasets*[2]. For randomly chosen 90% samples from each dataset, we apply a dimension reduction method and then perform clustering by the k-means algorithm[3]. Then the clustering accuracy is evaluated. Figure 3 depicts the mean and standard deviation of the clustering accuracy over 10 runs, showing that the proposed method overall performs well.

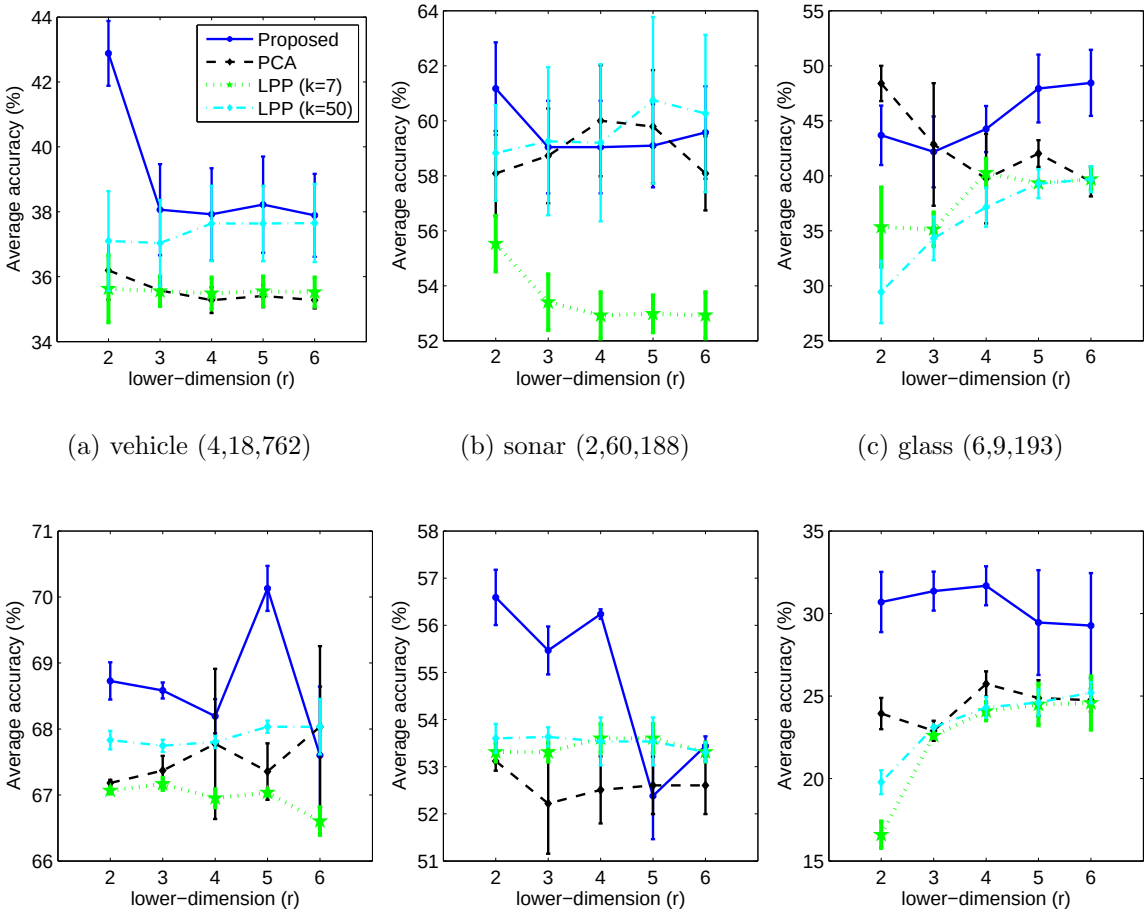
5 Conclusion

Tuning parameter selection has been an important problem in unsupervised dimension reduction. In this paper, we addressed this issue by applying least-squares quadratic mutual information (LSQMI) to unsupervised dimension reduction, which allows objective model selection based on cross-validation. Thanks to the high robustness of QMI against outliers and the computationally efficient implementation based on eigendecomposition, the proposed method was demonstrated to be useful through experiments.

[1] http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html

[2] http://www.ics.uci.edu/~mlearn/MLRepository.html

[3] We run the k-means algorithm 10 times with random initialization and chose the best solution with the minimum objective value.



(a) vehicle (4,18,762) (b) sonar (2,60,188) (c) glass (6,9,193)

(d) pima (2,8,692) (e) liver-disorders (2,6,311) (f) vowel (11,13,892)

Figure 3: Mean and standard deviation of the clustering accuracy over 10 runs. The three digits (c, d, n) show the original dimension d, the number of clusters c, and the number of samples n.

Acknowledgments

JS was supported by the Ph.D. scholarship from Prince of Songkla University, and MS is supported by MEXT KAKENHI 25700022 and AOARD.

References

[1] X. He and P. Niyogi, "Locality preserving projections," Advances in Neural Information Processing Systems. pp. 153–160, 2004.

[2] S. Janya and M. Sugiyama, "Direct approximation of quadratic mutual information and its application to dependence-maximization clustering," IEICE Transactions on Information & Systems, Vol. E96-D, No. 10, pp. 2282–2285, 2013.

[3] I. Jolliffe, "Principal component analysis," Springer-Verlag, 1986.

[4] J. A. Lee and M. Verleysen, "Unsupervised dimensionality reduction: overview and recent advances," Proceedings of International Joint Conference on Neural Networks, pp. 1–8, 2010.

[5] M. Sugiyama, T. Suzuki, T. Kanamori, M.C. du Plessis, S. Liu, and I. Takeuchi, "Density-difference estimation," Neural Computation. vol. 25, no. 10, pp. 2734–2775, 2013.

[6] T. Suzuki and M. Sugiyama, "Sufficient dimension reduction via squared-loss mutual information estimation," Neural Computation. vol. 25, no. 3, pp. 725–758, 2013.

[7] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," Journal of Machine Learning Research, vol. 3, pp. 1415–1438, 2003.

[8] M. Yamada, G. Niu, J. Takagi, and M. Sugiyama, "Computationally efficient sufficient dimension reduction via squared-loss mutual information," JMLR Workshop and Conference Proceedings, vol. 20, pp. 247-262, 2011.