Bias Reduction and Metric Learning for Nearest-Neighbor Estimation of Kullback-Leibler Divergence

Yung-Kyun Noh[1], **Masashi Sugiyama**[2], **Song Liu**[2], **Marthinus C. du Plessis**[2], **Frank Chongwoo Park**[3], **and Daniel D. Lee**[4]
[1]KAIST, Korea; [2]Tokyo Institute of Technology, Japan;
[3]Seoul National University, Korea; [4]University of Pennsylvania, USA

Abstract

Asymptotically unbiased nearest-neighbor estimators for KL divergence have recently been proposed and demonstrated in a number of applications. With small sample sizes, however, these nonparametric methods typically suffer from high estimation bias due to the non-local statistics of empirical nearest-neighbor information. In this paper, we show that this non-local bias can be mitigated by changing the distance metric, and we propose a method for learning an optimal Mahalanobis-type metric based on global information provided by approximate parametric models of the underlying densities. In both simulations and experiments, we demonstrate that this interplay between parametric models and nonparametric estimation methods significantly improves the accuracy of the nearest-neighbor KL divergence estimator.

1 Introduction

We consider the problem of estimating the *Kullback-Leibler (KL) divergence* between probability density functions $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$,

$$\mathrm{KL}(p_1\|p_2) = -\int p_1(\mathbf{x}) \log\left(\frac{p_2(\mathbf{x})}{p_1(\mathbf{x})}\right) d\mathbf{x}, \quad \mathbf{x} \in \mathbb{R}^D, \quad (1)$$

based on two sets of i.i.d. samples $\mathcal{X}_1 = \{\mathbf{x}_1, \ldots, \mathbf{x}_{N_1}\}$ and $\mathcal{X}_2 = \{\mathbf{x}_{N_1+1}, \ldots, \mathbf{x}_{N_1+N_2}\}$ generated from $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$, respectively.

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

KL divergence estimates can be used for a variety of machine learning applications, such as homogeneity testing of an underlying density function [7], dependency testing for feature selection [1, 2, 15, 23], state change detection [8, 11], and model parameter estimation [19]. Moreover, a number of other information-theoretic measures, such as entropy and Jensen-Shannon divergence, can be expressed using the KL divergence. Therefore, a properly designed KL divergence estimator is useful for a range of general statistical applications.

Current approaches to estimating the KL divergence can roughly be divided into parametric and nonparametric approaches. The parametric approach uses prespecified density models, and computes a closed-form approximation to the KL divergence by substituting estimated parameters into the density models. For example, under the Gaussian assumption on both p_1 and p_2, estimated means $\widehat{\mu}_1, \widehat{\mu}_2 \in \mathbb{R}^D$, and covariance matrices $\widehat{\Sigma}_1, \widehat{\Sigma}_2 \in \mathbb{R}^{D\times D}$ for p_1 and p_2 can be plugged into the following closed-form estimator:

$$\widehat{\mathrm{KL}}(p_1\|p_2) = \frac{1}{2}\Big[\log|\widehat{\Sigma}_2| - \log|\widehat{\Sigma}_1| + \mathrm{tr}\left[\widehat{\Sigma}_1\widehat{\Sigma}_2^{-1}\right] + \mathrm{tr}\left[(\widehat{\mu}_1-\widehat{\mu}_2)(\widehat{\mu}_1-\widehat{\mu}_2)^T\widehat{\Sigma}_2^{-1}\right] - D\Big], \quad (2)$$

where $|\widehat{\Sigma}|$ denotes the determinant of matrix $\widehat{\Sigma}$. This approximation and its simplified variant [1, 3] are computationally efficient and stable, and they are accurate when the true distributions are close to Gaussians. However, when the parametric assumption is violated, such plug-in estimators fail to capture the information that cannot be expressed using model parameters, resulting in failure to asymptotically converge to the true divergence.

In contrast, the nonparametric approach does not make any assumptions about appropriate density models. A popular non-parametric approach utilizes

nearest-neighbor distances [10, 16, 22]:

$$\widehat{\mathrm{KL}}(p_1\|p_2) = \frac{1}{N_1}\sum_{i=1}^{N_1}\log\frac{u_2(\mathbf{x}_i)}{u_1(\mathbf{x}_i)}, \tag{3}$$

where $u_1(\mathbf{x}_i) = (N_1-1)d_1(\mathbf{x}_i)^D$ and $u_2(\mathbf{x}_i) = N_2 d_2(\mathbf{x}_i)^D$ with $d_1(\mathbf{x}_i)$ and $d_2(\mathbf{x}_i)$ denoting the distance from $\mathbf{x}_i$ in $\mathcal{X}_1$ to the nearest neighbors in $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively. Previous work has shown that $\frac{1}{u_1(\mathbf{x})}$ and $\frac{1}{u_2(\mathbf{x})}$ are consistent estimators of $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ up to a common normalization factor [12]. However, these individual convergences do not necessarily imply convergence of the ratio; recently, it has been proved that the estimator (3) satisfies both the almost sure convergence and the L_2 convergence to the true KL divergence (1) [9, 10, 17, 18]. A related approach that directly estimates the density ratio p_1/p_2 nonparametrically was shown to achieve the optimal non-parametric convergence rate in the minimax sense [13, 21].

Unfortunately, these nonparametric methods still suffer high bias caused by the finite number of samples. For this reason, when only a small number of samples are available, parametric methods are typically more reliable than nonparametric methods. In this paper, we show that a nonparametric estimator can be significantly improved when using a finite number of samples using parametric model information. More specifically, the bias of a nonparametric estimator can be reduced using parametric models by learning an appropriate *Mahalanobis-type metric* for nearest neighbor selection. Due to metric-invariance of the KL divergence itself, convergence of the nearest-neighbor estimator to the true divergence is guaranteed regardless of the metric. The existing convergence property of the estimator and its proof for the Euclidean distance can be applied without any additional assumptions.

The remainder of the paper is organized as follows. In Section 2, we explain KL divergence estimation and review current state-of-the-art nonparametric methods for estimating the divergence. Then, we derive the finite sampling bias error and show how a metric for minimizing the bias can be learned. In Section 3, we provide experimental results using many synthetic and real datasets showing how our proposed method can improve the accuracy over other methods. Finally, we conclude in Section 4 with a discussion.

2 Metric Learning for Nearest-Neighbor KL Divergence Estimation

In this section, we analyze the bias of nearest-neighbor estimation for the KL divergence and show that it depends on the distance metric. We then propose a metric learning algorithm to minimize this bias using parametric information.

2.1 Bias of Nearest-Neighbor Density Estimation

Let us consider a D-dimensional density estimation problem from samples $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ following $p(\mathbf{x})$. A simple method of approximating the KL divergence is introduced in Poczos et al. [17] which uses a plug-in method of the probability density estimator of Loftsgaarden et al. [12]. The nearest-neighbor density estimator $\widehat{p}(\mathbf{x})$ is given by

$$\widehat{p}(\mathbf{x}) = \frac{1}{\gamma u(\mathbf{x})}, \tag{4}$$

$$\text{where}\quad \gamma = \frac{\pi^{D/2}}{\Gamma(D/2+1)}, \tag{5}$$

$$u(\mathbf{x}) = N d(\mathbf{x})^D, \tag{6}$$

$$d(\mathbf{x}) = \|\mathbf{x}_{\mathrm{NN}} - \mathbf{x}\|, \tag{7}$$

where $\Gamma(\cdot)$ denotes the Gamma function, and $\mathbf{x}_{\mathrm{NN}}$ denotes the nearest neighbor in $\mathcal{X}$ from $\mathbf{x}$, and $d(\mathbf{x})$ is the distance to the nearest neighbor.

In order to obtain a bias of the KL divergence estimator, we first consider how a bias perturbs the Loftsgaarden's estimator. When the underlying probability density function $p(\mathbf{x})$ is twice-differentiable, the Taylor expansion of $p(\mathbf{x}_{\mathrm{NN}})$ around $\mathbf{x}$ gives

$$p(\mathbf{x}_{\mathrm{NN}}) \simeq p(\mathbf{x}) + \nabla p(\mathbf{x})^\top(\mathbf{x}_{\mathrm{NN}} - \mathbf{x}) + \tag{8}$$

$$\frac{1}{2}(\mathbf{x}_{\mathrm{NN}} - \mathbf{x})^\top \nabla\nabla p(\mathbf{x})(\mathbf{x}_{\mathrm{NN}} - \mathbf{x}), \tag{9}$$

where $\nabla\nabla p(\mathbf{x})$ denotes the Hessian of $p(\mathbf{x})$. In this expansion, we consider the nearest neighbors to be at a nonzero distance from the point of interest $\mathbf{x}$, which is a situation where we have a finite number of data.

In this finite sample situation, the bias of the nearest-neighbor density estimator $\widehat{p}(\mathbf{x})$ can be obtained by utilizing the average densities on the surface of the D-dimensional hyper-sphere with center $\mathbf{x} \in \mathbb{R}^D$ and radius $\|\mathbf{x}_{\mathrm{NN}} - \mathbf{x}\|$. A detailed calculation of the average over the surface can be found in the Appendix A, and from the average, the bias can be obtained:

$$\mathrm{Bias}[\widehat{p}(\mathbf{x})] := \mathbb{E}[p(\mathbf{x}_{\mathrm{NN}})] - p(\mathbf{x}) \tag{10}$$

$$\simeq \frac{\mathbb{E}[\|\mathbf{x}_{\mathrm{NN}} - \mathbf{x}\|^2]}{2D}\nabla^2 p(\mathbf{x}), \tag{11}$$

where $\mathbb{E}[\cdot]$ denotes the expectation over nearest neighbor $\mathbf{x}_{\mathrm{NN}}$ from $\mathbf{x}$ and $\nabla^2 p(\mathbf{x}) = \mathrm{tr}(\nabla\nabla p(\mathbf{x}))$ denotes the Laplacian at $\mathbf{x}$.

The bias can be further calculated using the density function of the nearest-neighbor distance. It is known

[6, 22] that the value $u(\mathbf{x}_{\mathrm{NN}})$ converges in distribution to the exponential distribution $\rho(u) = \lambda \exp(-\lambda u)$ with rate parameter $\lambda = \gamma p(\mathbf{x}_{\mathrm{NN}})$. Using this asymptotic density function, the expectation can be calculated:

$$\mathbb{E}[\|\mathbf{x}_{\mathrm{NN}} - \mathbf{x}\|^2] = E_{u\sim\rho(u)}\left[\left(\frac{u}{N}\right)^{\frac{2}{D}}\right] \quad (12)$$

$$= \frac{1}{(\gamma N p(\mathbf{x}))^{2/D}}, \quad (13)$$

and thus the bias can be approximated as

$$\mathrm{Bias}[\widehat{p}(\mathbf{x})] \simeq \alpha \nabla^2 p(\mathbf{x}), \quad (14)$$

$$\text{where} \quad \alpha = \frac{1}{2D(\gamma N p(\mathbf{x}))^{2/D}}. \quad (15)$$

According to the equation, the bias depends on the curvature of the underlying density function, $\nabla^2 p(\mathbf{x})$, and α vanishes if N goes to infinity and $p(\mathbf{x})$ is away from zero. If either the underlying density function has a small curvature or N is large, the bias tends to be small.

2.2 Bias of Nearest-Neighbor KL Divergence Estimation

Turning back to the KL divergence estimation, the plug-in of Eq.(4) into the definition of $\mathrm{KL}(p_1\|p_2)$ yields the nearest-neighbor KL divergence estimator in Eq.(3) using two sets of i.i.d. samples $\mathcal{X}_1 = \{\mathbf{x}_1, \ldots, \mathbf{x}_{N_1}\} \sim p_1(\mathbf{x})$ and $\mathcal{X}_2 = \{\mathbf{x}_{N_1+1}, \ldots, \mathbf{x}_{N_1+N_2}\} \sim p_2(\mathbf{x})$.

Based on the bias analysis of nearest-neighbor density estimation shown above, we can analyze the bias of nearest-neighbor KL divergence estimation using perturbation as

$$\log \frac{u_2(\mathbf{x})}{u_1(\mathbf{x})} \longrightarrow -\log \frac{p_2(\mathbf{x}) + \alpha_2 \nabla^2 p_2(\mathbf{x})}{p_1(\mathbf{x}) + \alpha_1 \nabla^2 p_1(\mathbf{x})} \quad (16)$$

$$= -\log \frac{p_2(\mathbf{x})}{p_1(\mathbf{x})} + \left(\log \frac{p_1(\mathbf{x}) + \alpha_1 \nabla^2 p_1(\mathbf{x})}{p_1(\mathbf{x})} - \log \frac{p_2(\mathbf{x}) + \alpha_2 \nabla^2 p_2(\mathbf{x})}{p_2(\mathbf{x})}\right) \quad (17)$$

$$\simeq -\log \frac{p_2(\mathbf{x})}{p_1(\mathbf{x})} + \left(\frac{\alpha_1 \nabla^2 p_1(\mathbf{x})}{p_1(\mathbf{x})} - \frac{\alpha_2 \nabla^2 p_2(\mathbf{x})}{p_2(\mathbf{x})}\right), \quad (18)$$

where we use $\log(1+t) \simeq t$, and[1]

$$\alpha_1 = \frac{1}{2D(\gamma(N_1 - 1)p_1(\mathbf{x}))^{2/D}}, \quad (19)$$

$$\alpha_2 = \frac{1}{2D(\gamma N_2 p_2(\mathbf{x}))^{2/D}}. \quad (20)$$

Thus, the bias of $\widehat{g}(\mathbf{x}) = \log \frac{u_2(\mathbf{x})}{u_1(\mathbf{x})}$ is given as

$$\mathrm{Bias}(\widehat{g}(\mathbf{x})) \simeq \frac{\alpha_1 \nabla^2 p_1(\mathbf{x})}{p_1(\mathbf{x})} - \frac{\alpha_2 \nabla^2 p_2(\mathbf{x})}{p_2(\mathbf{x})}. \quad (21)$$

Note that this bias is pointwise, and the total bias can be approximated as the expectation of them. The coefficients α_1 and α_2 become zero with infinite samples $N_1 = N_2 = \infty$, yielding a consistent estimation of the true KL divergence, but the convergence rate is slow in high dimensional space; with N number of data the bias decreases with rate $N^{-\frac{2}{D}}$. However, with proper choice of metric, the pointwise bias can be significantly reduced even with small N_1 and N_2 by changing the Laplacians $\nabla^2 p_1(\mathbf{x})$ and $\nabla^2 p_2(\mathbf{x})$.

2.3 Metric Learning for Bias Reduction

The above analysis shows that the bias of nearest-neighbor KL divergence estimation not only depends on the number of data but also on the curvatures of the underlying density functions p_1 and p_2. In this section, we show how the bias can be reduced by appropriately learning the distance metric.

We use a Mahalanobis-type distance metric parameterized by a positive definite symmetric matrix $A \in \mathbb{R}^{D\times D}$: the distance between $\mathbf{x} \in \mathbb{R}^D$ and $\mathbf{x}_{NN} \in \mathbb{R}^D$ is

$$d(\mathbf{x}, \mathbf{x}_{\mathrm{NN}}) = \sqrt{(\mathbf{x} - \mathbf{x}_{\mathrm{NN}})^\top A (\mathbf{x} - \mathbf{x}_{\mathrm{NN}})}, \quad (22)$$

$$A^\top = A, \quad A \succ 0. \quad (23)$$

The bias changes according to the choice of the metric[2], and it is straightforward that the bias is minimized by solving the following semidefinite program:

$$\min_A \left(\mathrm{tr}[A^{-1}B]\right)^2 \quad (24)$$

$$\text{s.t.} \quad A^\top = A, \ |A| = 1, \ \text{and} \ \ A \succ 0, \quad (25)$$

where the symmetric matrix B is defined using the Hessians:

$$B = \frac{1}{((N_1 - 1)p_1(\mathbf{x}))^{\frac{2}{D}}} \frac{\nabla\nabla p_1(\mathbf{x})}{p_1(\mathbf{x})} \quad (26)$$

$$- \frac{1}{(N_2 p_2(\mathbf{x}))^{\frac{2}{D}}} \frac{\nabla\nabla p_2(\mathbf{x})}{p_2(\mathbf{x})}. \quad (27)$$

[1] We used $N_1 - 1$ instead of N_1 because one degree of freedom was used for estimating the expectation over p_1 in Eq.(1).

[2] We note that the KL divergence itself is metric-invariant. Due to the positive definiteness of A, we can always find a full rank matrix L such that $A = LL^\top$. Then the metric change can be regarded as a linear transformation of variables $\mathbf{z} = L^\top \mathbf{x}$ which uses the Euclidean metric in the transformed $\mathbf{z}$-space. In this case, because $p(\mathbf{z}) = p(\mathbf{x})/|L|$ and $d\mathbf{z} = |L| d\mathbf{x}$, we have $-\int p_1(\mathbf{z}) \log\left(\frac{p_2(\mathbf{z})}{p_1(\mathbf{z})}\right) d\mathbf{z} = -\int p_1(\mathbf{x}) \log\left(\frac{p_2(\mathbf{x})}{p_1(\mathbf{x})}\right) d\mathbf{x}$.

The solution of this semidefinite program can be obtained analytically. However, it is not unique, and the solution we selected is the following as in [14]:

$$A = \beta \begin{bmatrix} U_+ & U_- \end{bmatrix} \begin{pmatrix} d_+\Lambda_+ & 0 \\ 0 & -d_-\Lambda_- \end{pmatrix} \begin{bmatrix} U_+ & U_- \end{bmatrix}^\top, \quad (28)$$

where $\Lambda_+ \in \mathbb{R}^{d_+ \times d_+}$ and $\Lambda_- \in \mathbb{R}^{d_- \times d_-}$ are the diagonal matrices containing d_+ positive and d_- negative eigenvalues of B, respectively. The matrices A and B share the same eigenvectors, and $U_+ \in \mathbb{R}^{D \times d_+}$ is the collection of eigenvectors that correspond to the eigenvalues in Λ_+, and $U_- \in \mathbb{R}^{d_- \times d_-}$ corresponds to the eigenvalues in Λ_-.

The solution of Eq.(25) is not unique, but the metric Eq.(28) is a well-behaving solution that provides the minimum deviation for any number of positive and negative eigenvalues. The scale constant β does not affect the estimation result, but we use β to satisfy $|A| = 1$ for numerical stability of optimization. The detailed procedure of derivation for obtaining this particular matrix is in the Appendix B.

In order to obtain matrix B, we use (rough) parametric models of the underlying densities $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$. For example, when Gaussian models are considered for $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$, we can explicitly obtain the matrix B using

$$\frac{\nabla\nabla p_c(\mathbf{x})}{p_c(\mathbf{x})} = \widehat{\Sigma}_c^{-1} (\mathbf{x} - \widehat{\mu}_c) (\mathbf{x} - \widehat{\mu}_c)^\top \widehat{\Sigma}_c^{-1} - \widehat{\Sigma}_c^{-1}, \quad \text{for } c = 1, 2, \quad (29)$$

where $\widehat{\mu}_1$, $\widehat{\mu}_2$, $\widehat{\Sigma}_1$, and $\widehat{\Sigma}_2$ are estimates of means and covariance matrices. Throughout the experiment, we use maximum-likelihood estimated solutions.

At each sample $\mathbf{x}_i$ in $\mathcal{X}_1$, we calculate the metric to obtain $\widehat{g}(u_1, u_2) = \log \frac{u_2}{u_1}$, and then use the Monte-Carlo summation to estimate the KL divergence. The procedure of the estimation is summarized in Algorithm 1.

Algorithm 1 Nearest-Neighbor Estimation of KL Divergence with Metric Learning

Input: $\mathcal{X}_1 = \{\mathbf{x}_1, \ldots, \mathbf{x}_{N_1}\} \sim p_1(\mathbf{x})$ and $\mathcal{X}_2 = \{\mathbf{x}_{N_1+1}, \ldots, \mathbf{x}_{N_1+N_2}\} \sim p_2(\mathbf{x})$
Output: $\widehat{\mathrm{KL}}(p_1 \| p_2)$
Procedure:
Estimate parameters of generative models
$KL = 0$
for $i = 1$ **to** N_1 **do**
  Calculate A at $\mathbf{x}_i$ with estimated parameters
  Calculate u_1 and u_2 using A
  $KL = KL + \frac{1}{N_1} \log u_2/u_1$
end for
$\widehat{\mathrm{KL}}(p_1 \| p_2) = KL$
End procedure:

3 Experiments

In this section, we provide experimental results to illustrate how our method estimates the KL divergence in various problems.

3.1 KL Divergence Estimation from Synthetic Data

We compare the performance of the following four non-parametric methods:

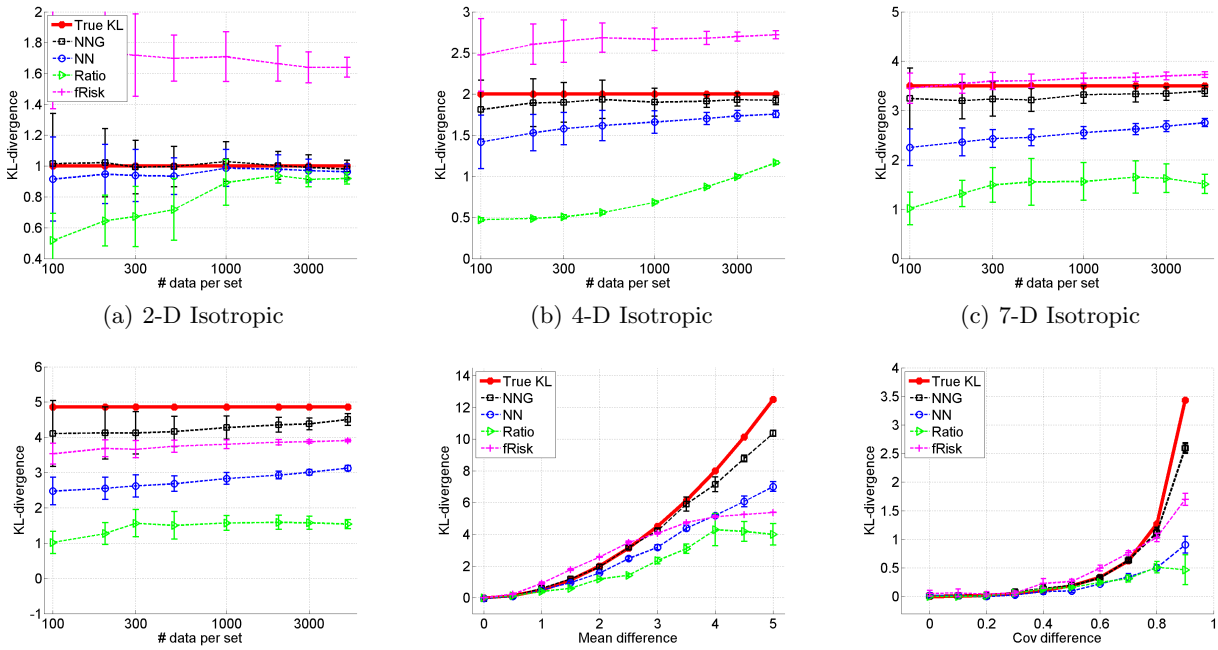
- Proposed nearest-neighbor KL divergence estimator with the Gaussian metric (NNG).
- Plain nearest-neighbor KL divergence estimator without metric learning (NN).
- State-of-the-art density-ratio KL divergence estimator (Ratio) [13].
- Risk-based nearest-neighbor KL divergence estimator (fRisk) [5].

First, we consider the estimation of KL divergence between two Gaussian densities in various shapes and with various dimensionalities. Fig. 1(a)–(d) depicts the experimental results for isotropic-isotropic and isotropic-correlated Gaussians with increasing numbers of samples. Fig. 1(e)–(f) shows how the estimation results increase with the increase of the mean difference and covariance difference for a fixed number of data samples ($N_1 = N_2 = 500$). Compared with the true KL divergence for two Gaussian densities, the proposed NNG always provides a significant improvement in accuracy, and NN and Ratio tend to under-estimate the true KL divergence.

Next we consider KL divergence estimation between two Student-t distributions. A one-dimensional Student's t density function is defined as

$$p(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi\sigma^2}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{(x-\mu)^2}{\nu\,\sigma^2}\right)^{-\frac{\nu+1}{2}}, \quad (30)$$

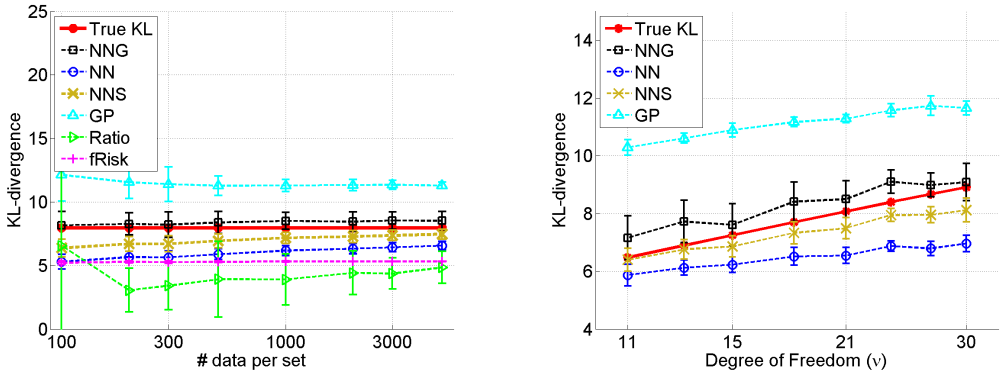
with location and scale parameters μ and σ^2. The overall shape of the Student-t distribution differs according to the parameter ν, representing the degree of freedom. With high ν, the Student-t distribution is known to have a shape close to Gaussian. As ν decreases, the function has heavier tails. We used a 5-dimensional density function with independent marginal functions, each of which is represented by Eq.(30). In addition to NNG, NN, and Ratio, we include in our comparison the following:



(a) 2-D Isotropic (b) 4-D Isotropic (c) 7-D Isotropic

(d) 7-D Correlated (e) For Various Mean Differences (f) For Various Covariance Differences

Figure 1: KL divergence estimation for two Gaussian probability densities. True KL divergence is calculated using analytical integration of two Gaussians. Here, NNG uses the true generative model with estimated parameters in this experiment.



(a) Degree of Freedom 20 (b) For Various Degrees of Freedoms

Figure 2: KL divergence estimation for two 5-D Student-t probability densities. NNS uses the true generative model in this experiment.

- Proposed nearest-neighbor KL divergence estimator with the Student-t metric (NNS).
- Gaussian parametric method (GP).

Here, GP uses Eq.(2) with maximum-likelihood estimated parameters.

Fig. 2(a) shows the estimation results for 5-dimensional Student-t data with $\nu = 20$, as functions of the number of data samples. As can be expected, NNS shows quick convergence to the true KL divergence. NN shows better convergence than Ratio in this 5-dimensional experiment. The large deviation of GP is due to the inaccuracy of the Gaussian assumption. However, even in this situation, NNG still shows comparable accuracies to NNS which use the true model. This robustness using an inaccurate model illustrates the usefulness of the proposed method in practice.

In Fig. 2(b), we show the estimation results for different degrees of freedom in Student-t distributions.

Table 1: Area under the ROC curve (AUC) of discriminating change points. In addition to original data, data corrupted with Gaussian and Poisson noise are used. The methods with the best accuracy are starred, and the accuracies within p-value=0.05 of single-sided T-test are written in bold.

	NNG	NN	GP	fRisk	KLIEP
Original Data	***0.810 (0.052)**	0.759 (0.051)	**0.787 (0.032)**	**0.793 (0.047)**	0.576 (0.055)
Gaussian noise	**0.550 (0.055)**	0.510 (0.082)	***0.608 (0.078)**	0.543 (0.078)	0.491 (0.061)
Poisson noise	***0.735 (0.028)**	**0.709 (0.032)**	0.599 (0.060)	**0.729 (0.043)**	0.511 (0.041)

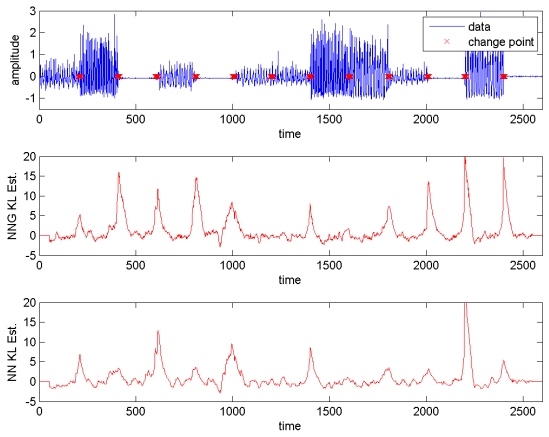


Figure 3: One example of time series in HASC and the estimated KL divergence for NNG (middle figure) and NN (bottom figure)

With other parameters fixed, the increase in the degree of freedom in Student-t distributions lessens the overlap between the two density functions; hence, the true KL divergence (red) grows with the increase of ν. The NNS estimator, using the true model with estimated scale and location, achieves a very accurate estimation of the true KL divergence.

NN performs poorly if ν is large. Considering that the curvature increases as ν gets large, our theoretical analysis that the bias comes from the curvature is also experimentally supported. Finally, GP deviation reduces as we increase ν because the true Student-t distribution approaches Gaussian. However, we note that NNG approaches the true KL divergence even more quickly with the increase of ν.

3.2 Change Detection in Time Series

The objective of discovering change points is to detect abrupt changes in the time series property. The changing property can be simply mean or variance, but sometimes more complex properties can change even with the same mean and the same variance due to the change in dependency.

We consider a column vector of length m, $\mathbf{y}(t) \in \mathbb{R}^m$ to represent a segment of time series at time t, and a collection of r such vectors is obtained from a sliding window: $\mathbf{Y}(t) := \{\mathbf{y}(t), \mathbf{y}(t+1), \ldots, \mathbf{y}(t+r-1)\}$. According to [8], we can consider an underlying density function that generates the retrospective r number of vectors in $\mathbf{Y}(t)$. We measure the KL divergence of the underlying density functions of the two sets, $\mathbf{Y}(t)$ and $\mathbf{Y}(t+r+m)$ for every t, and a point $t_0 + r + m$ is determined as a change point if the KL divergence for $\mathbf{Y}(t_0)$ and $\mathbf{Y}(t_0 + r + m)$ is greater than a predefined threshold.

We use the *Human Activity Sensing Consortium (HASC) Challenge 2011* collection[3] which provides human activity information collected by a portable three-axis accelerometer. Our task is to segment different behaviors such as "stay," "walk," "jog," and "skip." Because the orientation of the accelerometer is not necessarily fixed, we took the ℓ_2-norm of the 3-dimensional accelerometer data.

Fig.3 shows one example of the time series in HASC and its change points. We measured the KL divergence using NNG, NN, GP, fRisk, and KLIEP, and compared the classification accuracies to determine whether or not a point is a change point within a small tolerance region (± 10 from change point). The classification is performed for various thresholds and the area under the ROC curve (AUC) scores are reported in Table 1.

In Fig.3, the estimated KL divergences for NNG and its original algorithm, NN, show similar tendencies, but there are apparent differences in several change point regimes, where NNG always captures the change point more clearly.

In Table 1, we show the accuracies of NNG, NN, GP, fRisk, and KLIEP for change point detection using the HASC dataset. We use these measures on behalf of KL divergence and expect that the estimators more accurately measuring the KL divergence will perform better. Indeed, our NNG method outperforms other methods not only with the original dataset but also

[3] `http://hasc.jp/hc2011/`

with the data corrupted with noise. With Gaussian noise, all algorithms show similar accuracy decrease, while accuracy drop in GP is relatively small indicating that the information in mean and covariance difference is relatively intact by Gaussian noise. On the other hand, we can observe that the Poisson noise severely corrupts the mean and covariance information making the parametric GP estimator fail.

Surprisingly, fRisk performed well in this change point detection experiment, whereas it mostly failed in estimating the KL divergence with synthetic data. The difference of fRisk from other estimators is that this estimator is generally insensitive to the change of KL divergence when the true KL divergence is high as seen in Fig.1(e),(f), and this property can work as a regularizer preventing the estimator from making severe mistakes.

3.3 Feature Selection Using Jensen-Shannon Divergence Estimation

Finally, KL divergence estimation is applied to the selection of relevant features for classification.

3.3.1 Setup

For feature selection, t-score has been conveniently used as a selection criterion for the relevant feature selection by providing the mean difference $|\widehat{\mu}_1 - \widehat{\mu}_2|$ of two classes relative to the size of the variances $\widehat{\sigma}_1$ and $\widehat{\sigma}_2$:

$$\text{t-score} = \frac{|\widehat{\mu}_1 - \widehat{\mu}_2|}{\sqrt{\frac{\widehat{\sigma}_1^2}{N_1} + \frac{\widehat{\sigma}_2^2}{N_2}}}. \tag{31}$$

However, this criterion cannot capture the correlation and redundancy between features, and recently, a parametric approximation of the Jensen-Shannon (JS) divergence, mIMR, has been suggested as the feature selection criterion [1]. The JS divergence is an information-theoretic measure, which is also known as the Shannon mutual information between the labels $y \in \{1, 2\}$ and the data $\mathcal{X}$:

$$\text{JS}(\mathcal{X}; \mathbf{y}) = -\sum_{y=1}^{2} \int p(\mathbf{x}, y) \log \frac{p(\mathbf{x})p(y)}{p(\mathbf{x}, y)} d\mathbf{x}. \tag{32}$$

In this work, we consider another form of this measure, which is the sum of two KL divergences:

$$\text{JS}(\mathcal{X}; \mathbf{y}) = \tag{33}$$

$$\gamma_1 \text{KL}(p_1(\mathbf{x}) \| p(\mathbf{x})) + \gamma_2 \text{KL}(p_2(\mathbf{x}) \| p(\mathbf{x})), \tag{34}$$

where $p(\mathbf{x}) = \gamma_1 p_1(\mathbf{x}) + \gamma_2 p_2(\mathbf{x})$ with class-priors γ_1 and γ_2 and class-conditional densities $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ for class 1 and class 2, respectively.

In contrast to the previous synthetic Gaussian experiment, JS divergence does not have an analytic solution for two Gaussian class-conditional densities. The calculation of KL divergence is now between a Gaussian and a Gaussian mixture. However, even in this problem, our method can be used without any further approximation, because our method does not need any analytic integration but only needs twice-differentiation of densities. Throughout experiments in this section, we choose the Gaussian model for class-conditional densities, and we use the differentiation of a Gaussian for the differentiation of a Gaussian mixture.

3.3.2 Feature selection in high-dimensional Gaussian

We first prepare two 1000-dimensional (1000-D) Gaussian densities of which only 30 dimensionalities differ. After we obtain a 1000-D random Gaussian for common use of both classes, additional 30-D random mean and random covariance are later added only to the first 30 dimensionalities of one class.

With the entire 1000-dimensional features, the true discriminative information is easily lost, and many algorithms simply learn from data only a little better than the random choice of classes.

In Fig.4, we depict the classification accuracies for different Gaussian configurations after feature selection using JS divergence from three methods: NNG, NN, and mIMR, and one conventional criterion: t-score. At each realization, we increased the mean difference of two Gaussians of the informative dimensionalities with the same Gaussian configuration (covariance structure), and the results are averaged for each mean distance. Fig.4 shows the graph for the mean difference vs. the averaged accuracy.

Regardless of the mean difference, NNG and NN capture many dimensionalities among 30 informative dimensionalities yielding good classification results, while mIMR and t-score capture the informative dimensionalities only when the means of two Gaussians are separated substantially. Though NN generally captures the informative dimensionalities, NNG always finds better dimensionalities and shows substantial classification improvements.

3.3.3 Gene selection with microarray data

Two different datasets are collected from previous gene expression research for breast cancer prognosis studies [4, 20], and samples in each dataset are classified using selected genes according to the JS divergence and a univariate t-score.

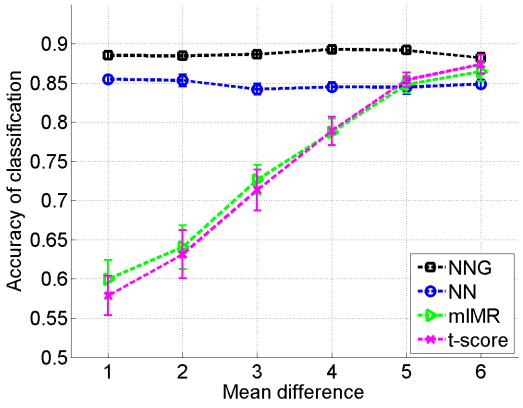


Figure 4: Feature selection in Gaussian example

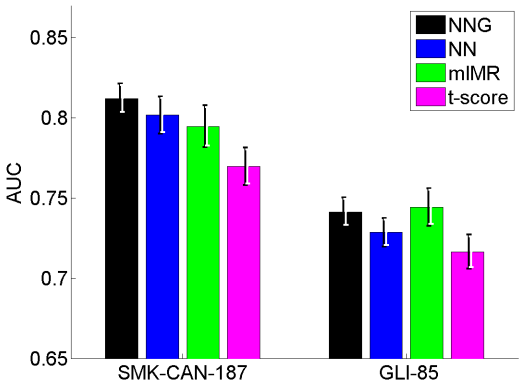


Figure 5: Gene expression classification using selected features for datasets SMK-CAN-183 [4] and GLI-85 [20]

Gaussianity is used again in $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ for NNG. As in the previous section, the metric can be obtained by assuming the Gaussianity even though the JS divergence is non-integrable.

We use the forward selection strategy for feature selection and compare the average AUC of classification for the proposed NNG method, for the plain NN method, for mIMR which is a parametric method of estimating the JS divergence [1], and for a simple univariate t-score. The results are reported in Fig. 5, showing that the proposed NNG method compares favorably with other methods.

4 Conclusions

In this work, we showed how nonparametric nearest-neighbor estimation can be significantly affected by the choice of a metric and that the metric dependency is related to the finite sample effect. Typically, the finite sampling causes a bias of the estimator, and this bias can be alleviated using an appropriate metric. In a small sample situation, the proposed perturbative derivation is at its weakest for adjusting the bias. Nevertheless, it is shown that the estimator effectively enhances its reliability of estimation by minimizing only the leading-order of the perturbed deviation.

We should note that the chosen metric in Eq. (28) reduces the bias effectively in empirical situations, but this metric is not a unique solution for minimizing the derived leading-order bias term. The estimation results may be different for other solutions due to the higher-order deviation, but the proposed metric still works reliably in most situations. In our future work, we will consider the higher-order bias to see if we improve the metric for a small data situation.

Finally, the dependency of parametric model needs to be investigated more systematically. In this work, we tried to show that a Gaussian model can be used as a rough model capturing the global configuration of data, but we can use more complex models that can capture the specific components of data. The simple extensions include the Gaussian mixture extension using local estimates of Σ and μ.

Acknowledgements

We acknowledge the support from KAKENHI 23120004 for YKN, from KAKENHI 25700022 and AOARD for MS, from JSPS KAKENHI 253189 and DC2 program for SL, and from BMRR and the SNU-MAE BK21+ program for FCP.

References

[1] G. Bontempi and P. E. Meyer. Causal filter selection in microarray data. In *Proceedings of the 27th International Conference on Machine Learning*, pages 95–102, 2010.

[2] G. Brown. A new perspective for information theoretic feature selection. *Journal of Machine Learning Research - Proceedings Track*, 5:49–56, 2009.

[3] K. Das and Z. Nenadic. Approximate information discriminant analysis: A computationally simple heteroscedastic feature extraction technique. *Pattern Recognition*, 41(5):1548–1557, 2008.

[4] W. A. Freije, F. E. Castro-Vargas, Z. Fang, S. Horvath, T. Cloughesy, L. M. Liau, P. S. Mischel, and S. F. Nelson. Gene expression profiling of gliomas strongly predicts survival. *Cancer Research*, 15;64(18):6503–6510, 2004.

[5] D. Garcia-Garcia, U. von Luxburg, and R. Santos-Rodriguez. Risk-based generalizations of f-divergences. In *Proceedings of 28th International Conference on Machine Learning*, pages 417–424, 2011.

[6] M. N. Goria, N. N. Leonenko, V. V. Mergel, and P. Inverardi. A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *Journal of Nonparametric Statistics*, 17(3):277–297, 2005.

[7] T. Kanamori, T. Suzuki, and M. Sugiyama. f-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. *IEEE Transactions on Information Theory*, 58(2):708–720, 2012.

[8] Y. Kawahara and M. Sugiyama. Sequential change-point detection based on direct density-ratio estimation. *Statistical Analysis and Data Mining*, 5(2):114–127, 2012.

[9] N. Leonenko and L. Pronzato. Correction: A class of Rényi information estimators for multidimensional densities. *Annals of Statistics*, 38:3837–3838, 2010.

[10] N. Leonenko, L. Pronzato, and V. Savani. A class of Rényi information estimators for multidimensional densities. *Annals of Statistics*, 36:2153–2182, 2008.

[11] S. Liu, M. Yamada, N. Collier, and M. Sugiyama. Change-point detection in time-series data by relative density-ratio estimation. Technical Report 1203.0453, arXiv, 2012.

[12] D. O. Loftsgaarden and C. P. Quesenberry. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36(3):1049–1051, June 1965.

[13] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Advances in Neural Information Processing Systems*, 2007.

[14] Y. Noh, B. Zhang, and D. D. Lee. Generative local metric learning for nearest neighbor classification. In *Advances in Neural Information Processing Systems 23*, pages 1822–1830. 2010.

[15] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, August 2005.

[16] F. Perez-Cruz. Kullback-Leibler divergence estimation of continuous distributions. In Frank R. Kschischang and En-Hui Yang, editors, *Proceedings of IEEE International Symposium on Information Theory*, pages 1666–1670. IEEE, 2008.

[17] B. Poczos and J. Schneider. On the estimation of alpha-divergences. In *International Conference on AI and Statistics*, pages 609–617, 2011.

[18] B. Poczos and J. Schneider. Nonparametric estimation of conditional information and divergences. In *International Conference on AI and Statistics*, JMLR Workshop and Conference Proceedings, 2012.

[19] B. Ranneby. The maximum spacing method. an estimation method related to the maximum likelihood method. *Scandinavian Journal of Statistics*, 1(2):93–112, 1984.

[20] A. Spira, et al. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nature Medicine*, 13(3):361–6, 2007.

[21] M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.

[22] Q. Wang, S. R. Kulkarni, and S. Verdu. A nearest-neighbor approach to estimating divergence between continuous random vectors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 55(5):2392–2405, 2006.

[23] L. Yu and H. Liu. Efficient Feature Selection via Analysis of Relevance and Redundancy. *Journal of Machine Learning Research*, 5:1205–1224, 2004.