Linear Kernel Tests via Empirical Likelihood for High Dimensional Data

Abstract

We propose a framework for analyzing and comparing distributions without imposing any parametric assumptions via empirical likelihood methods. Our framework is used to study two statistical test problems: the two sample test problem and the goodness of fit test problem. For the two sample test, we need to determine whether two groups of samples are from different distributions; for the goodness of fit test, we examine how likely a set of samples are generated from a known target distribution. Specifically, we propose two empirical likelihood ratio (ELR) statistics respectively for the two sample test and the goodness of fit test, both of which are of linear time complexity and show higher power (i.e., the probability of correctly rejecting the null hypothesis) than the existing linear statistics on high dimensional data. We prove the nonparametric Wilks' theorems for the ELR statistics, which illustrate that the limiting distributions of the proposed ERL statistics are both chi-square distributions. The limiting distributions can avoid bootstrap or simulation to get the threshold for rejecting the null hypothesis, which makes the ELR statistics are more efficient than the recently proposed linear statistic, finite set Stein discrepancy (FSSD). Additionally, we have experimentally found and theoretically analyzed that FSSD have poor performance or even fail to test for high dimensional data. Finally, we conduct a series of experiments to evaluate the performance of our ELR statistics as compared to state-of-the-art ones.

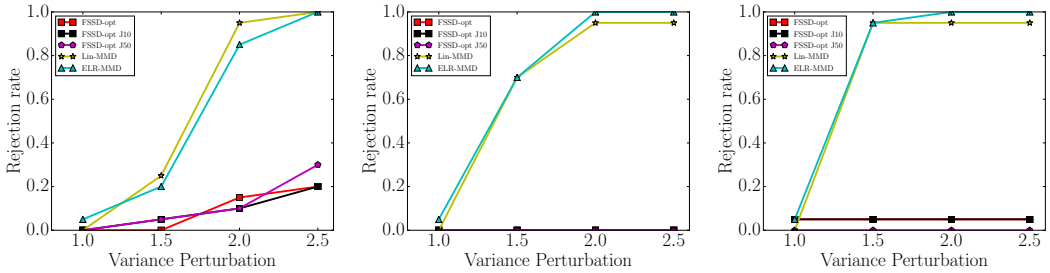
Introduction

Comparing samples from two probability distributions or evaluating the goodness of fit of models over observed samples without imposing any parametric assumptions on their distributions are fundamental tasks in machine learning and statistics, and have wide spectra of applications in various areas (Lloyd and Ghahramani, 2015; Li et al., 2017; Bińkowski et al., 2018). The goal of the two sample test problem is to determine whether two distributions p and q are different on the basis of samples $\mathcal{D}_{\boldsymbol{x}} = \{\boldsymbol{x}_i\}_{i=1}^n \subset \mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{D}_{\boldsymbol{y}} = \{\boldsymbol{y}_j\}_{j=1}^m \subset \mathcal{Y} \subseteq \mathbb{R}^d$ independently drawn from p and q, respectively. The aim of the goodness of fit test problem is to determine how well a given model density p fits a set of given samples $\mathcal{D}_{\boldsymbol{x}} = \{\boldsymbol{x}_i\}_{i=1}^n \subset \mathcal{X} \subseteq \mathbb{R}^d$ from an unknown distribution q. These two problems can both be solved via a hypothesis test, where the null hypothesis $H_0 : p = q$ is tested against the alternative hypothesis $H_1 : p \neq q$. The main difference is that the distribution of p is known for the goodness of fit test.

The two sample test and the goodness of fit test are generally difficult and challenging in practice, since the underneath distributions are generally unknown apriori. Kernel methods provide a way to implicitly transform the data into a new feature space, and the corresponding reproducing kernel Hilbert spaces (RKHSs) have strong representative power (Cucker and Smale, 2002). In this paper, we adopt the unit balls in universal RKHSs as the function classes (Muandet et al., 2017) to study these two test problems, since these classes are rich enough to represent all bounded continuous functions defined on a metric space (Fukumizu, Bach, and Jordan, 2004; Sriperumbudur et al., 2010; Steinwart, 2001; Micchelli, Xu, and Zhang, 2006).

For the two sample test problem, the popular discrepancy, maximum mean discrepancy (MMD), was designed to measure two distributions by embedding them in an RKHS (Gretton et al., 2012a). The MMD has been attracting much attention in recent two sample test research due to its solid theoretical foundation (Sriperumbudur et al., 2009; Gretton et al., 2007, 2012a,b; Song et al., 2012; Zaremba, Gretton, and Blaschko, 2013; Zhao and Meng, 2015; Ramdas et al., 2014). The minimum variance unbiased estimator $\mathrm{MMD}_{\mathrm{Unb}}$ of MMD was first proposed in (Gretton et al., 2007) on the basis of n samples observed from each of p and q, which is a U-statistic. However, the asymptotic distribution takes the form of an infinite weighted sum of independent χ^2 variables, so the estimation of the null distribution requires bootstrap or moment matching, which costs at least $O(n^2)$. Later, an $O(n)$ unbiased estimate $\mathrm{MMD}_{\mathrm{Lin}}$ of MMD was proposed (Gretton et al., 2012a), by using a subsampling of the terms in the sum. $\mathrm{MMD}_{\mathrm{Lin}}$ has higher variance than $\mathrm{MMD}_{\mathrm{Unb}}$, but it is computationally much more appealing.

For the goodness of fit test, the traditional methods need to calculate or compare the likelihoods or cumulative distribution functions (CDF) of the models, but for the large graphical models or deep generative models (Koller and Friedman, 2009; Salakhutdinov, 2015), it is often computationally intractable (Chandrasekaran, Srebro, and Harsha, 2008). Recently, Stein's method (Stein and others, 1972; Oates, Girolami, and Chopin, 2017) has been introduced into the ker-



(a) $d = 10$ (b) $d = 50$ (c) $d = 100$

Figure 1: Rejection rates of FSSD, MMD_{Lin} and ELR-MMD on two different normal distributions $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|0, \mathbf{I}_d)$ and $q(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|0, v\mathbf{I}_d)$ with the variance changed in the set $v \in \{1, 1.5, 2, 2.5\}$ for $d = 10, 50, 100$.

nel domain (Chwialkowski, Strathmann, and Gretton, 2016; Liu, Lee, and Jordan, 2016), by combining Stein's identity with the RKHS theory, which is a likelihood-free method and depends on p only through the log derivatives. The asymptotic distribution of the statistics under the null hypothesis is an infinite weighted sum of independent χ^2 variables. The bootstrap is adopted for calculating the approximate rejection threshold, whose time complexity is $O(n^2)$. A linear time statistic, KSD_{Lin}, was proposed by using half sampling (Liu, Lee, and Jordan, 2016), which has a zero-mean Gaussian limit under the null hypothesis. To improve the performance of the existing linear time statistics, Jitkrittum et al. (2017) proposed a novel statistic, the finite set Stein discrepancy (FSSD), by introducing a witness function on a finite set, which can conduct testing in linear time and show excellent performance on low dimensional data.

In this paper, we introduce the method of empirical likelihood into the domain of linear kernel tests for the first time, and propose two novel empirical likelihood ratio (ELR) statistics for the two sample test and the goodness of fit test, respectively. The empirical likelihood method (Owen, 1990, 2001) owns its broad usage and fast research development to a number of important advantages. Generally speaking, it combines the reliability of nonparametric methods with the effectiveness of the likelihood approach. The regions behave better than confidence regions based on asymptotic normality when the sample size is not large enough. Taking into consideration the asymptotic normality of the linear unbiased estimate MMD_{Lin} (Gretton et al., 2012a), we first propose an empirical likelihood ratio (ERL) statistic based on the formulation of MMD_{Lin}, named ERL-MMD, for the two sample test problem. We optimize an empirical distribution on the set of the one-dimensional pairwise discrepancies with the constraint that the empirical mean of all discrepancies is 0. We establish the nonparametric Wilks' theorem for the statistic ERL-MMD, which shows that the proposed ERL-MMD has a limiting chi-square distribution. For the goodness of fit test, we propose an ERL statistic based on the linear unbiased estimate KSD_{Lin}, called ERL-KSD, by enforcing an empirical distribution on the the pairwise discrepancies, and derive the nonparametric Wilks' theorem to show the limiting distribution of ERL-KSD. The proposed ELR-MMD and ELR-KSD statistics show better performance than MMD_{Lin} and KSD_{Lin}, and remarkably higher discriminability (power) when comparing two distributions with subtle difference. There are two possible reasons for the impressive performance of the ERL statistics. First, enforcing a probability on each pairwise discrepancy can help discriminate the subtle difference between two distributions. Second, the rejection regions of the proposed statistics are obtained by contouring a log likelihood ratio that may be the most powerful test for a fixed significance level α by Neyman-Pearson lemma (Neyman and Pearson, 1933).

Another contribution of this paper is that we have experimentally found that the recently proposed FSSD have poor performance or even fail to test for high dimensional data. In Figure 1[1], we investigate the power of FSSD as compared to MMD_{Lin} and ELR-MMD, on two different normal distributions $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|0, \mathbf{I}_d)$ and $q(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|0, v\mathbf{I}_d)$ with the variance changed in the set $v \in \{1, 1.5, 2, 2.5\}$ for $d = 10, 50, 100$. We can find that, MMD_{Lin} and ELR-MMD can work well for $d = 10, 50, 100$, but FSSD shows poor rejection rates for both $d = 10$, and fails to reject the null hypothesis $H_0 : p = q$ for $d = 50, 100$ even when the variance v is very large. We also increase an important parameter of FSSD, the number of test locations J, to further verify the performance of FSSD, but the results are similar that have also been shown in 1. We further provide a deeper understanding of FSSD from the perspective of empirical likelihood and analyze the possible reasons why FSSD shows poor performance or even fails on high dimensional data. Since FSSD has shown good performance on low dimensional data (Jitkrittum et al., 2017), the proposed ERL statistics can be considered as complements to FSSD for high dimensional data.

Empirical Likelihood Ratio for Two Sample Test

In this section, we will propose an empirical likelihood ratio (ERL) statistic for the two sample test problem and derive its limiting distribution by Wilks' Theorem.

Assume that the data domain is a connected open set $\mathcal{X} \in \mathbb{R}^d$. Let $\mathcal{H}_\kappa$ be a reproducing kernel Hilbert space

[1]Comprehensive results are given in the section of empirical study.

(RKHS) defined on $\mathcal{X}$ with reproducing kernel κ, and p a Borel probability measure on $\mathcal{X}$. In this paper, we consider the function class $\mathcal{F}$ as a unit ball in a universal RKHS $\mathcal{H}_\kappa$, since this class is rich enough to show the equivalence between the zero expectation of the statistics and the equality of two distributions (Fukumizu, Bach, and Jordan, 2004; Sriperumbudur et al., 2010; Steinwart, 2001; Micchelli, Xu, and Zhang, 2006). Universality requires that κ is continuous and $\mathcal{H}_\kappa$ is dense in $C(\mathcal{X})$ with respect to the L_∞ norm. It has been proved that the Gaussian and Laplace RKHSs are universal (Steinwart, 2001).

The mean embedding of p in $\mathcal{F}$, written as $\mu_\kappa(p) \in \mathcal{F}$, is defined such that $\mathbf{E}_{\boldsymbol{x}\sim p} f(\boldsymbol{x}) = \langle f, \mu_\kappa(p)\rangle$ for all $f \in \mathcal{F}$ and exists for all Borel probability measures when κ is bounded and continuous. The MMD between a Borel probability measure p and a second Borel probability measure q is the squared RKHS distance between their respective mean embeddings,

$$\begin{aligned}\eta_\kappa(p,q) &= \|\mu_\kappa(p) - \mu_\kappa(q)\|^2_{\mathcal{F}} \\ &= \mathbf{E}_{\boldsymbol{x}\boldsymbol{x}'}\kappa(\boldsymbol{x},\boldsymbol{x}') + \mathbf{E}_{\boldsymbol{y}\boldsymbol{y}'}\kappa(\boldsymbol{y},\boldsymbol{y}') - 2\mathbf{E}_{\boldsymbol{x}\boldsymbol{y}}\kappa(\boldsymbol{x},\boldsymbol{y}),\end{aligned}$$

where $\boldsymbol{x}'$ denotes an independent copy of $\boldsymbol{x}$. Denoting $\boldsymbol{z} = (\boldsymbol{x},\boldsymbol{y})$, we write $\eta_\kappa(p,q) = \mathbf{E}_{\boldsymbol{z}\boldsymbol{z}'} h_\kappa(\boldsymbol{z},\boldsymbol{z}')$, with $h(\boldsymbol{z},\boldsymbol{z}') = \kappa(\boldsymbol{x},\boldsymbol{x}') + \kappa(\boldsymbol{y},\boldsymbol{y}') - \kappa(\boldsymbol{x},\boldsymbol{y}') - \kappa(\boldsymbol{x}',\boldsymbol{y})$. It has been proved that for the unit ball in a universal RKHS $\mathcal{F}$, $\eta_\kappa(p,q) = 0$ if and only if $p = q$ (Gretton et al., 2012a).

For the given two samples $\mathcal{D}_{\boldsymbol{x}} = \{\boldsymbol{x}_i\}_{i=1}^n \subset \mathcal{X} \subseteq \mathbb{R}^d$, where $\boldsymbol{x}_i \sim p$ i.i.d., and $\mathcal{D}_{\boldsymbol{y}} = \{\boldsymbol{y}_j\}_{j=1}^m \subset \mathcal{Y} \subseteq \mathbb{R}^d$, where $\boldsymbol{y}_i \sim q$ i.i.d., if we assume $m = n$, the minimum variance unbiased estimate $\mathrm{MMD}_{\mathrm{Unb}}$ of $\eta_\kappa(p,q)$ can be simply represented as

$$\mathrm{MMD}^2_{\mathrm{Unb}}[\mathcal{F},\mathcal{D}_{\boldsymbol{x}},\mathcal{D}_{\boldsymbol{y}}] = \frac{1}{n(n-1)}\sum_{i\neq j}^{n} h(\boldsymbol{z}_i,\boldsymbol{z}_j).$$

$\mathrm{MMD}_{\mathrm{Unb}}$ requires $O(n^2)$ time to compute h on all interacting pairs. The asymptotic distribution of $\mathrm{MMD}_{\mathrm{Unb}}$ takes the form of an infinite weighted sum of independent χ^2 variables, so the bootstrap or moment matching for estimation of the null distribution costs at least $O(n^2)$. A linear time unbiased estimate $\mathrm{MMD}_{\mathrm{Lin}}$ of $\eta_\kappa(p,q)$ was proposed in (Gretton et al., 2012a),

$$\mathrm{MMD}^2_{\mathrm{Lin}}[\mathcal{F},\mathcal{D}_{\boldsymbol{x}},\mathcal{D}_{\boldsymbol{y}}] = \frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor} h(\boldsymbol{z}_{2i-1},\boldsymbol{z}_{2i}).$$

In the following, we will present the empirical likelihood ratio statistic based on $\mathrm{MMD}_{\mathrm{Lin}}$. We write $h_i = h(\boldsymbol{z}_{2i-1},\boldsymbol{z}_{2i})$ and $N = \lfloor n/2\rfloor$. When calculating h_i, $i = 1,\ldots,N$, different independent samples are used for different i. We consider $h_1, h_2, \ldots, h_N$ as i.i.d observations from a univariate distribution ρ. We can define an empirical likelihood function as

$$L(\rho) = \prod_{i=1}^{N} d\rho(h_i) = \prod_{i=1}^{N} p_i,$$

where $p_i = d\rho(h_i) = \Pr(H = h_i)$. $L(\rho)$ is maximized by the empirical distribution function $\rho_N(h) = N^{-1}\sum_{i=1}^N I(h_i < h)$. The empirical likelihood ratio is then defined as

$$R(\rho) = L(\rho)/L(\rho_N),$$

and it is easy to show that this can be written as

$$R(\rho) = \prod_{i=1}^{N} Np_i.$$

In the two sample test problem, under the null hypothesis $H_0 : p = q$, the mean μ of ρ should be 0. To obtain the confidence regions, we define the empirical likelihood ratio function

$$R(\mu) = \sup_{\{p_i\geq 0\}_{i=0}^N}\left\{\prod_{i=1}^{N} Np_i \,\middle|\, \sum_{i=1}^{N} p_i = 1, \sum_{i=1}^{N} p_i h_i = \mu\right\}. \tag{1}$$

As noted by (Owen, 1988, 1990, 2001), a unique value for the right-hand side of (1) exists, provided that μ is inside the convex hull of the points $h_1,\ldots,h_N$. An explicit expression for $R(0)$ can be derived by a Lagrange multiplier argument: the maximum of $\prod_{i=1}^N Np_i$ subject to the constraints $p_i \geq 0$, $\sum_{i=1}^N p_i = 1$ and $\sum_{i=1}^N p_i h_i = 0$ is attained when

$$p_i = p_i(0) = \frac{1}{N}\frac{1}{1+\lambda h_i},$$

where λ is the solution to

$$\sum_{i=1}^{N}\frac{h_i}{1+\lambda h_i} = 0.$$

The empirical likelihood ratio test statistic is proposed as $W(0) = -2\log R(0)$, that is,

$$W(0) = 2\sum_{i=1}^{N}\log\{1+\lambda h_i\}.$$

We derive the following theorem, which shows that the proposed statistic $W(0)$ have a limiting chi-square distribution.

Theorem 1 (Wilks' Theorem)**.** *Under $H_0 : p = q$, if $\mathbf{E}_{\boldsymbol{x},\boldsymbol{x}'}[\kappa^2(\boldsymbol{x},\boldsymbol{x}')] \leq \infty$, the empirical likelihood ratio test statistic*

$$W(0) \xrightarrow{d} \chi^2_{(1)}.$$

Proof. Every h_i is defined on an independent pairwise discrepancy $(\boldsymbol{z}_{2i-1},\boldsymbol{z}_{2i})$ for $i = 1,2,\ldots,N$, where $\boldsymbol{x}_{2i-1}$, $\boldsymbol{y}_{2i-1}$, $\boldsymbol{x}_{2i}$ and $\boldsymbol{y}_{2i}$ are used only for h_i. Therefore, $\{h_i, i = 1,2,\ldots,N\}$ is a sequence of i.i.d. random variables. If $p = q$, we have $\mathbf{E}[h_i] = 0$. Let $\bar{h} := \frac{1}{N}\sum_{i=1}^m h_i$ and $S := \frac{1}{N}\sum_{i=1}^m h_i^2$. From Lemma 11.1 of (Owen, 2001), we have $\lambda = \bar{h}/S + o_p(N^{-1/2})$. By Taylor's expansion, we

have

$$\begin{aligned}
W(0) &= 2\sum_{i=1}^{N}\log(1+\lambda h_i) \\
&= 2\sum_{i=1}^{N}\lambda h_i - \sum_{i=1}^{N}(\lambda h_i)^2 + \sum_{i=1}^{N}\eta_i \\
&= 2N\lambda\bar{h} - N\lambda^2 S + \sum_{i=1}^{N}\eta_i \\
&= 2N\bar{h}^2/S - N\bar{h}^2/S + 2N\bar{h}\cdot o_p(N^{-1/2}) \\
&\qquad + N\cdot o_p(N^{-1})/S + \sum_{i=1}^{N}\eta_i \\
&= N\bar{h}^2/S + 2N\bar{h}\cdot o_p(N^{-1/2}) \\
&\qquad + N\cdot o_p(N^{-1})/S + \sum_{i=1}^{N}\eta_i,
\end{aligned}$$

where for some finite constant $C > 0$,

$$\Pr\left(|\eta_i| \le C|\lambda h_i|^3\right) \to 1,\ 1 \le i \le N.$$

Standard Central Limit Theorem and Continuous Mapping Theorem imply that

$$N\bar{h}^2/S = (\sqrt{N}\bar{h}/\sqrt{S})^2 \xrightarrow{d} \chi^2_{(1)}.$$

Now, since $N\bar{h}\cdot o_p(N^{-1/2}) = o_p(1)$, $N\cdot o_p(N^{-1})/S = o_p(1)$ and

$$\begin{aligned}
\left|\sum_{i=1}^{N}\eta_i\right| &\le C|\lambda|^3\sum_{i=1}^{N}|h_i|^3 \\
&= O_p(N^{-3/2})\cdot o_p(N^{3/2}) = o_p(1),
\end{aligned}$$

the proof is completed. □

Based on Theorem 1, we can conduct two sample test in this way: we will reject the null hypothesis H_0, when $W(0) \ge \chi^2_\alpha$, where χ^2_α is defined such that

$$\Pr(\chi^2_{(1)} \ge \chi^2_\alpha) = \alpha.$$

Since the limiting distribution is $\chi^2_{(1)}$, we do not require bootstrap or simulation to get the threshold for rejection. The main computational burden for $W(0)$ is the calculation of $h_i, i = 1, \ldots, N$. Therefore, the time complexity of the empirical likelihood ratio statistic $W(0)$ is linear in the number of examples. The empirical likelihood method can also be applied to B-tests (Zaremba, Gretton, and Blaschko, 2013), since the statistics η_i, $i = 1, \ldots, \frac{n}{B}$, for different blocks are independent with each other, where B is the block size. However, in each block of B-tests, the pairwise discrepancies are not independent with each other.

Empirical Likelihood Ratio for Goodness of Fit Test

In this section, we will propose an empirical likelihood ratio (ERL) statistic for the goodness of fit test problem and derive its limiting distribution by Wilks' Theorem.

We first introduce the Stein operator (Stein and others, 1972; Oates, Girolami, and Chopin, 2017), which depends on the distribution p only through log derivative and avoids the calculation of the normalization constant. A Stein operator T_p takes a multivariate function $\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \ldots, f_d(\boldsymbol{x}))^{\mathrm{T}} \in \mathbb{R}^d$ as input and outputs a function $(T_p f)(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}$. The function $T_p f$ has the key property that for all fs in an appropriate function class,

$$\mathbf{E}_{\boldsymbol{x}\sim q}[(T_p f)(\boldsymbol{x})] = 0$$

if and only if $q = p$. Thus, the expectation can be used to test the goodness of fit: how well a model density $p(\boldsymbol{x})$ fits a set of given samples $\mathcal{D}_{\boldsymbol{x}} = \{\boldsymbol{x}_i\}_{i=1}^n \subset \mathcal{X} \subseteq \mathbb{R}^d$ from an unknown distribution q.

Here we consider the function class $\mathcal{F}^d$, where $\mathcal{F}$ is a unit-norm ball in universal RKHS (Gretton et al., 2012a). More precisely, assume that $f_i \in \mathcal{F}$ for all $i = 1, \ldots d$ so that $f \in \mathcal{F}\times\cdots\times\mathcal{F} := \mathcal{F}^d$ where $\mathcal{F}^d$ is equipped with the standard inner product $\langle f, g\rangle_{\mathcal{F}^d} := \sum_{i=1}^d \langle f_i, g_i\rangle_{\mathcal{F}}$.

According to the reproducing property of $\mathcal{F}$, $f_i(\boldsymbol{x}) = \langle f_i, \kappa(\boldsymbol{x},\cdot)\rangle_{\mathcal{F}}$, and that $\frac{\partial\kappa(\boldsymbol{x},\cdot)}{\partial\boldsymbol{x}_i} \in \mathcal{F}$, we can define $\omega_p(\boldsymbol{x},\cdot) = \frac{\partial\log p(\boldsymbol{x})}{\partial\boldsymbol{x}}\kappa(\boldsymbol{x},\cdot) + \frac{k(\boldsymbol{x},\cdot)}{\partial\boldsymbol{x}}$, which is in $\mathcal{F}^d$. Therefore, the kernelized Stein operator can be written as

$$\begin{aligned}
(T_p f)(\boldsymbol{x}) &= \sum_{i=1}^{d}\left(\frac{\partial\log p(\boldsymbol{x})}{\partial\boldsymbol{x}_i}f_i(\boldsymbol{x}) + \frac{\partial f_i(\boldsymbol{x})}{\partial\boldsymbol{x}_i}\right) \\
&= \langle f, \omega_p(\boldsymbol{x},\cdot)\rangle_{\mathcal{F}^d}.
\end{aligned}$$

Under the condition $\lim_{\|\boldsymbol{x}\|\to\infty} p(\boldsymbol{x})f_i(\boldsymbol{x}) = 0$ for all $i = 1, \ldots, d$, it can be shown using integration by parts that $\mathbf{E}_{\boldsymbol{x}\sim p}(T_p f)(\boldsymbol{x}) = 0$ for any $f \in \mathcal{F}^d$. Based on the Stein operator, the kernelized Stein discrepancy is defined as

$$S_p(q) = \sup_{\|f\|_{\mathcal{F}^d}\le 1}\langle f, \mathbf{E}_{\boldsymbol{x}\sim p}\omega_p(\boldsymbol{x},\cdot)\rangle = \|g(\cdot)\|_{\mathcal{F}^d}, \qquad (2)$$

where $g(\cdot) = \mathbf{E}_{\boldsymbol{x}\sim q}\omega_p(\boldsymbol{x},\cdot)$ is called the Stein witness function. The Stein witness function plays a crucial role in the FSSD test.

When $\mathbf{E}_{\boldsymbol{x}\sim p}\|\nabla_{\boldsymbol{x}}\log p(\boldsymbol{x}) - \nabla_{\boldsymbol{x}}\log q(\boldsymbol{x})\| < \infty$, it can be shown that $S_p(q) = 0$ if and only if $p = q$. The statistic of the KSD can be represented as

$$S_p^2(q) = \mathbf{E}_{\boldsymbol{x}\sim q}\mathbf{E}_{\boldsymbol{x}'\sim q}h_p(\boldsymbol{x},\boldsymbol{x}'),$$

where $h_p(\boldsymbol{x},\boldsymbol{y}) = s_p^{\mathrm{T}}(\boldsymbol{x})s_p(\boldsymbol{y})\kappa(\boldsymbol{x},\boldsymbol{y}) + s_p^{\mathrm{T}}\nabla_{\boldsymbol{x}}\kappa(\boldsymbol{x},\boldsymbol{y}) + s_p^{\mathrm{T}}\nabla_{\boldsymbol{y}}\kappa(\boldsymbol{x},\boldsymbol{y}) + \sum_{i=1}^d\frac{\partial^2\kappa(\boldsymbol{x},\boldsymbol{y})}{\partial\boldsymbol{x}_i\partial\boldsymbol{y}_i}$, and $s_p(\boldsymbol{x}) = \nabla_{\boldsymbol{x}}\log p$ which is called the score function. An unbiased empirical estimator of $S_p^2(q)$, written as

$$\hat{S}^2 = \frac{2}{n(n-1)}\sum_{i<j}h_p(\boldsymbol{x}_i,\boldsymbol{x}_j),$$

is a degenerate U-statistic under H_0. For the goodness of fit test, the rejection threshold is computed by a bootstrap procedure. It can be seen that the computational cost of $\hat{S}^2$ is $O(n^2)$. To reduce this cost, a linear time estimator was proposed (Liu, Lee, and Jordan, 2016),

$$\hat{S}_m^2 = \frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}h_p(\boldsymbol{x}_{2i-1},\boldsymbol{x}_{2i}).$$

We write $h_{p,i} = h_p(\boldsymbol{x}_{2i-1}, \boldsymbol{x}_{2i})$ and $N = \lfloor n/2 \rfloor$. When calculating $h_{p,i}$, $i = 1, \ldots, N$, different independent samples are used for different i. For the null hypothesis H_0, we consider $h_{p,1}, h_{p,2}, \ldots, h_{p,N}$ as i.i.d observations from a univariate distribution. By using $\hat{S}_m^2$. we define the empirical likelihood ratio function

$$R(\mu) = \sup_{\{p_i \geq 0\}_{i=0}^N} \left\{ \prod_{i=1}^N N p_i \middle| \sum_{i=1}^N p_i = 1, \sum_{i=1}^N p_i h_{p,i} = \mu \right\}.$$

An explicit expression for $R(0)$ can be derived by a Lagrange multiplier argument,

$$p_i = p_i(0) = \frac{1}{N} \frac{1}{1 + \lambda h_{p,i}},$$

where λ is the solution to

$$\sum_{i=1}^N \frac{h_{p,i}}{1 + \lambda h_{p,i}} = 0.$$

The empirical likelihood ratio test statistic is thus

$$W_p(0) = 2 \sum_{i=1}^N \log\{1 + \lambda h_{p,i}\}.$$

We derive the Wilks' theorem (Theorem 2) for $W_p(0)$, which shows a limiting chi-square distribution of $W_p(0)$. The proof of Theorem 2 is similar to that of Theorem 1.

Theorem 2 (Wilks' Theorem)**.** *Under $H_0 : p = q$, if $\mathbf{E}_{\boldsymbol{x},\boldsymbol{x}'}[\kappa^2(\boldsymbol{x}, \boldsymbol{x}')] \leq \infty$, the empirical likelihood ratio test statistic*

$$W_p(0) \xrightarrow{d} \chi^2_{(1)}.$$

Based on Theorem 2, we will reject the null hypothesis H_0, when $W_p(0) \geq \chi^2_\alpha$ with χ^2_α satisfying $\Pr(\chi^2_{(1)} \geq \chi^2_\alpha) = \alpha$. The main computational burden for $W_p(0)$ is the calculation of $h_{p,i}$, $i = 1, \ldots, N$. Therefore, the time complexity of $W_p(0)$ is linear in the number of examples.

Comparisons with FSSD

In this section, we compare FSSD with existing linear statistics and our ELR statistics and analyze the possible reasons why FSSD shows poor performance or even fails on high dimensional data.

We first briefly introduce FSSD (Jitkrittum et al., 2017). Let $V = \{v_1, \ldots, v_J\} \subset \mathbb{R}^d$ be random vectors drawn i.i.d. from a distribution η which has a density. The statistic of FSSD can be defined as

$$\text{FSSD}_p^2(q) = \frac{1}{dJ} \sum_{i=1}^d \sum_{j=1}^J g_i^2(v_j),$$

where $g(\cdot)$ is the Stein witness function given in (2). It has been proved (Jitkrittum et al., 2017) that if the following conditions are satisfied 1) $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is C_0-universal and real analytic i.e., for all $\boldsymbol{x}' \in \mathcal{X}$, $f(\boldsymbol{x}) = \kappa(\boldsymbol{x}, \boldsymbol{x}')$ is a real analytic function on $\mathcal{X}$; 2) $\mathbf{E}_{\boldsymbol{x}\sim q}\mathbf{E}_{\boldsymbol{x}'\sim p} h_p(\boldsymbol{x}, \boldsymbol{x}') < 0$; 3) $\mathbf{E}_{\boldsymbol{x}\sim q}\|\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} \log q(\boldsymbol{x})\|^2 < \infty$; and 4) $\lim_{\|\boldsymbol{x}\|\to\infty} p(\boldsymbol{x}) g(\boldsymbol{x}) = 0$; for any $J \geq 1$, η-almost surely $\text{FSSD}_p^2(q) = 0$ if and only if $p = q$. Let $\Omega(\boldsymbol{x}) \in \mathbb{R}^{d\times J}$ such that

$$[\Omega(\boldsymbol{x})]_{i,j} = \omega_{p,i}(\boldsymbol{x}, \boldsymbol{v}_j)/\sqrt{dJ},$$

$\tau(\boldsymbol{x}) = \text{vec}(\Omega(\boldsymbol{x})) \in \mathbb{R}^{dJ}$ where $\text{vec}(\cdot)$ denotes the vectorization, and $\Delta(\boldsymbol{x}, \boldsymbol{y}) = \tau(\boldsymbol{x})^{\mathrm{T}}\tau(\boldsymbol{y}) = \text{tr}(\Omega(\boldsymbol{x})^{\mathrm{T}}\Omega(\boldsymbol{y}))$. The unbiased estimator of $\text{FSSD}_p^2(q)$ is

$$\widehat{\text{FSSD}}^2 = \frac{2}{n(n-1)} \sum_{i<j} \Delta(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

In the following, we explain the reason why FSSD is different from MMD_{Lin}, KSD_{Lin}, ERL-MMD and ERL-KSD and further the reason why FSSD shows poor performance on high dimensional data.

For MMD_{Lin}, KSD_{Lin}, ERL-MMD and ERL-KSD one data point only corresponds to a one-dimensional statistical value, but for $\widehat{\text{FSSD}}^2$, one data point corresponds to a $d \times J$ matrix $\Omega(\boldsymbol{x})$ or a dJ-dimensional vector $\tau(\boldsymbol{x})$. The underlying reason for the higher dimensional correspondence of FSSD is the introduction of the finite set. The finite set makes the kernel function $\kappa(\boldsymbol{x}, \cdot)$ no longer only appearing in the dot product form with another function $f \in \mathcal{F}$, which is different from the forms in MMD_{Lin}, KSD_{Lin}, ERL-MMD and ERL-KSD. The higher dimensional correspondence makes the empirical likelihood difficult to be applied in FSSD. The elements in $\tau(\boldsymbol{x})$ for FSSD are not independent, so if we enforce a probability distribution on the set of $\tau(\boldsymbol{x}_i)$, $i = 1, \ldots, n$, the empirical likelihood ratio does not have a limiting χ^2_{dJ} distribution.

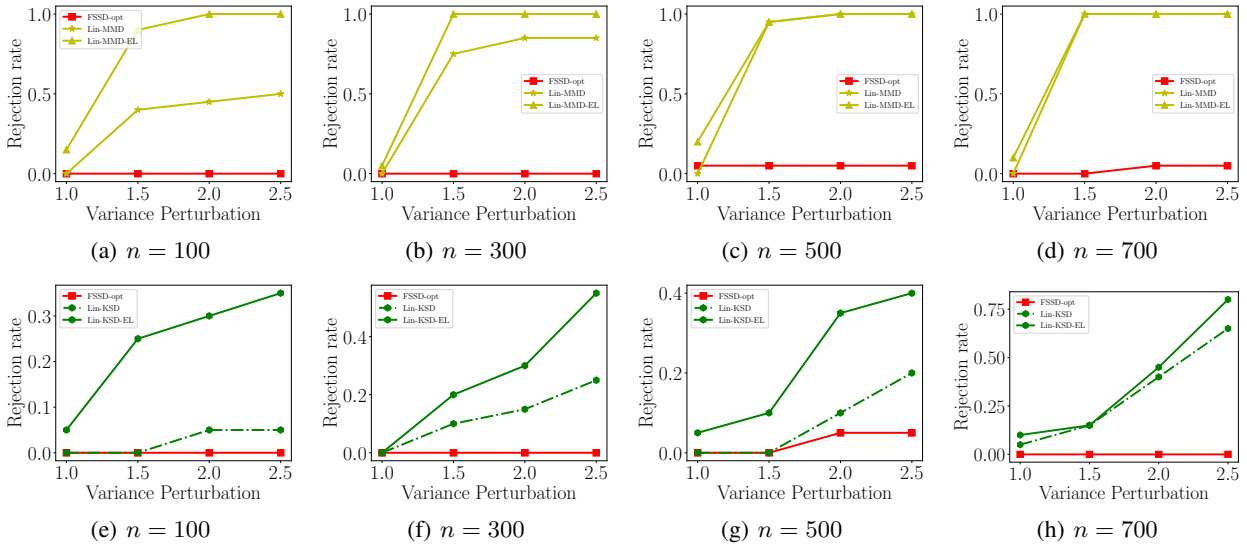
According to Proposition 2 in (Jitkrittum et al., 2017), under the alternative hypothesis $H_1 : p \neq q$, if $\sigma_{H_1} = 4\mu^{\mathrm{T}}\boldsymbol{\Sigma}_q\mu > 0$, then

$$n\widehat{\text{FSSD}}^2 \sim \sqrt{n}\mathcal{N}(0, \sigma_{H_1}) + n\text{FSSD}^2,$$

where $\mu = \mathbf{E}_{\boldsymbol{x}\sim q}[\tau(\boldsymbol{x})]$ and $\boldsymbol{\Sigma}_q = \text{cov}_{\boldsymbol{x}\sim q}[\tau(\boldsymbol{x})] \in \mathbb{R}^{dJ\times dJ}$. From the above equation, we know that $n\widehat{\text{FSSD}}^2$ is highly dependent on the dimension of the data. When the dimension increases, the dimension of $\boldsymbol{\Sigma}_q$ will increase, and then the variance σ_{H_1} becomes larger. When the variance becomes larger, the results will become unstable or even fail. For MMD_{Lin}, KSD_{Lin}, ERL-MMD and ERL-KSD, the kernel function $\kappa(\boldsymbol{x}, \cdot)$ only appears in the dot product form, and thus the statistics are all less dependent on the dimension of data. In addition, under the null hypothesis $H_0 : p = q$, the asymptotic distribution of $n\widehat{\text{FSSD}}^2$ is a finite weighted sum of independent χ^2 variables rather than the normal distributions of MMD_{Lin} and KSD_{Lin}.

Experiments

In this section, we conduct a series of experiments to demonstrate the performance of the proposed ERL statistics and understand the conditions under which the proposed statistics can perform well.



(a) $n = 100$ (b) $n = 300$ (c) $n = 500$ (d) $n = 700$

(e) $n = 100$ (f) $n = 300$ (g) $n = 500$ (h) $n = 700$

Figure 2: Rejection rates of Lin-MMD, ELR-MMD, Lin-KSD, ELR-KSD and FSSD on two different normal distributions $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|0, \mathbf{I}_d)$ and $q(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|0, v\mathbf{I}_d)$ with the variance changed in the set $v \in \{1, 1.5, 2, 2.5\}$ for $d = 100$.

Since this paper focuses on the linear kernel-based statistics, we adopt three recently proposed linear statistics as baselines, including MMD_{Lin} (Lin-MMD) (Gretton et al., 2012a), KSD_{Lin} (Lin-KSD) (Liu, Lee, and Jordan, 2016) and FSSD. Since Gaussian kernels are universal (Steinwart, 2001), we adopt Gaussian kernels $\kappa(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\gamma\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2\right)$ with variable width $\gamma \in \{2^{-10}, 2^{-9}, \ldots, 2^{10}\}$ as our candidate kernel set. The kernel parameter is tuned using 50% of the sample size n. For all evaluations, we set the significance level $\alpha = 0.05$. All experiments are repeated 100 times. All implementations are in Python and R.

Here we first investigate the power of Lin-MMD, Lin-KSD, ERL-MMD, ERL-KSD and FSSD, and further provide deep insights of the proposed statistics. The first set of experiments are conducted on two Gaussians $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|0, \mathbf{I}_d)$ and $q(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|0, v\mathbf{I}_d)$ with variable variance $v \in \{1, 1.5, 2, 2.5\}$. We adopt a fixed dimension $d = 100$. To investigate the influence of the number of examples on the gap between the empirical likelihood-based statistics (ELR-MMD and ELR-KSD) and the original linear statistics (Lin-MMD and Lin-KSD), we observe the rejection rates of the statistics for different numbers of examples. The results are shown in Figure 2. We can find that the gap between Lin-MMD and ELR-MMD or between Lin-KSD and ELR-KSD becomes smaller as the number of examples becomes larger. This is because the influence of the enforced constraint $\sum_{i=1}^{N} p_i h_i = 0$ will be smaller under the null hypothesis for the larger number of examples. By comparing the y-axes of the two figures, we can see that the rejection rates of Lin-MMD and ELR-MMD are higher than those of Lin-KSD and ELR-KSD. Lin-KSD has the worst performance among all linear statistics. In this experiment, FSSD fails to reject the null hypothesis nearly in all cases, while the existing linear statistics Lin-MMD and Lin-KSD and the empirical likelihood ratio statistics ELR-MMD and ELR-KSD can perform well, independent of the data dimension. These results are in agreement with the analyses given in the end of the last section. It is known that FSSD has shown excellent performance on low dimensional data (Jitkrittum et al., 2017). The proposed empirical likelihood ratio statistics can be considered as a complement to FSSD for high dimensional data, since they have shown higher power than the existing linear statistics.

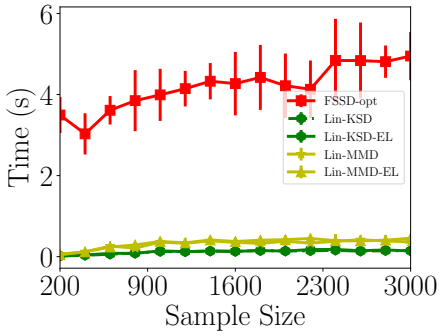
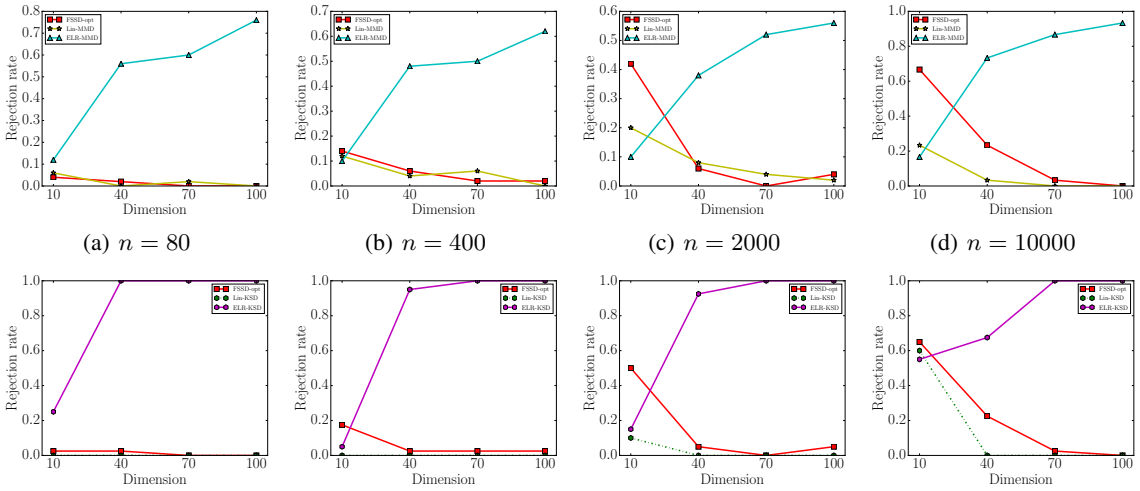


Figure 4: Running time comparison of FSSD, Lin-MMD, ELR-MMD, Lin-KSD and ELR-KSD on Gaussian $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|0, \mathbf{I}_d)$ and Laplacian $q(\boldsymbol{x}) = \prod_{i=1}^{d} \text{Laplace}(\boldsymbol{x}_i|0, 1/\sqrt{2})$ with $d = 100$ with variable size $n \in \{200, 400, \ldots, 3000\}$.

In the second experiment, we adopt two distributions Gaussian $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|0, \mathbf{I}_d)$ and Laplacian $q(\boldsymbol{x}) = \prod_{i=1}^{d} \text{Laplace}(\boldsymbol{x}_i|0, 1/\sqrt{2})$, in which the parameters are set to make p and q have the same mean and variance. We change the dimension d from 1 to 100 to observe the influence of the dimension on different statistics. The results



(a) $n = 80$ (b) $n = 400$ (c) $n = 2000$ (d) $n = 10000$

(e) $n = 80$ (f) $n = 400$ (g) $n = 2000$ (h) $n = 10000$

Figure 3: Comparison of rejection rates between Lin-MMD, ELR-MMD, Lin-KSD, ELR-KSD and FSSD on Gaussian $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|0, \mathbf{I}_d)$ and Laplacian $q(\boldsymbol{x}) = \prod_{i=1}^{d} \text{Laplace}(\boldsymbol{x}_i|0, 1/\sqrt{2})$ with variable dimension $d \in \{10, 40, 70, 100\}$ for $n = 80, 400, 2000, 10000$.

for different sample sizes $n \in \{80, 400, 2000\}$ are shown in Figure 3. We observe that FSSD has high power when the dimension $d = 1$, but the power quickly drops as the dimension increases. When the dimension $d \geq 10$, FSSD has poor performance (the power is less than 0.5), and when the dimension $d \geq 40$, FSSD fails to reject the null hypothesis. In this experiment, the difference between p and q is subtle, because they have the same mean and variance. Lin-MMD and Lin-KSD nearly fail to detect this subtle difference even with a large sample size, whereas ELR-MMD and ELR-KSD show remarkably good performance. There are two reasons for the impressive performance of the empirical likelihood ratio statistics. First, enforcing a probability on each pairwise discrepancy can help discriminate the subtle difference between two distributions. Second, the rejection regions of the proposed statistics are obtained by contouring a log likelihood ratio that may be the most powerful test for a fixed significance level α by Neyman-Pearson lemma (Neyman and Pearson, 1933). This point still needs to be theoretically proved though.

In the third experiment, we compare the running time of all these linear statistics. The results are shown in Figure 4. We observe that the running time of the empirical likelihood-based statistics ELR-MMD and ELR-KSD are almost the same as that of the linear statistics Lin-MMD and Lin-KSD, and all these linear statistics are much faster than FSSD. There are two reasons for the low efficiency of FSSD. First, under the null hypothesis, the asymptotic distribution of FSSD takes the form of a finite weighted sum of independent χ^2 variables, so it requires bootstrap or simulation to get the threshold for rejecting the null hypothesis (Jitkrittum et al., 2017), which is time-consuming. Second, FSSD optimizes the test locations $V = \{v_1, \ldots, v_J\} \subset \mathbb{R}^d$ via gradient ascent to get better performance than FSSD-rand[2](Jitkrittum et al., 2017).

In the fourth experiment, we check the Type-I errors (false rejection rates) of all tests. We consider a 10-dimensional Gaussian distribution and a Gaussian-Bernoulli restricted Boltzmann machine (RBM), which is a hidden variable graphical model consisting of a continuous observable variable $\boldsymbol{x} \in \mathbb{R}^d$ and a binary hidden variable $h \in \{\pm 1\}^{d_h}$, with joint probability

$$p(\boldsymbol{x}, h) = \frac{1}{Z} \exp(\boldsymbol{x}^{\mathrm{T}} \mathbf{B} h + b^{\mathrm{T}} \boldsymbol{x} + c^{\mathrm{T}} \boldsymbol{x} - \frac{1}{2} \|\boldsymbol{x}\|^2).$$

The results are demonstrated in Figure 5, which shows the rejection rates of all the tests as the sample size increases when p and q are the same RBM or Gaussian distribution. All the tests have roughly the right false rejection rates at the set significance level $\alpha = 0.05$

Conclusions

In this paper, we introduce the empirical likelihood method into the kernel test domain for the first time and propose two novel empirical likelihood ratio (ELR) statistics for the two sample test and the goodness of fit test, respectively. The proposed statistics have better performance than the existing MMD_{Lin} and KSD_{Lin}, and alleviate the high dimensionality curse faced by the best linear statistic, FSSD. We demonstrate the limiting chi-square distributions of the proposed novel statistics by proving the Wilks' theorems. We empirically verify the performance of the ELR statistics for high dimensional data and provide deep insights of the proposed statistics as compared to the state-of-the-art statistics.

[2]In FSSD-rand, the test locations are set to random draws from a multivariate normal distribution fitted to the data (Jitkrittum et al., 2017).

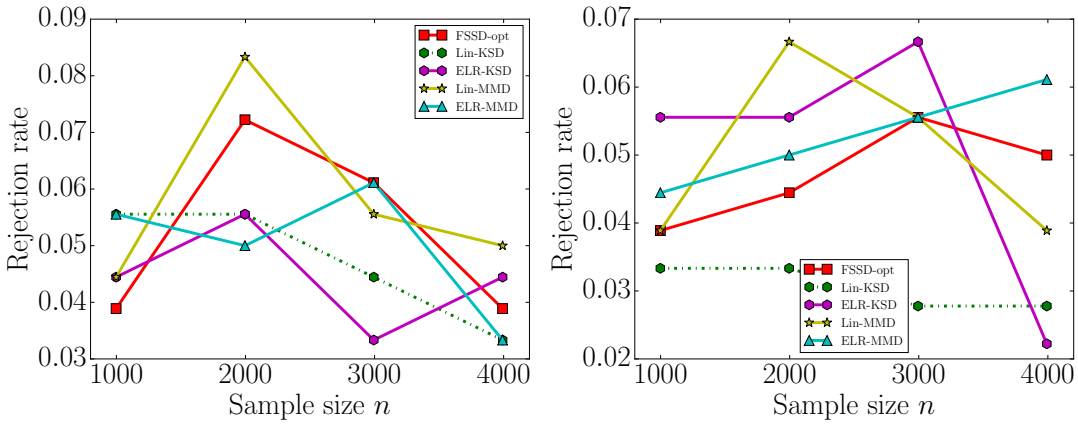


Figure 5: Type I errors of all tests. The left one is for two Gaussian distributions and the right one is for two RBMs.

In the future, we will exploit the higher dimensional empirical likelihood statistics to further improve the performance of the kernel tests and apply the empirical likelihood method to other statistical test problems, such as the independence test (Zhang et al., 2018) and conditional independence test (Ji et al., 2017; Strobl, Zhang, and Visweswaran, 2017).

References

Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.

Borgwardt, K. M.; Gretton, A.; Rasch, M. J.; Kriegel, H.-P.; Schölkopf, B.; and Smola, A. J. 2006. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* 22(14):e49–e57.

Chandrasekaran, V.; Srebro, N.; and Harsha, P. 2008. Complexity of inference in graphical models. In *UAI 2018*, 70–78. AUAI Press.

Chwialkowski, K.; Strathmann, H.; and Gretton, A. 2016. A kernel test of goodness of fit. In *ICML 2016*, 2606–2615.

Cucker, F., and Smale, S. 2002. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society* 39(1):1–49.

Fukumizu, K.; Bach, F. R.; and Jordan, M. I. 2004. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research* 5:73–99.

Gretton, A.; Borgwardt, K. M.; Rasch, M.; Schölkopf, B.; and Smola, A. J. 2007. A kernel method for the two-sample-problem. In *NIPS 19*, 513–520.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. J. 2012a. A kernel two-sample test. *Journal of Machine Learning Research* 13:723–773.

Gretton, A.; Sejdinovic, D.; Strathmann, H.; Balakrishnan, S.; Pontil, M.; Fukumizu, K.; and Sriperumbudur, B. K. 2012b. Optimal kernel choice for large-scale two-sample tests. In *NIPS 25*, 1205–1213.

Ji, S.; Ning, J.; Qin, J.; and Follmann, D. 2017. Conditional independence test by generalized kendalls tau with generalized odds ratio. *Statistical methods in medical research* 0962280217695345.

Jitkrittum, W.; Xu, W.; Szabó, Z.; Fukumizu, K.; and Gretton, A. 2017. A linear-time kernel goodness-of-fit test. In *NIPS 2017*, 261–270.

Koller, D., and Friedman, N. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.

Li, C.-L.; Chang, W.-C.; Cheng, Y.; Yang, Y.; and Póczos, B. 2017. Mmd gan: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, 2203–2213.

Liu, Q.; Lee, J.; and Jordan, M. 2016. A kernelized Stein discrepancy for goodness-of-fit tests. In *ICML 2016*, 276–284.

Lloyd, J. R., and Ghahramani, Z. 2015. Statistical model criticism using kernel two sample tests. In *NIPS 28*, 829–837.

Micchelli, C. A.; Xu, Y.; and Zhang, H. 2006. Universal kernels. *Journal of Machine Learning Research* 7:2651–2667.

Muandet, K.; Fukumizu, K.; Sriperumbudur, B.; Schölkopf, B.; et al. 2017. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning* 10(1-2):1–141.

Neyman, J., and Pearson, E. S. 1933. On the problem of the most efficient tests of statistical inference. *Biometrika A* 20:175–240.

Oates, C. J.; Girolami, M.; and Chopin, N. 2017. Control functionals for monte carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79(3):695–718.

Owen, A. B. 1988. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 75(2):237–249.

Owen, A. B. 1990. Empirical likelihood ratio confidence regionsn. *Annals of Statistics* 18(1):90–120.

Owen, A. B. 2001. *Empirical Likelihood*. Chapman and Hall/CRC, New York.

Ramdas, A.; Reddi, S. J.; Poczos, B.; Singh, A.; and Wasserman, L. 2014. On the high-dimensional power of linear-time kernel two-sample testing under mean-difference alternatives. *arXiv preprint arXiv:1411.6314*.

Salakhutdinov, R. 2015. Learning deep generative models. *Annual Review of Statistics and Its Application* 2:361–385.

Song, L.; Smola, A. J.; Gretton, A.; Bedo, J.; and Borgwardt, K. 2012. Feature selection via dependence maximization. *Journal of Machine Learning Research* 13:1393–1434.

Sriperumbudur, B. K.; Fukumizu, K.; Gretton, A.; Lanckriet, G. R.; and Schölkopf, B. 2009. Kernel choice and classifiability for RKHS embeddings of probability distributions. In *NIPS 22*, 1750–1758.

Sriperumbudur, B. K.; Gretton, A.; Fukumizu, K.; Schölkopf, B.; and Lanckriet, G. R. G. 2010. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research* 11:1517–1561.

Stein, C., et al. 1972. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. The Regents of the University of California.

Steinwart, I. 2001. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research* 2:67–93.

Strobl, E. V.; Zhang, K.; and Visweswaran, S. 2017. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *arXiv preprint arXiv:1702.03877*.

Zaremba, W.; Gretton, A.; and Blaschko, M. 2013. B-test: A non-parametric, low variance kernel two-sample test. In *NIPS 26*, 755–763.

Zhang, Q.; Filippi, S.; Gretton, A.; and Sejdinovic, D. 2018. Large-scale kernel methods for independence testing. *Statistics and Computing* 28(1):113–130.

Zhao, J., and Meng, D. 2015. FastMMD: Ensemble of circular discrepancy for efficient two-sample test. *Neural Computation* 27(6):1345–1372.