Linear Kernel Tests via Empirical Likelihood for High-Dimensional Data

Lizhong Ding,[1,*] Zhi Liu,[3] Yu Li,[2] Shizhong Liao,[4] Yong Liu,[5]
Peng Yang,[2] Ge Yu,[6] Ling Shao,[1] Xin Gao[2,*]
[1]Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, UAE
[2]King Abdullah University of Science and Technology (KAUST), Saudi Arabia
[3]University of Macau, China, [4]Tianjin University, China, [5]Institute of Information Engineering, CAS, China
[6]Technology and Engineering Center for Space Utilization, CAS, China
[*]Corresponding authors (email: lizhong.ding@inceptioniai.org, xin.gao@kaust.edu.sa)

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 fundamental statistical test problems: the two-sample test and the goodness-of-fit test. 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 it is that a set of samples is generated from a known target distribution. Specifically, we propose empirical likelihood ratio (ELR) statistics 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 for high-dimensional data. We prove the nonparametric Wilks' theorems for the ELR statistics, which illustrate that the limiting distributions of the proposed ELR statistics are chi-square distributions. With these limiting distributions, we can avoid bootstraps or simulations to determine the threshold for rejecting the null hypothesis, which makes the ELR statistics more efficient than the recently proposed linear statistic, finite set Stein discrepancy (FSSD). We also prove the consistency of the ELR statistics, which guarantees that the test power goes to 1 as the number of samples goes to infinity. In addition, we experimentally demonstrate and theoretically analyze that FSSD has poor performance or even fails 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 linear statistics.

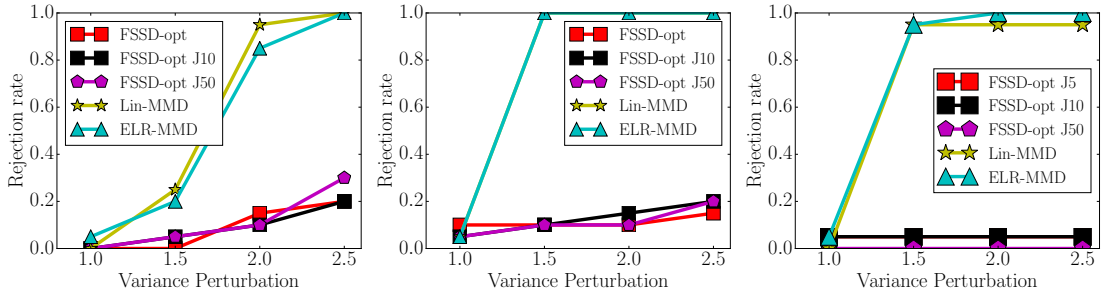
Introduction

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 a wide spectra of applications in various areas (Lloyd and Ghahramani 2015; Li et al. 2017; Yang 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 $\mathcal{D}_x = \{x_i\}_{i=1}^n \subset \mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{D}_y = \{y_j\}_{j=1}^m \subset \mathcal{Y} \subseteq \mathbb{R}^d$ 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 $\mathcal{D}_x = \{x_i\}_{i=1}^n \subset \mathcal{X} \subseteq \mathbb{R}^d$ from an unknown distribution q. Both of these two problems can be formulated as a hypothesis test, where the null hypothesis $H_0 : p = q$ is tested against the alternative hypothesis $H_1 : p \neq q$. The knowledge of p is what distinguishes the goodness-of-fit test from the two-sample test.

The two-sample test and the goodness-of-fit test are generally difficult in practice, since the underlying distributions (or one of the distributions) are unknown apriori. Kernel methods provide an effective way to implicitly transform data into a new feature space with the carefully-chosen kernel functions and kernel parameters (Liu and Liao 2015; Liu et al. 2017; Ding and Liao 2014a; 2017; Ding et al. 2019). The corresponding reproducing kernel Hilbert spaces (RKHSs) have strong representative power (Cucker and Smale 2002; Li et al. 2018). We adopt the unit balls in universal RKHSs as function classes (Muandet et al. 2017; Ding and Liao 2014b) 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 statistic, maximum mean discrepancy (MMD), was designed to measure two distributions by embedding them in an RKHS (Gretton et al. 2012). MMD has been attracting much attention in two-sample test research due to its solid theoretical foundation (Sriperumbudur et al. 2009; Gretton et al. 2012; Song et al. 2012; Zaremba, Gretton, and Blaschko 2013; Ding et al. 2018). The minimum variance unbiased estimator $\mathrm{MMD}_{\mathrm{Unb}}$ of MMD was first proposed in (Gretton et al. 2012) on the basis of n samples being observed from each of p and q. However, the estimation of the asymptotic distribution of $\mathrm{MMD}_{\mathrm{Unb}}$ under the null distribution requires bootstrap or moment matching to determine the test threshold, which costs at least $O(n^2)$. Later, an $O(n)$ unbiased estimator $\mathrm{MMD}_{\mathrm{Lin}}$ was proposed (Gretton et al. 2012), using a subsampling of the terms in $\mathrm{MMD}_{\mathrm{Unb}}$. $\mathrm{MMD}_{\mathrm{Lin}}$ has higher variance than $\mathrm{MMD}_{\mathrm{Unb}}$, but it is computationally much more appealing.

For the goodness-of-fit test, traditional methods need to calculate the likelihoods of the models. However, for large graphical models or deep generative models, this is often computationally intractable due to the complex-



(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 $p(x) = \mathcal{N}(x|0, \mathbf{I}_d)$ and $q(x) = \mathcal{N}(x|0, v\mathbf{I}_d)$ with the variance changed in the set $v \in \{1, 1.5, 2, 2.5\}$ for $d = 10, 50, 100$. The abbreviation "opt" in FSSD-opt means that all parameters in FSSD are optimized, including the kernel parameter and all test locations (Jitkrittum et al. 2017). The results of FSSD-opt for the number of test locations $J = 5, 10, 50$ are all shown in this figure.

ity of the probabilistic models. Recently, Stein's method (Stein and others 1972; Oates, Girolami, and Chopin 2017) has been introduced into the kernel domain (Gorham and Mackey 2017), by combining Stein's identity with the RKHS theory, which is a *likelihood-free* method and depends on p only through logarithmic derivatives. The proposed statistic is referred to as kernel Stein discrepancy (KSD) (Chwialkowski, Strathmann, and Gretton 2016; Liu, Lee, and Jordan 2016). Since the null distribution of the unbiased estimator KSD_{Unb} of KSD does not have an analytical form, the bootstrap was adopted to calculate the approximate rejection threshold, whose time complexity is $O(n^2)$. A linear statistic, KSD_{Lin}, was proposed using half-sampling, which has a zero-mean Gaussian limit under the null hypothesis (Liu, Lee, and Jordan 2016). To improve the performance of the existing linear 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) owes its broad usage and fast research development to a number of important advantages in statistics. Generally speaking, it combines the reliability of nonparametric methods with the effectiveness of the likelihood approach. Taking into consideration the asymptotic normality of the linear unbiased estimator MMD_{Lin} (Gretton et al. 2012), we first propose an ELR statistic based on the formulation of MMD_{Lin}, named ELR-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 ELR-MMD, which shows that the proposed ELR-MMD has a limiting chi-square distribution. For the goodness-of-fit test, we propose an ELR statistic based on the linear unbiased estimator KSD_{Lin}, called ELR-KSD, by enforcing an empirical distribution on the pairwise discrepancies. We derive the nonparametric Wilks' theorem to show the limiting distribution of ELR-KSD. The proposed ELR-MMD and ELR-KSD statistics show better performance than MMD_{Lin} and KSD_{Lin}, and remarkably higher discriminability (power) when testing two distributions with subtle differences. There are two possible reasons for the impressive performance of the ELR statistics. First, enforcing a probability on each pairwise discrepancy can help discriminate the subtle difference between two distributions. Second, the rejection regions of the ELR statistics are obtained by contouring a logarithmic likelihood ratio in what may be the most powerful test for a fixed significance level α by Neyman-Pearson lemma (Neyman and Pearson 1933). We further prove the consistency of the proposed ELR statistics, which guarantees that the test power (i.e., the probability of correctly rejecting H_0 when H_1 holds) goes to 1, as the number of samples goes to infinity.

Another contribution of this paper is that we experimentally demonstrate that the recently proposed FSSD has poor performance or even fails 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 normal distributions $p(x) = \mathcal{N}(x|0, \mathbf{I}_d)$ and $q(x) = \mathcal{N}(x|0, v\mathbf{I}_d)$ with the variance changed in the set $v \in \{1, 1.5, 2, 2.5\}$ for $d = 10, 50, 100$. We find that both the existing statistic MMD_{Lin} and the proposed ELR-MMD work well for $d = 10, 50, 100$, but FSSD shows poor rejection rates for $d = 10$, and fails to reject the null hypothesis 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 almost the same (see Figure 1). We will further provide a deeper understanding of FSSD and analyze the possible reasons why FSSD shows poor performance on high-dimensional data. Since FSSD has shown good performance on low-dimensional data (Jitkrittum et al. 2017), the proposed ELR statistics can be considered as complements to

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

FSSD for high-dimensional data.

Empirical Likelihood Ratio for Two Sample Test

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

Assume that the data domain is a compact set $\mathcal{X} \in \mathbb{R}^d$. Let $\mathcal{H}_\kappa$ be a reproducing kernel Hilbert space (RKHS) defined on $\mathcal{X}$ with the reproducing kernel $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and p a Borel probability measure on $\mathcal{X}$. We adopt a unit ball in a universal RKHS $\mathcal{H}_\kappa$ as the function class $\mathcal{F}$, 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 the space of bounded continuous functions $C(\mathcal{X})$ with respect to the L_∞ norm. Gaussian and Laplace RKHSs are universal[2] (Steinwart 2001). Kernel parameters can be chosen via cross validation (Ding and Liao 2011; 2012; Liu, Jiang, and Liao 2014; Liu et al. 2018).

The mean embedding of a distribution p in $\mathcal{F}$, written as $\mu_\kappa(p) \in \mathcal{F}$, is defined such that $\mathbf{E}_{x\sim p} f(x) = \langle f, \mu_\kappa(p)\rangle$ for all $f \in \mathcal{F}$. The squared MMD between two distributions p and q is the squared RKHS distance between their respective mean embeddings,

$$\mathrm{MMD}^2[\mathcal{F}, p, q] = \|\mu_\kappa(p) - \mu_\kappa(q)\|^2_{\mathcal{F}} = \mathbf{E}_{zz'} h(z, z'),$$

where $z = (x, y)$, $z' = (x', y')$ and $h(z, z') = \kappa(x, x') + \kappa(y, y') - \kappa(x, y') - \kappa(x', y)$. It has been proved that for a unit ball $\mathcal{F}$ in a universal RKHS, $\mathrm{MMD}[\mathcal{F}, p, q] = 0$ if and only if $p = q$ (Gretton et al. 2012).

For two sets of samples $\mathcal{D}_x = \{x_i\}_{i=1}^n \subset \mathcal{X} \subseteq \mathbb{R}^d$, where $x_i \sim p$ i.i.d., and $\mathcal{D}_y = \{y_j\}_{j=1}^m \subset \mathcal{Y} \subseteq \mathbb{R}^d$, where $y_i \sim q$ i.i.d., if we assume $m = n$, the minimum variance unbiased estimator of $\mathrm{MMD}^2[\mathcal{F}, p, q]$ can be represented as

$$\mathrm{MMD}^2_{\mathrm{Unb}}[\mathcal{F}, \mathcal{D}_x, \mathcal{D}_y] = \frac{1}{n(n-1)} \sum_{i \neq j}^{n} h(z_i, z_j).$$

$\mathrm{MMD}_{\mathrm{Unb}}$ requires $O(n^2)$ time to compute h on all interacting pairs. The null distribution of $\mathrm{MMD}_{\mathrm{Unb}}$ does not have an analytical form, so the bootstrap or moment matching are required with $O(n^2)$ time complexity. A linear time unbiased estimator $\mathrm{MMD}_{\mathrm{Lin}}$ was proposed in (Gretton et al. 2012),

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

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

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

where $p_i = d\rho(h_i) = \Pr(H = h_i)$. Only distributions with an atom of probability on each h_i have nonzero likelihood (Owen 1988) and $L(\rho)$ is maximized by the empirical distribution function $\rho_N(h) = N^{-1} \sum_{i=1}^{N} I(h_i < h)$, where I is an indicator function (Qin and Lawless 1994). The empirical likelihood ratio (ELR) is then defined as $R(\rho) = L(\rho)/L(\rho_N)$, and it is easy to show that $R(\rho) = \prod_{i=1}^{N} N p_i$.

Now we define an ELR function $\Psi_{\mathrm{tst}}(\mu)$ for the two-sample test problem in Equation (1), in which we enforce a probability $\{p_i \geq 0\}_{i=1}^N$ on the pairwise discrepancies $\{h_i \geq 0\}_{i=1}^N$. There are two virtues of $\Psi_{\mathrm{tst}}(\mu)$. First, the constraint $\sum_{i=1}^N p_i h_i = \mu$ forces the empirical mean to be the expectation μ, which makes the maximum of the empirical likelihood ratio more trustful. For example, in the two-sample test, under $H_0 : p = q$, we have $\mathbf{E}[h_i] = 0$ and the constraint $\sum_{i=1}^N p_i h_i = 0$ guarantees the empirical mean of all discrepancies h_i to be 0. Second, optimizing the probability p_i on each pairwise discrepancy h_i can help discriminate the subtle difference between two distributions.

$$\Psi_{\mathrm{tst}}(\mu) = \max_{\{p_i \geq 0\}_{i=1}^N} \left\{ \prod_{i=1}^{N} N p_i \,\middle|\, \sum_{i=1}^{N} p_i = 1, \sum_{i=1}^{N} p_i h_i = \mu \right\}. \quad (1)$$

A unique value for the right-hand side of Equation (1) exists, provided that μ is inside the convex hull of the points $h_1, \ldots, h_N$ (Owen 1990; 2001). An explicit expression for $\Psi_{\mathrm{tst}}(0)$ can be derived by a Lagrange multiplier argument: the maximum of $\prod_{i=1}^N N p_i$ 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

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

where λ is the solution to $\sum_{i=1}^N \frac{h_i}{1+\lambda h_i} = 0$.

Now we propose an ELR test statistic for the two-sample test problem as

$$W_{\mathrm{tst}}(0) = -2 \log \Psi_{\mathrm{tst}}(0) = 2 \sum_{i=1}^{N} \log(1 + \lambda h_i).$$

We derive the Wilks' theorem (Theorem 1) for the ELR test statistic $W_{\mathrm{tst}}(0)$, which shows that $W_{\mathrm{tst}}(0)$ has a limiting chi-square distribution.

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

$$W_{\mathrm{tst}}(0) \xrightarrow{d} \chi^2_{(1)},$$

where $\chi^2_{(1)}$ is the chi-square distribution with 1 degree of freedom.

[2]Universal kernels can be used to approximate any target function in $C(\mathcal{X})$, That is, the corresponding RKHSs are dense in $C(\mathcal{X})$.

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

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

Since the limiting distribution is $\chi^2_{(1)}$, we can obtain the threshold for rejection directly from the chi-square table, without needing time-consuming bootstraps or simulations. The main computational burden for $W_{\text{tst}}(0)$ is the calculation of h_i, $i = 1, \ldots, N$. Therefore, the time complexity of $W_{\text{tst}}(0)$ is linear in the number of samples. The proposed ELR statistic can easily be extended to B-test (Zaremba, Gretton, and Blaschko 2013), since the statistics for different blocks in B-test are independent from each other.

Theorem 2 guarantees the test consistency of the proposed ELR statistic $W_{\text{tst}}(0)$, that is, when the number of samples are large enough, $W_{\text{tst}}(0)$ can always correctly reject the null hypothesis. We write $\Pr_{H_1}$ for the distribution of $W_{\text{tst}}(0)$ under H_1.

Theorem 2. *Under $H_1 : p \neq q$, if $\mathbf{E}_{x,x'}[\kappa^2(x, x')] < \infty$, the test power,*

$$\Pr_{H_1}\left(W_{\text{tst}}(0) \geq \chi^2_\alpha\right) \to 1,$$

as $n \to \infty$.

Empirical Likelihood Ratio for Goodness of Fit Test

In this section, we will propose an ELR 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 logarithmic derivatives. A Stein operator T_p takes a multivariate function $f(x) = (f_1(x), \ldots, f_d(x))^{\mathrm{T}} \in \mathbb{R}^d$ as input and outputs a function $(T_p f)(x) : \mathbb{R}^d \to \mathbb{R}$. The function $T_p f$ has the key property that for all fs in an appropriate function class,

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

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

We consider the function class $\mathcal{F}^d := \mathcal{F} \times \cdots \times \mathcal{F}$, where $\mathcal{F}$ is a unit-norm ball in a universal RKHS. Assume that $f_i \in \mathcal{F}$ for all $i = 1, \ldots d$ so that $f \in \mathcal{F}^d$ with the inner product $\langle f, f' \rangle_{\mathcal{F}^d} := \sum_{i=1}^d \langle f_i, f_i' \rangle_{\mathcal{F}}$. According to the reproducing property of $\mathcal{F}$, $f_i(x) = \langle f_i, \kappa(x, \cdot)\rangle_{\mathcal{F}}$, and that $\frac{\partial \kappa(x,\cdot)}{\partial x_i} \in \mathcal{F}$, we can define $\omega_p(x, \cdot) = \frac{\partial \log p(x)}{\partial x}\kappa(x, \cdot) + \frac{\kappa(x,\cdot)}{\partial x}$. The kernel Stein operator can be written as

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

Kernel Stein discrepancy (KSD) is defined as

$$\begin{aligned}\text{KSD}[\mathcal{F}^d, \mathcal{D}_x, p] &= \sup_{\|f\|_{\mathcal{F}^d} \leq 1} \langle f, \mathbf{E}_{x\sim q}\omega_p(x, \cdot)\rangle \\ &:= \|g(\cdot)\|_{\mathcal{F}^d},\end{aligned} \tag{2}$$

where $g(\cdot) = \mathbf{E}_{x\sim q}\omega_p(x, \cdot)$. When $\mathbf{E}_{x\sim p}\|\nabla_x \log p(x) - \nabla_x \log q(x)\| < \infty$, it can be shown that $\text{KSD}[\mathcal{F}^d, \mathcal{D}_x, p] = 0$ if and only if $p = q$. The squared KSD can be written as

$$\text{KSD}^2[\mathcal{F}^d, \mathcal{D}_x, p] = \mathbf{E}_{x\sim q}\mathbf{E}_{x'\sim q} h_p(x, x'),$$

where $h_p(x, x') = s_p^{\mathrm{T}}(x)s_p(x')\kappa(x, x') + s_p^{\mathrm{T}}\nabla_x\kappa(x, x') + s_p^{\mathrm{T}}\nabla_{x'}\kappa(x, x') + \sum_{i=1}^d \frac{\partial^2\kappa(x,x')}{\partial x_i \partial x_i'}$, and $s_p(x) = \nabla_x \log p$, which is called the score function. We denote the unbiased empirical estimator of KSD (Liu, Lee, and Jordan 2016) as

$$\text{KSD}^2_{\text{Unb}}[\mathcal{F}^d, \mathcal{D}_x, p] = \frac{2}{n(n-1)} \sum_{i<j} h_p(x_i, x_j).$$

The computational cost of $\text{KSD}^2_{\text{Unb}}$ is $O(n^2)$. To reduce this cost, a linear time estimator was proposed in (Liu, Lee, and Jordan 2016) and we write it as

$$\text{KSD}^2_{\text{Lin}}[\mathcal{F}^d, \mathcal{D}_x, p] = \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} h_p(x_{2i-1}, x_{2i}).$$

Now we derive an ELR statistic for the goodness-of-fit test problem. We write $h_{p,i} = h_p(x_{2i-1}, x_{2i})$ and $N = \lfloor n/2 \rfloor$. When calculating $h_{p,i}$, $i = 1, \ldots, N$, different independent samples are used for different i. $h_{p,1}, h_{p,2}, \ldots, h_{p,N}$ are i.i.d observations from a univariate distribution. Now we define an ELR function

$$\begin{aligned}&\Psi_{\text{goft}}(\mu) \\ &= \sup_{\{p_i \geq 0\}_{i=1}^N} \left\{ \prod_{i=1}^N N p_i \,\middle|\, \sum_{i=1}^N p_i = 1, \sum_{i=1}^N p_i h_{p,i} = \mu \right\}.\end{aligned}$$

Under $H_0 : p = q$, we have $\mathbf{E}[h_{p,i}] = 0$ and set $\mu = 0$. We use the Lagrange multiplier method to derive the explicit expression of $\Psi_{\text{goft}}(0)$. For $i = 1, \ldots, N$, we have

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

where λ is the solution to $\sum_{i=1}^N \frac{h_{p,i}}{1+\lambda h_{p,i}} = 0$.

Now we define an ELR test statistic for the goodness of test problem as follows,

$$W_{\text{goft}}(0) = -2 \log \Psi_{\text{goft}}(0) = 2 \sum_{i=1}^N \log\left(1 + \lambda h_{p,i}\right).$$

We derive the Wilks' theorem for $W_{\text{goft}}(0)$, which shows a limiting chi-square distribution of $W_{\text{goft}}(0)$.

Theorem 3 (Wilks' Theorem)**.** *Under $H_0 : p = q$, if $\mathbf{E}_{x,x'}[\kappa^2(x, x')] < \infty$, the ELR test statistic*

$$W_{\text{goft}}(0) \xrightarrow{d} \chi^2_{(1)}.$$

Based on Theorem 3, we will reject H_0, when $W_{\text{goft}}(0) \geq \chi^2_\alpha$ with χ^2_α satisfying $\Pr(\chi^2_{(1)} \geq \chi^2_\alpha) = \alpha$. The main computational burden for $W_{\text{goft}}(0)$ is the calculation of $h_{p,i}$, $i = 1, \ldots, N$. Therefore, the time complexity of $W_{\text{goft}}(0)$ is linear in the number of samples.

Theorem 4 guarantees the test consistency of $W_{\text{goft}}(0)$.

Theorem 4. *Under $H_1 : p \neq q$, if $\mathbf{E}_{x,x'}[\kappa^2(x, x')] < \infty$, the test power,*

$$\Pr_{H_1}\left(W_{\text{goft}}(0) \geq \chi^2_\alpha\right) \to 1,$$

as $n \to \infty$.

Comparisons with FSSD

In this section, we compare FSSD (Jitkrittum et al. 2017) with existing linear statistics and our ELR statistics, and analyze the possible reasons why FSSD shows poor performance or even fails for high-dimensional data.

We first briefly introduce FSSD. Let $V = \{v_1, \ldots, v_J\} \subset \mathbb{R}^d$ be random vectors drawn from a distribution. The statistic of FSSD is defined as

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

where $g(\cdot)$ is referred to as the Stein witness function, given in Equation (2). It has been proved (Jitkrittum et al. 2017) that if the following conditions are satisfied, 1) κ is a universal and analytic function; 2) $\mathbf{E}_{x\sim q}\mathbf{E}_{x'\sim p} h_p(x, x') < \infty$; 3) $\mathbf{E}_{x\sim q}\|\nabla_x \log p(x) - \nabla_x \log q(x)\|^2 < \infty$; and 4) $\lim_{\|x\|\to\infty} p(x)g(x) = 0$; for any $J \geq 1$, almost surely $\text{FSSD}^2_p(q) = 0$ if and only if $p = q$. Let $\Omega(x) \in \mathbb{R}^{d\times J}$ such that $[\Omega(x)]_{i,j} = \omega_{p,i}(x, v_j)/\sqrt{dJ}$, $\tau(x) = \text{vec}(\Omega(x)) \in \mathbb{R}^{dJ}$, where $\text{vec}(\cdot)$ denotes the vectorization, and $\Delta(x, y) = \tau(x)^{\mathrm{T}}\tau(y)$. The unbiased estimator of $\text{FSSD}^2_p(q)$ is

$$\widehat{\text{FSSD}}^2 = \frac{2}{n(n-1)} \sum_{i<j} \Delta(x_i, x_j).$$

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

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

According to Proposition 2 in (Jitkrittum et al. 2017), under the $H_1 : p \neq q$,

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

if $\sigma_{H_1} = 4\mu^{\mathrm{T}}\mathbf{\Sigma}_q\mu > 0$, where $\mu = \mathbf{E}_{x\sim q}[\tau(x)]$ and $\mathbf{\Sigma}_q = \text{cov}_{x\sim q}[\tau(x)] \in \mathbb{R}^{dJ\times dJ}$. From the above equation, we know that $n\widehat{\text{FSSD}}^2$ is highly dependent on the dimension of the data: when the dimension d increases, the dimension of $\mathbf{\Sigma}_q$ will increase, and then the variance σ_{H_1} becomes larger. When the variance becomes larger, the resulting values of the statistic will become unstable. For MMD_{Lin}, KSD_{Lin}, ELR-MMD and ELR-KSD, the kernel function $\kappa(x, \cdot)$ only appears in the dot product form, and thus the statistics are less dependent on the dimension d of data.

In addition, under $H_0 : p = q$, the asymptotic distribution of $n\widehat{\text{FSSD}}^2$ is not an analytical form, but the existing linear statistics MMD_{Lin} and KSD_{Lin}, and the ELR statistics ELR-MMD and ELR-KSD all have analytical limiting distributions.

Experiments

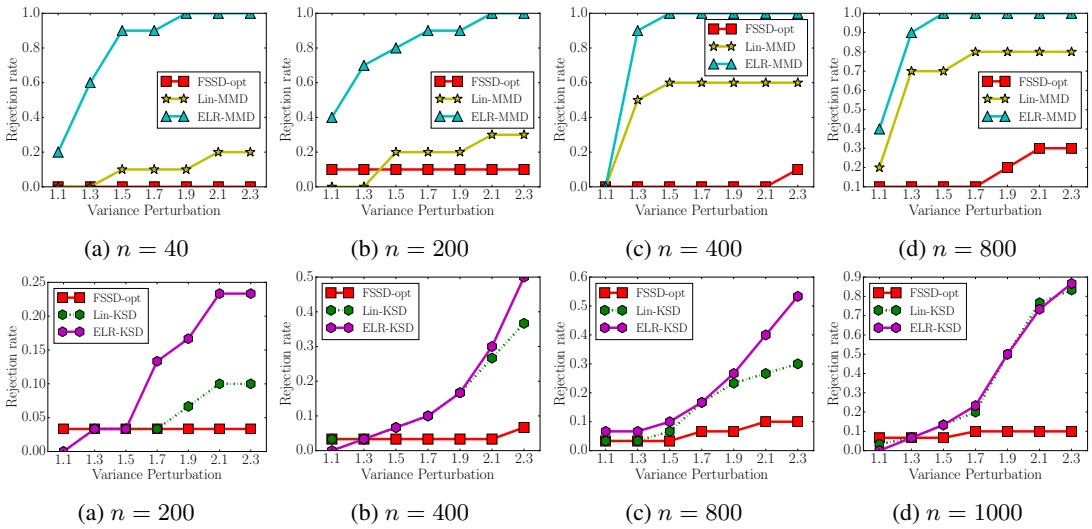
Here we conduct a series of experiments to evaluate the performance of the proposed ELR statistics and exploit the conditions under which the proposed statistics can perform well.

We compare the ELR statistics, ELR-MMD and ELR-KSD, with three existing linear nonparametric statistics, including MMD_{Lin} (Lin-MMD) (Gretton et al. 2012), KSD_{Lin} (Lin-KSD) (Liu, Lee, and Jordan 2016) and FSSD (FSSD-opt)[3] (Jitkrittum et al. 2017). Because Gaussian kernels are universal (Steinwart 2001), we adopt Gaussian kernels $\kappa(x, x') = \exp\left(-\gamma\|x - x'\|_2^2\right)$ with variable width $\gamma \in \{2^{-10}, 2^{-9}, \ldots, 2^{10}\}$ as our candidate kernel set. For all evaluations, we set the significance level $\alpha = 0.05$. All experiments are repeated 100 times. All implementations are in Python and R.

We investigate the power of Lin-MMD, Lin-KSD, ELR-MMD, ELR-KSD and FSSD, and provide deep insights into the proposed statistics.

The first set of experiments are conducted on two Gaussians $p(x) = \mathcal{N}(x|0, \mathbf{I}_d)$ and $q(x) = \mathcal{N}(x|0, v\mathbf{I}_d)$, with variable variance $v \in \{1.1, 1.3, \ldots, 2.3\}$. We adopt a fixed dimension $d = 100$. To investigate the influence of the number of samples on the gap between the ELR statistics (ELR-MMD and ELR-KSD) and the existing linear statistics (Lin-MMD and Lin-KSD), we observe the rejection rates of the statistics for different numbers of samples. 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 samples becomes larger.

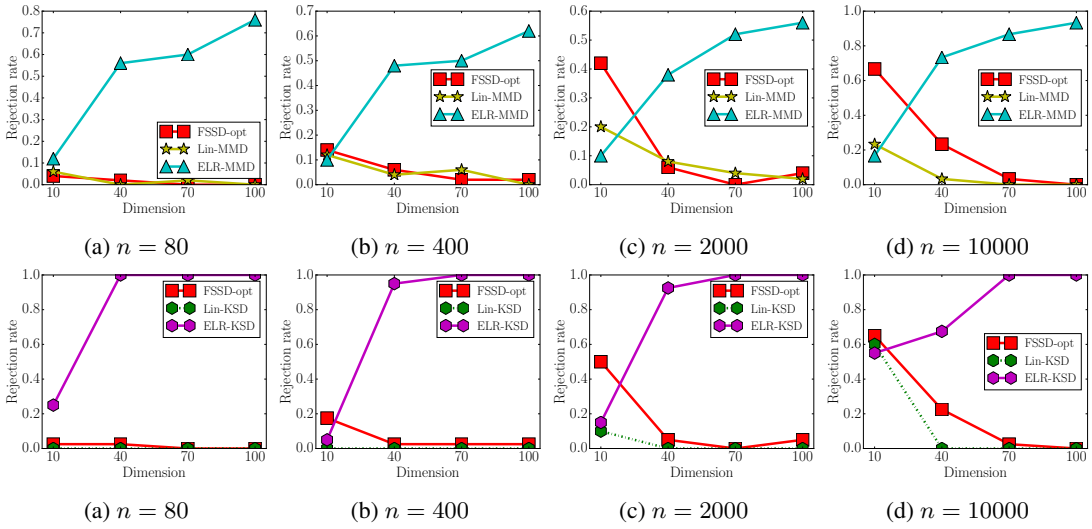
[3]The abbreviation "opt" in FSSD-opt means that all parameters in FSSD are optimized, including the kernel parameter and all test locations. We set the number of test locations $J = 5$ as in (Jitkrittum et al. 2017)



(a) $n = 40$ (b) $n = 200$ (c) $n = 400$ (d) $n = 800$

(a) $n = 200$ (b) $n = 400$ (c) $n = 800$ (d) $n = 1000$

Figure 2: Rejection rates of Lin-MMD, ELR-MMD, Lin-KSD, ELR-KSD and FSSD on two different normal distributions $p(x) = \mathcal{N}(x|0, \mathbf{I}_d)$ and $q(x) = \mathcal{N}(x|0, v\mathbf{I}_d)$ with the variance changed in the set $v \in \{1.1, 1.3, \ldots, 2.3\}$ for $d = 100$.



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

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

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

The possible reason is that, under $H_0 : p = q$, the influence of the enforced constraint $\sum_{i=1}^{N} p_i h_i = 0$ will become smaller as number of samples increases, since the expectation of the discrepancy h_i is 0. We can also see that the rejection rates of Lin-MMD and ELR-MMD are higher than those of Lin-KSD and ELR-KSD. In this experiment, FSSD shows low rejection rates or fails to test nearly in all cases, while the existing linear statistics Lin-MMD and Lin-KSD and the proposed statistics ELR-MMD and ELR-KSD can perform normally. These results are in agreement with the analyses given in the last section.

In the second experiment, we adopt the distributions Gaussian $p(x) = \mathcal{N}(x|0, \mathbf{I}_d)$ and Laplacian $q(x) = \prod_{i=1}^{d} \text{Laplace}(x_i|0, 1/\sqrt{2})$, in which the parameters are set to make p and q have the same mean and variance. We change the dimension d from 10 to 100 to observe the influence of the dimension on different statistics. The results for different sample sizes $n \in \{80, 400, 2000, 10000\}$ are shown in Figure 3. We observe that the power of FSSD quickly drops as the dimension increases. When the dimension $d = 40$, FSSD has poor performance (the power is less than 0.5), and when the dimension $d > 40$, 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. In the first experiment, Lin-MMD and Lin-KSD work well for the two Gaussian distributions, but they nearly fail to detect the subtle difference

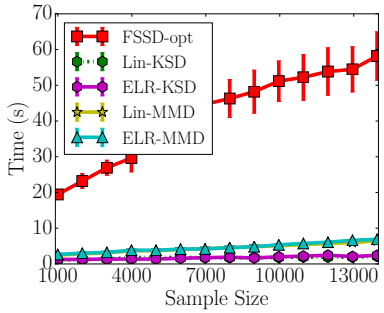


Figure 4: Running time of FSSD, Lin-MMD, ELR-MMD, Lin-KSD and ELR-KSD on Gaussian $p(x) = \mathcal{N}(x|0, \mathbf{I}_d)$ and Laplacian $q(x) = \prod_{i=1}^{d} \text{Laplace}(x_i|0, 1/\sqrt{2})$ with $d = 100$ with variable size $n \in \{1000, 2000, \ldots, 14000\}$.

in this experiment 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 ELR statistics. First, enforcing a probability on each pairwise discrepancy can help discriminate the subtle difference between two distributions. Second, the rejection thresholds of the ELR statistics are determined by contouring a logarithmic likelihood ratio, which may be the most powerful test for a fixed α. This point still needs to be theoretically supported by proving the empirical version of the Neyman-Pearson lemma (Neyman and Pearson 1933). It is known that FSSD has shown good performance on low-dimensional data (Jitkrittum et al. 2017). The proposed ELR statistics can be considered as complements to FSSD for high-dimensional data, since they have shown higher power than the existing linear statistics.

In the third experiment, we compare the running time of all linear statistics. The results are shown in Figure 4. We observe that the running time of the ELR 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 is not an analytical form, so it requires bootstraps or simulations to calculate 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[4](Jitkrittum et al. 2017).

In the fourth experiment, we check the Type I errors (false rejection rates) of all linear statistics. We consider a 10-dimensional Gaussian distribution and a Gaussian-Bernoulli restricted Boltzmann machine (RBM) (Liu, Lee, and Jordan 2016), which is a hidden variable graphical model consisting of a continuous observable variable $x \in \mathbb{R}^d$ and a binary hidden variable $r \in \{\pm 1\}^{d_h}$, with joint probability

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

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

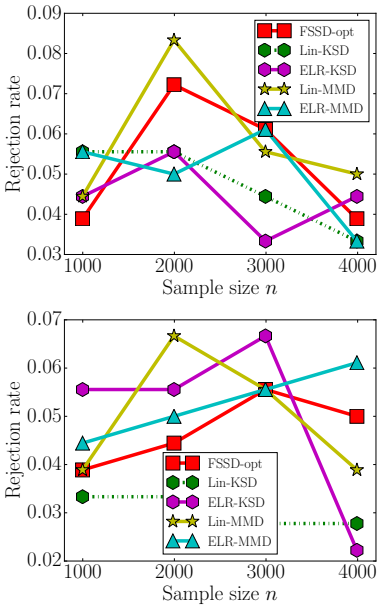


Figure 5: Type I errors (false rejection rates) of all linear tests with variable size $n \in \{1000, 2000, 3000, 4000\}$. The first one is for the 10-dimensional Gaussian distribution and the second one is for RBM.

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

Conclusions

In this paper, we established the first connection between the empirical likelihood and nonparametric kernel tests, and derived novel empirical likelihood ratio (ELR) statistics for the two-sample test and the goodness-of-fit test. We provided theoretical insights indicating that the ERL statistics have limiting chi-square distributions under the null hypothesis, and that their test consistencies hold under the alternative hypothesis. The new ELR statistics have empirically shown stronger test power than the existing linear statistics for high-dimensional data while preserving high computational efficiency. In the near future, we will develop ELR statistics for other nonparametric statistical test problems, including the independence test and the conditional independence test.

Acknowledgments

This publication is based upon work supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. URF/1/3007-01-01 and BAS/1/1624-01-01, National Natural Science Foundation of China (No. 61673293), National Natural Science Foundation of China (No. 61703396) and Shenzhen Government (GJHZ20180419190732022).

References

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

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

Ding, L., and Liao, S. 2011. Approximate model selection for large scale LSSVM. *Journal of Machine Learning Research - Proceedings Track* 20:165–180.

Ding, L., and Liao, S. 2012. Nyström approximate model selection for LSSVM. In *Advances in Knowledge Discovery and Data Mining — Proceedings of the 16th Pacific-Asia Conference (PAKDD)*, 282–293.

Ding, L., and Liao, S. 2014a. Approximate consistency: Towards foundations of approximate kernel selection. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Database (ECML PKDD)*, 354–369.

Ding, L., and Liao, S. 2014b. Model selection with the covering number of the ball of RKHS. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM)*, 1159–1168.

Ding, L., and Liao, S. 2017. An approximate approach to automatic kernel selection. *IEEE Transactions on Cybernetics* 47(3):554–565.

Ding, L.; Liao, S.; Liu, Y.; Yang, P.; and Gao, X. 2018. Randomized kernel selection with spectra of multilevel circulant matrices. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, 2910–2917.

Ding, L.; Liu, Y.; Liao, S.; Li, Y.; Yang, P.; Pan, Y.; Huang, C.; Shao, L.; and Gao, X. 2019. Approximate kernel selection with strong approximate consistency. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*.

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.

Gorham, J., and Mackey, L. 2017. Measuring sample quality with kernels. In *ICML*, 1292–1301.

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

Jitkrittum, W.; Xu, W.; Szabó, Z.; Fukumizu, K.; and Gretton, A. 2017. A linear-time kernel goodness-of-fit test. In *NIPS 30*, 261–270.

Li, C.-L.; Chang, W.-C.; Cheng, Y.; Yang, Y.; and Póczos, B. 2017. MMD GAN: Towards deeper understanding of moment matching network. In *NIPS 30*, 2203–2213.

Li, J.; Liu, Y.; Yin, R.; Zhang, H.; Ding, L.; and Wang, W. 2018. Multi-class learning: from theory to algorithm. In *NeurIPS 31*, 1593–1602.

Liu, Y., and Liao, S. 2015. Eigenvalues ratio for kernel selection of kernel methods. In *AAAI*, 2814–2820.

Liu, Y.; Liao, S.; Lin, H.; Yue, Y.; and Wang, W. 2017. Infinite kernel learning: generalization bounds and algorithms. In *AAAI*, 2280–2286.

Liu, Y.; Lin, H.; Ding, L.; Wang, W.; and Liao, S. 2018. Fast cross-validation. In *IJCAI*, 2497–2503.

Liu, Y.; Jiang, S.; and Liao, S. 2014. Efficient approximation of cross-validation for kernel methods using Bouligand influence function. In *ICML*, 324–332.

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

Qin, J., and Lawless, J. 1994. Empirical likelihood and general estimating equations. *Annals of Statistics* 300–325.

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 6th Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*.

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

Yang, P.; Zhao, P.; Zheng, V. W.; Ding, L.; and Gao, X. 2018. Robust asymmetric recommendation via min-max optimization. In *SIGIR*, 1077–1080.

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